DiffDVC: Accurate Event Detection for Dense Video Captioning via Diffusion Models

Wei Chen[1], **Jianwei Niu**[123], **Xuefeng Liu** [12*], **Zhendong Wang**[1], **Shaojie Tang**[4], **Guogang Zhu**[1]

[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]Zhengzhou University Research Institute of Industrial Technology, Zhengzhou University, Zhengzhou, China
[4]Department of Management Science and Systems, University at Buffalo, Buffalo, New York, United States
{chenweibuaa, niujianwei, liu_xuefeng, zhendong.wang, buaa_zgg}@buaa.edu.cn
shaojiet@buffalo.edu

Abstract

Dense video captioning (DVC) aims to describe multiple events within a video, and its performance is greatly affected by the accuracy of video event detection. Video event detection involves predicting the proposal boundaries (start and end times) and the classification score of each event in a video. Recently, a few methods have applied diffusion models originally designed for image object detection to detect events in DVC. These methods add noise to the ground-truth event proposal boundaries, and subsequently learn the denoising process. However, these methods often overlook the fundamental differences between videos and images. We observe that, whereas in images the important information for object classification is normally around the boundaries of the ground-truth boxes, in videos the key information for event classification is typically centered in the middle of ground-truth event proposals. As a result, the classification module in these existing diffusion models becomes insensitive to boundary changes introduced by the added noise, leading to suboptimal performance. This paper introduces DiffDVC, an innovative diffusion model for DVC. The core of DiffDVC is a boundary-sensitive detector. The detector increases the sensitivity of the classification module to boundary changes by focusing on frames within a specific range around the start and end times of noisy event proposals. Additionally, this range is dynamically adjusted to suit different event proposals. Comprehensive experiments on ActivityNet-1.3, ActivityNet Captions, and YouCook2 datasets show DiffDVC achieving superior performance.

Introduction

Dense video captioning (DVC) focuses on generating descriptions for multiple events in a video, which has garnered significant interest in recent years (Wang et al. 2021; Yang et al. 2023; Zhang, Song, and Jin 2022). DVC is typically divided into two sub-tasks: video event detection and video event captioning. Video event detection involves predicting the temporal boundaries (i.e., start and end times) containing events in a video and generating the corresponding foreground classification confidence. Most existing methods for video event detection in DVC are based on discriminative learning, which relies on regressing anchor proposals (Zhou et al. 2018), predicting the starting and ending probabilities of each frame (Zhang, Wu, and Li 2022; Shi et al. 2023), or learning proposal embedding by learned queries (Wang et al. 2021; Tan et al. 2021; Liu et al. 2022).

Figure 1: An illustration of the comparative study. The red arrows indicate the direction in which object boxes and event proposals are scaled down. The blue dotted lines represent the ground-truth box or proposal (GT), while dotted lines in other colors indicate the scaled-down boxes and proposals.

Recently, several methods have applied generative learning-based models, such as diffusion models (Ho, Jain, and Abbeel 2020) to video event detection (Nag et al. 2023). These models were originally designed for image object detection (Chen et al. 2023a). These methods add noise to the ground-truth event proposal boundaries and subsequently learn a denoising process to reconstruct them. Although the above methods have shown promising results, they do not fully address the differences between videos and images. Specifically, in images, the crucial information for object classification is generally concentrated in regions close to the boundary of the ground-truth box. In contrast, for event detection in videos, the crucial information for event classification is generally concentrated in a region located in the middle of the event proposal.

To statistically demonstrate this observation, we conduct a comparative study on image object detection and video event detection shown in Figure 1. By gradually scaling down ground-truth object boxes in the COCO validation dataset (Lin et al. 2014) and ground-truth event proposals in the THUMOS14 validation dataset (Idrees et al. 2017), we progressively remove boundary information. We then assess the significance of the removed parts by applying the pre-trained detection models, Mask R-CNN (He et al. 2017) and TadTR

*Corresponding Author.

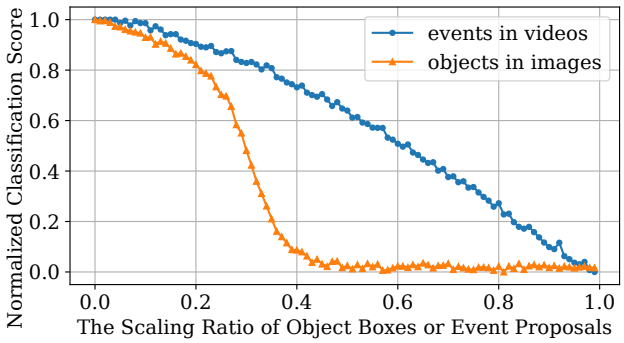

Figure 2: The statistics of the comparative study. The horizontal axis is the scaling ratio. For example, 0.05 means trimming 5% from the boundary of ground-truth boxes or proposals, and scaling their size to 95% of the original. The vertical axis is the normalized classification score.

(Liu et al. 2022) to predict the classification scores for the scaled-down boxes and proposals, respectively. The results are shown in Figure 2, where the horizontal axis is the scaling ratio, and the vertical axis is the normalized classification score. For object classification, the score drops more sharply during early size reductions. At a scaling ratio of 0.2, removing 20% of boundary information significantly reduces the score. By a ratio of 0.4, the score drops 90%, highlighting the importance of boundary regions for classification. For event classification, scores decrease noticeably as the scaling ratio increases to 0.2 and continue to drop until 0.9, where the score drops 90%. This indicates that regions located in the middle of proposals hold critical information, with a sharper decline starting at larger scaling ratios.

As a result, the event classification module becomes insensitive to boundary changes introduced by the added noise. High sensitivity to noise is essential for diffusion models since they learn the added noise to achieve good denoising performance (Ho, Jain, and Abbeel 2020). Therefore, applying diffusion models designed for image object detection to video event detection can lead to suboptimal performance. This raises a natural question: *Does the accuracy of video event detection significantly impact the performance of event captioning?*

We approach this question by shifting the ground-truth event proposals in ActivityNet Captions (Krishna et al. 2017) and produce captions for ground-truth proposals and shifted proposals based on a pre-trained model PDVC (Wang et al. 2021). The captioning results in Table 1 indicate that even minor changes to proposal boundaries greatly impact captioning performance.

Based on the previous discussion, we propose a boundary-sensitive detector for event detection in video. The detector focuses on the frames around the boundaries of noisy proposals to enhance the sensitivity of the classification model to changes at boundaries. Our design faces two key challenges: 1) Identifying the optimal range of frames around the boundaries that require focus, and 2) Deciding how much attention should be allocated to each frame within that range. To address these challenges, the boundary-sensitive detector

Metrics	GT	5%L	5%R	10%L	10%R	15%L	15%R
B ↑	2.98	2.96	2.87	2.37	1.94	2.14	1.77
C ↑	52.38	51.80	51.52	40.52	36.85	36.48	33.93
S ↑	9.53	8.79	8.86	8.16	8.18	7.58	7.54

Table 1: Dense captioning on ground truths and shifted proposals. Evaluation metrics include BELU@4 (B), CIDEr (C), and SODA_c (S). L and R denote the left and right shifts of ground truth, respectively. x% indicates a shift distance equivalent to x% of the length of ground-truth proposals.

utilizes the proposal boundaries (start and end times) as reference points and dynamically adjusts the attention range around them. Subsequently, the detector learns the attention weights of frames within the range and aggregates these frames according to the weights to form proposal features. Using these features, the boundary-sensitive detector predicts temporal boundaries of events and generates the corresponding foreground classification confidence.

Building on the boundary-sensitive detector, we design an innovative diffusion model for DVC, called DiffDVC. DiffDVC utilizes the diffusion model to predict accurate event proposals, resulting in enhanced caption quality.

In summary, our key contributions are as follows:

- One key observation we make is that, whereas in images, important information for classification often resides near the boundaries of the ground-truth boxes, in video event classification tasks, this information typically resides in the middle of the event proposals. This has not been accounted for by previous approaches in adapting the diffusion model from image object detection to video event detection.
- We design a boundary-sensitive detector that focuses on a set of frames around the boundaries of the noisy proposals. This approach enhances the diffusion model's sensitivity to boundary changes, making it more effective for video event detection.
- We introduce an innovative diffusion model designed for dense video captioning, called DiffDVC. DiffDVC achieves superior performance compared to state-of-the-art methods on the ActivityNet-1.3, ActivityNet Captions, and YouCook2 datasets.

Related Work

Video Event Detection

Video event detection, also known as temporal action detection, identifies the start and end times of events in untrimmed videos. Existing methods are primarily based on discriminative learning and fall into three types: anchor-based, boundary-based, and query-based. Anchor-based methods pre-define anchors to classify potential events (Zhou et al. 2018). Boundary-based methods predict classification score for each frame and combine high-confidence frames to form proposals (Zhang, Wu, and Li 2022; Shi et al. 2023; Shao et al. 2023). Query-based methods, inspired by DETR (Carion et al. 2020), use learnable embeddings called action

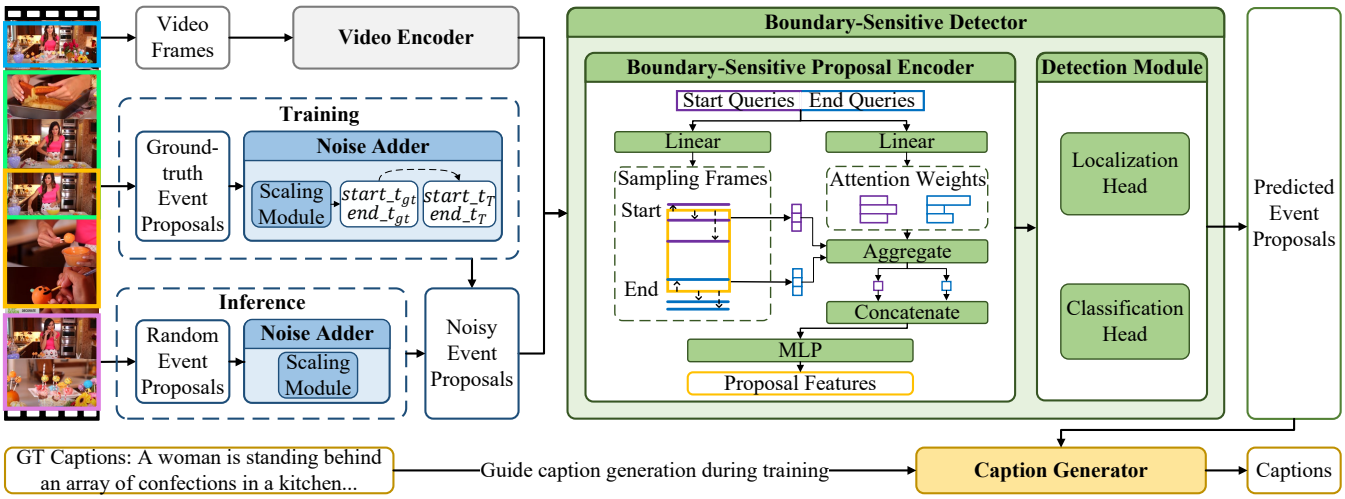

Figure 3: An overview of DiffDVC. DiffDVC comprises four components: a video encoder, a noise adder, a boundary-sensitive detector, and a caption generator. The video encoder extracts multi-level features from an input video. The noise adder adds noise to ground-truth event proposals during training and scales proposals sampled from a Gaussian distribution during inference. The boundary-sensitive detector predicts temporal boundaries (start and end times) and foreground classification scores from the noisy proposals. Finally, the caption generator produces descriptions for the video based on the predicted proposals.

queries to learn event features and decode them into proposals (Liu et al. 2022; Tan et al. 2021; Wang et al. 2021). Recently, a few works explore generative learning-based models, such as diffusion models for event detection (Nag et al. 2023), which are originally developed for image object detection (Chen et al. 2023a). This work considers the differences between images and videos to better adapt diffusion models for video event detection.

Diffusion Model

The diffusion model (Song, Meng, and Ermon 2020) is a powerful class of deep generative models known for its outstanding in various applications. Diffusion models are widely used in various generative tasks in computer vision, including image generation (Ho, Jain, and Abbeel 2020; Wang et al. 2024; Shang et al. 2024) and video generation (Ho et al. 2022; Mei and Patel 2023; Ma et al. 2024). Furthermore, the capabilities of diffusion models are further explored in discriminative tasks, such as image segmentation (Rahman et al. 2023; Chen et al. 2023b) and detection (Nag et al. 2023; Chen et al. 2023a). We address the challenge that event detection methods based on diffusion models (Nag et al. 2023) are insensitive to the noise added to proposals, improving the effectiveness of event detection and further enhancing dense video captioning performance.

Method

Preliminaries

Diffusion Model. Diffusion models are composed of a diffusion process and a reverse process. The diffusion process is mathematically defined as follows (Ho, Jain, and Abbeel 2020):

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \quad (1)$$

which adds Gaussian noise to the sample data x_0, transforming it into a noisy latent sample x_t at time step t within the range $\{1, 2, \ldots, T\}$. Here, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s = \prod_{s=1}^{t}(1-\beta_s)$, where β_s denotes the schedule for noise variance. The reverse process learns to recover the original sample data. During training, a neural network is optimized to predict x_0 based on x_t. During inference, the network reconstructs x_0 from Gaussian noise x_T through an iterative updating rule $x_T \rightarrow x_{T-\Delta} \rightarrow \ldots \rightarrow x_0$ (Ho, Jain, and Abbeel 2020).

We reformulate event detection as a proposal generation task through the diffusion model. The sample data x_0 are M ground-truth event proposals $\{(t_m^s, t_m^e)\}_{m=1}^{M}$, where t_m^s and t_m^e indicate the start time and end time of an event, respectively. The boundary-sensitive detector predicts x_0 from Gaussian noisy proposals x_t, conditioned on the corresponding video. Based on the predicted proposals, we apply a caption generator to generate the corresponding captions.

Overview

The structure of DiffDVC is illustrated in Figure 3. DiffDVC comprises four parts: a video encoder, a noise adder, a boundary-sensitive detector, and a caption generator.

Video Encoder

The purpose of the video encoder is to obtain multi-level features from videos to facilitate event detection and captioning. Initially, video frames are encoded by pre-trained action recognition methods (Wang et al. 2018). Next, Y temporal convolutional layers are employed to extract multi-level frame features. Finally, a deformable transformer (Zhu et al. 2020) is utilized to capture interactions among frames and aggregate multi-level frame features into $F = \{\hat{v}^y\}_{y=1}^{Y}$, where Y represents the total number of feature levels.

Noise Adder

The noise adder builds the diffusion process by converting ground-truth event proposals into noisy proposals during training. It pads GTs with extra ones to form N proposals, denoted $P_{pad} = \{(t_n^s, t_n^e)\}_{n=1}^N$. These proposals are then scaled as follows (Chen et al. 2023a; Nag et al. 2023):

$$P_{pad} = (P_{pad} \times 2 - 1) \times s_p \tag{2}$$

The signal-to-noise ratio, set by the signal scaling factor s_p, affects diffusion model performance (Chen et al. 2023b). As per Eq.(1), Gaussian noise is introduced to the scaled proposals, producing noisy proposals $P_{crpt} = \{(t_n^{cs}, t_n^{ce})\}_{n=1}^N$. During inference, the noise adder samples multiple proposals from a Gaussian distribution and scales them as follows:

$$P_{crpt} = (P_{crpt}/s_p + 1)/2 \tag{3}$$

Boundary-Sensitive Detector

The boundary-sensitive detector is designed to increase the sensitivity of the classification module to changes at the boundaries. It comprises two key components: a boundary-sensitive proposal encoder and a detection module.

Boundary-Sensitive Proposal Encoder. The encoder receives multi-level frame features from the video encoder and noisy proposals P_{crpt} from the noise adder. The encoder attends to a set of sampling frames around the boundaries of noisy event proposals to capture changes at boundaries.

Our design faces two key issues: 1) Identifying the optimal range of frames around the proposal boundaries that require attention. Due to video frame redundancy, the issue is simplified to selecting which frames around the boundaries require attention, rather than focusing on each frame. 2) Deciding the level of attention for each selected frame.

For the first issue, we regard start times $P_s = \{t_n^{cs}\}_{n=1}^N$ and end times $P_e = \{t_n^{ce}\}_{n=1}^N$ of noisy proposals as reference points. We apply $2N$ learnable event queries $\{\hat{e}_n\}_{n=1}^{2N}$ to aggregate features of key frames at the noisy proposal boundaries, with N for start times and N for end times. Each query is linearly projected to predict offsets for K sampling points surrounding each reference point across all feature levels, identifying the frames to attend to. For example, the k-th sampling point p_{nyk}^s around the start time t_n^{cs} of the n-th proposal at the y-th level are obtained as follows:

$$p_{nyk}^s = \phi_y\left(t_n^{cs}\right) + \Delta p_{nyk}^s \tag{4}$$

where ϕ_y maps the reference points to frame features of the y-th level and Δp_{nyk}^s are the predicted offset. Similarly, the k-th sampling point around the end time p_{nyk}^e are obtained.

For the second issue, we apply a linear projection to each event query, generating attention weights for each sampling point. The features of these sampling points are then fused based on the attention weights. For example, the attention feature of the start time of the n-th proposal is calculated by

$$c_n^s = \sum_{y=1}^{Y} \sum_{k=1}^{K} a_{nyk} \hat{v}_{p_{nyk}^s}^y \tag{5}$$

The attention features of the end time c_n^e, are similarly calculated. Lastly, the features of start times $\{c_n^s\}_{n=1}^N$ and the features of end times $\{c_n^e\}_{n=1}^N$ are concatenated and processed by a multi-layer perceptron as proposal features $\{c_n\}_{n=1}^N$. The proposal features serve as the new event queries, and the above operation (in Eq.(4) and Eq.(5)) is repeated 6 times. This operation dynamically adjusts the sampling points or attention ranges as proposal positions change.

Detection Module. The detection module aims to predict the boundaries and classifications of event proposals. The proposal features $\{c_n\}_{n=1}^N$ are scaled and shifted based on time embedding before being passed to the localization and classification heads. The localization head applies a multi-layer perceptron to refine event proposals by adding predicted offsets to the noisy proposals. The classification head applies a linear projection to predict the foreground classification score for proposals. Based on the current predictions, the proposals for the subsequent time step are generated by DDIM (Song, Meng, and Ermon 2020) during inference.

Caption Generator

The caption generator generates captions based on the predicted proposals $\{(\hat{t}_n^s, \hat{t}_n^e)\}_{n=1}^N$ from the boundary-sensitive detector. N learnable event caption queries $\{\hat{s}_n\}_{n=1}^N$ and the predicted proposals are input to deformable DETR (Zhu et al. 2020) for extracting event features $\{\hat{s}_n\}_{n=1}^N$. To predict the i-th word $w_{n,i}$ for the n-th event, an LSTM takes the event features $\hat{s}_n$ and the previous word $w_{n,i-1}$ as input. Finally, we utilize an event counter similar to the one used in PDVC (Wang et al. 2021) to determine the caption number.

Experiments

Experimental Settings

Dataset. We perform experiments using the ActivityNet-1.3 (Caba Heilbron et al. 2015), ActivityNet Captions (Krishna et al. 2017), and YouCook2 (Zhou, Xu, and Corso 2018) datasets. ActivityNet-1.3 has 10,024 training, 4,926 validation, and 5,044 testing videos. ActivityNet Captions includes 10,009 training, 4,917 validation, and 5044 testing videos. YouCook2 contains 1,333 training, 457 validation, and 210 testing videos. Due to the inaccessibility of the testing sets of these datasets, we evaluate DiffDVC using the validation sets following previous methods (Tan et al. 2021; Zhang, Wu, and Li 2022; Shao et al. 2023; Liu et al. 2022; Shi et al. 2023; Tang et al. 2024; Nag et al. 2023; Yang et al. 2023; Zhang, Song, and Jin 2022; Zhou et al. 2018; Mun et al. 2019; Suin and Rajagopalan 2020; Wang et al. 2021).

Evaluation Metrics. We evaluate DiffDVC in two aspects: 1) Event detection: For ActivityNet-1.3, we evaluate the standard mean average precision (mAP) at IOU thresholds [0.3:0.2:0.9]. For ActivityNet Captions and YouCook2, we assess average precision (P), average recall (R), and F1 score at IOU thresholds [0.5:0.05:0.95]. 2) Dense video captioning: We employ the ActivityNet Challenge 2018 official evaluation tool, which includes BLEU@4 (B@4) (Papineni et al. 2002), METEOR (M) (Banerjee and Lavie 2005), and CIDEr (C) (Vedantam, Lawrence Zitnick, and Parikh 2015). And we utilize SODA_c (S) (Fujita et al. 2020), which is designed for evaluating video story descriptions in DVC.

Implementation Details. We adopt the same frame features as those employed by state-of-the-art methods for a

Method	Fea	0.5	0.75	0.95	Avg
Discriminative learning-based models					
RTD-Net	I3D	47.2	30.7	8.6	30.8
Actionformer	I3D	53.5	36.2	8.2	35.6
TadTR	I3D	49.1	32.6	8.5	32.3
ASL	I3D	54.1	37.4	8.0	36.2
TriDet	I3D	54.7	38.0	8.4	36.8
MFAM	I3D	54.7	37.3	8.6	36.6
Generative learning-based models					
DiffTAD	I3D	52.4	35.6	8.8	34.8
DiffDVC	I3D	**53.2**	**36.4**	**9.8**	**35.8**

Table 2: Event detection on the ActivityNet-1.3. 'Fea' refers to the frame features used. 'Avg' means average mAP.

Method	Fea	Recall	Precision	F1
SDVC	C3D	55.58	**57.57**	56.56
PDVC	C3D	55.20	57.36	56.26
DiffDVC	C3D	**57.29**	56.30	**56.79**
MFT	TSN	24.31	51.41	33.01
PDVC	TSN	56.21	**57.46**	56.83
DiffDVC	TSN	**58.20**	56.81	**57.55**
PDVC	I3D	55.79	**57.39**	56.58
DiffDVC	I3D	**57.52**	56.71	**57.11**
Vid2Seq (No pre)	CLIP	-	-	46.5
Vid2Seq (Pre, Multi)	CLIP	-	-	52.4

Table 3: Event detection on the ActivityNet Captions. 'Pre' refers to methods using pretraining. 'Multi' indicates training on multi-modal inputs.

fair comparison. For ActivityNet-1.3, we utilize I3D features (Zhang, Wu, and Li 2022). For ActivityNet Captions, we adopt C3D, TSN, and I3D+VGGish (I3D) features (Wang et al. 2021). For YouCook2, we utilize TSN features (Wang et al. 2021; Zhou et al. 2018). For ActivityNet-1.3 and ActivityNet Captions, we configure the number of event proposals or queries N to be 15, while for YouCook2, N is set to 100. During inference, the sample steps in DDIM are set to 1. We train word embeddings with 512 dimensions from scratch. The signal scaling factor is 1.0. We apply the Adam optimizer and the learning rate is initialized to 5e-5.

Comparison with State-of-the-art Methods

We compare DiffDVC with state-of-the-art event detection methods and dense video captioning methods.

Event Detection Performance. We compare DiffDVC with event detection methods, including RTD-Net (Tan et al. 2021), Actionformer (Zhang, Wu, and Li 2022), ASL(Shao et al. 2023), TadTR (Liu et al. 2022), TriDet (Shi et al. 2023), MFAM (Tang et al. 2024), and DiffTAD (Nag et al. 2023). Results on the ActivityNet-1.3 are shown in Table 2. Compared to discriminative learning-based methods, DiffDVC does not achieve the best performance but outperforms methods applying queries for detection, i.e., RTD-Net (Tan et al. 2021) and TadTR (Liu et al. 2022). Most importantly, DiffDVC performs better in all metrics than DiffTAD (Nag et al. 2023), which also employs a diffusion model.

Method	Fea	Recall	Precision	F1
Vid2Seq (No pre)	CLIP	-	-	15.4
Vid2Seq (Pre, Multi)	CLIP	27.9	27.8	27.8
PDVC *	TSN	24.33	33.78	28.29
DiffDVC	TSN	**28.60**	**34.37**	**31.22**

Table 4: Event detection on the YouCook2. $*$ denotes the calculation using the official codebase.

Method	Fea	B@4	M	C	S
DCE	C3D	0.17	5.69	12.43	-
SDVC	C3D	-	6.92	-	-
DVC	C3D	0.73	6.93	12.61	-
Efficient	C3D	1.35	6.21	13.82	-
ECHR	C3D	1.29	7.19	14.71	3.22
PDVC	C3D	**1.65**	**7.50**	25.87	5.26
DiffDVC	C3D	1.57	7.15	**28.20**	**5.71**
MT *	TSN	1.15	4.98	9.25	-
PDVC	TSN	1.78	**7.96**	28.96	5.44
DiffDVC	TSN	**1.85**	7.72	**29.90**	**6.01**
PDVC	I3D	**1.96**	**8.08**	28.59	5.42
DiffDVC	I3D	1.88	7.74	**30.61**	**6.91**
Vid2Seq (No pre)	CLIP	-	-	14.2	5.4
Vid2Seq (Pre, Multi)	CLIP	-	8.5	30.1	5.8

Table 5: Comparison of dense video captioning results using the ActivityNet Captions.

We also compare DiffDVC with dense video captioning methods on the ActivityNet Captions, including SDVC (Mun et al. 2019), PDVC (Wang et al. 2021), MFT (Xiong, Dai, and Lin 2018), and Vid2Seq (Yang et al. 2023). The results in Table 3 show that DiffDVC achieves higher recall and F1 scores on all features. For example, on three features, the recall scores of DiffDVC are 2.09, 1.99 and 1.73 higher respectively compared with the best method PDVC. Notably, DiffDVC outperforms Vid2Seq, which utilizes large-scale pretraining and multimodal inputs (transcribed speech and video). Furthermore, Table 4 shows DiffDVC achieving the best results with significant gains on the YouCook2.

Dense Video Captioning Performance. We compare DiffDVC with SOTA methods trained with cross-entropy loss, including DCE (Krishna et al. 2017), SDVC (Mun et al. 2019), DVC (Li et al. 2018), Efficient (Suin and Rajagopalan 2020), ECHR (Wang et al. 2020), PDVC (Wang et al. 2021), MT (Zhou et al. 2018), and Vid2Seq (No pre) (Yang et al. 2023). Results on the ActivityNet Captions are presented in Table 5. DiffDVC achieves better results on most metrics. Notably, DiffDVC significantly exceeds all methods on both the CIDEr and SODA_c score based on the proposals predicted by the boundary-sensitive detector. CIDEr is shown to align closely with human consensus when evaluating the overall quality of captions. On the CIDEr score, DiffDVC shows relative improvements of 9.01%, 3.25%, and 7.07% over state-of-the-art scores using C3D, TSN, and I3D features, respectively. On the SODA_c score, DiffDVC achieves

Method	Fea	B@4	M	C	S
MT	TSN	0.30	3.18	6.10	-
ECHR	TSN	-	3.82	-	-
PDVC	TSN	**0.80**	**4.74**	22.71	4.42
DiffDVC	TSN	0.71	4.29	**23.70**	**5.34**
Vid2Seq (No pre)	CLIP	-	-	15.6	3.0

Table 6: Comparison of DVC using the YouCook2.

Detector	Fea	Recall	Precision	F1
vanilla	C3D	56.19	55.10	55.64
b-sensitive	C3D	**57.29**	**56.30**	**56.79**
vanilla	TSN	57.32	54.88	56.07
b-sensitive	TSN	**58.20**	**56.81**	**57.55**
vanilla	I3D	57.05	55.11	56.06
b-sensitive	I3D	**57.52**	**56.71**	**57.11**

Table 7: Ablation study on the ActivityNet Captions to prove the importance of the boundary-sensitive detector (b-sensitive) for event detection.

relative improvements of 8.56%, 10.48%, and 27.49% respectively over the state-of-the-art scores based on three features. Furthermore, DiffDVC performs better than the pretraining method Vid2Seq. Table 6 shows that DiffDVC achieves better performance on the YouCook2 validation set for CIDEr and SODA_c. The results align with our observation that the accuracy of video event detection greatly affects the performance of dense video captioning. Moreover, this impact is more pronounced for CIDEr and SODA_c.

Ablation Studies

To explore DiffDVC in detail, we conduct ablation studies using the ActivityNet Captions and YouCook2 datasets, with 15 and 100 proposals, respectively, except when examining the effects of proposal numbers.

Boundary-Sensitive Detector. To evaluate the effectiveness of the boundary-sensitive detector (b-sensitive), we replace this module with a vanilla detector (vanilla) that does not focus on proposal boundaries. As shown in Table 7, when special attention is paid to the boundaries of noisy proposals, detection performance improves. Consequently, the model with the boundary-sensitive detector achieves better results on all dense video captioning metrics presented in Table 8. For example, when employing the boundary-sensitive detector based on C3D features, recall and precision scores increase by 1.1 and 1.2. Correspondingly, CIDEr and SODA_c scores increase by 1.36 and 0.49. Similarly, the ablation study results on YouCook2 presented in Table 9 indicate the importance of the boundary-sensitive detector for event detection and dense video captioning. In conclusion, focusing on frames around the boundaries of noisy proposals enhances event detection, which in turn leads to better video captions.

Detector	Fea	B@4	M	C	S
vanilla	C3D	1.45	6.61	26.84	5.22
b-sensitive	C3D	**1.57**	**7.15**	**28.20**	**5.71**
vanilla	TSN	1.77	7.40	29.09	5.93
b-sensitive	TSN	**1.85**	**7.72**	**29.90**	**6.01**
vanilla	I3D	1.79	7.53	30.30	6.75
b-sensitive	I3D	**1.88**	**7.74**	**30.61**	**6.91**

Table 8: Ablation study on the ActivityNet Captions to prove the importance of the boundary-sensitive detector for DVC.

Detector	R	P	F1	B@4	M	C	S
vanilla	28.52	33.14	30.66	0.63	**4.33**	23.51	5.08
b-sensitive	**28.60**	**34.37**	**31.22**	**0.71**	4.29	**23.70**	**5.34**

Table 9: Ablation study on the YouCook2 to prove the importance of the boundary-sensitive detector.

Detector	Factor	Recall	Precision	F1
b-sensitive	0.5	58.30	**56.81**	**57.55**
b-sensitive	1.0	57.52	56.71	57.11
b-sensitive	1.5	57.51	56.40	56.95
b-sensitive	2.0	**58.38**	56.20	57.27
b-sensitive	3.0	57.78	55.68	56.71
vanilla	0.5	**57.53**	**56.01**	**56.76**
vanilla	1.0	57.05	55.11	56.06
vanilla	1.5	56.82	55.50	56.15
vanilla	2.0	56.81	54.78	55.78
vanilla	3.0	56.80	53.94	55.33

Table 10: Comparison of event detection results for models with different signal scaling factors using I3D features on the ActivityNet Captions.

Signal Scaling Factor. To further prove the effectiveness of the boundary-sensitive detector, we perform experiments with different signal scaling factors. The factor s_p (in Eq.(2) and Eq.(3)) determines the signal-to-noise ratio during the denoising process. A larger factor results in smaller boundary changes caused by added noise, and vice versa. Results are presented in Table 10. Notably, as the factor increases, detecting boundary changes becomes progressively more difficult. DiffDVC with the boundary-sensitive detector exhibits a smaller decline in scores, compared to DiffDVC with the vanilla detector. This suggests that the boundary-sensitive detector can sensitively capture subtle changes at the boundaries caused by added noise. For example, when the factor increases from 0.5 to 3.0, DiffDVC with the boundary-sensitive detector sees a precision drop of 1.13, compared to a 2.07 drop with the vanilla detector. Moreover, a smaller factor enhances detection performance, because appropriately reducing the factor makes boundary changes more pronounced and easier to detect.

Number of Sampling Points. We examine the influence of the number of sampling points K at the boundaries on the performance of DiffDVC with a boundary-sensitive detector. As illustrated in Table 11, DiffDVC achieves optimal performance when K is 4.

K	Fea	R	P	F1	B@4	M	C	S
1	C3D	56.80	55.86	56.33	1.56	7.08	27.19	5.60
2	C3D	56.97	55.94	56.45	1.57	7.05	28.07	5.67
4	C3D	**57.29**	**56.30**	**56.79**	**1.57**	**7.15**	**28.20**	**5.71**
8	C3D	57.15	55.70	56.42	1.55	7.06	27.82	5.63
16	C3D	57.27	55.39	56.31	1.54	7.11	28.07	5.69
1	TSN	56.27	55.36	55.81	1.78	7.40	29.09	5.74
2	TSN	56.81	56.55	56.68	1.74	7.55	28.43	5.79
4	TSN	**58.20**	**56.81**	**57.50**	**1.85**	**7.72**	**29.90**	**6.01**
8	TSN	57.69	56.34	57.01	1.74	7.49	28.89	5.77
16	TSN	57.58	56.57	57.07	1.71	7.67	30.03	5.93
1	I3D	57.34	56.45	56.89	1.85	7.73	30.00	6.73
2	I3D	57.48	56.60	57.04	1.73	7.57	30.31	6.84
4	I3D	**57.52**	**56.71**	**57.11**	**1.88**	**7.74**	**30.61**	**6.91**
8	I3D	57.05	56.19	56.62	1.77	7.71	30.29	6.75
16	I3D	56.96	55.68	56.31	1.83	7.62	30.18	6.67

Table 11: Results using TSN features on the ActivityNet Captions with different numbers of sampling points (K).

Proposals	Fea	B@4	M	C	S
15	C3D	1.57	7.15	28.20	**5.71**
30	C3D	1.64	7.13	28.79	5.64
50	C3D	1.67	6.97	29.31	5.51
100	C3D	**1.68**	**7.17**	**29.76**	5.66
15	TSN	1.85	7.72	29.90	6.01
30	TSN	1.83	7.57	30.07	5.87
50	TSN	**1.86**	**7.79**	32.26	5.98
100	TSN	1.78	7.71	**32.28**	**6.15**
15	I3D	1.88	7.74	30.61	**6.91**
30	I3D	1.85	7.78	30.79	6.67
50	I3D	**2.00**	**7.88**	31.10	6.47
100	I3D	1.82	7.61	**31.54**	6.66

Table 12: Dense video captioning results with varying numbers of proposals (N) using the ActivityNet Captions.

Number of Proposals. We evaluate the impact of different numbers of proposals N (15, 30, 50, or 100) on DVC, keeping them consistent during training and inference. According to Table 12, higher proposal numbers improve the scores on the CIDEr. For instance, increasing the number from 15 to 100 improves CIDEr scores by 1.56, 2.38, and 0.93 for C3D, TSN, and I3D features, respectively.

Qualitative Examples

In Figure 4, we present examples that visually compare the dense video captioning results of DiffDVC and the best-performing PDVC. The example demonstrates that DiffDVC accurately predicts event boundaries and provides detailed captions. For video 2, DiffDVC accurately predicts the boundaries of all two events and provides accurate descriptions of the belly dancing.

Conclusion

In this work, we introduce a new diffusion model designed for dense video captioning, named DiffDVC. At the heart of DiffDVC is a boundary-sensitive detector, which is designed to increase the sensitivity of the classification module to boundary changes. The detector focuses on frames within a specific range around the start and end times of noisy event proposals, with the range dynamically adjusted to suit different event proposals. The proposed method considers that the distribution of important information for classification differs between image objects and video events, making the diffusion model more applicable to video event detection. Experiments on three datasets demonstrate that the boundary-sensitive detector predicts more accurate event proposal boundaries. With the help of the detector, DiffDVC exhibits improved performance in caption generation.

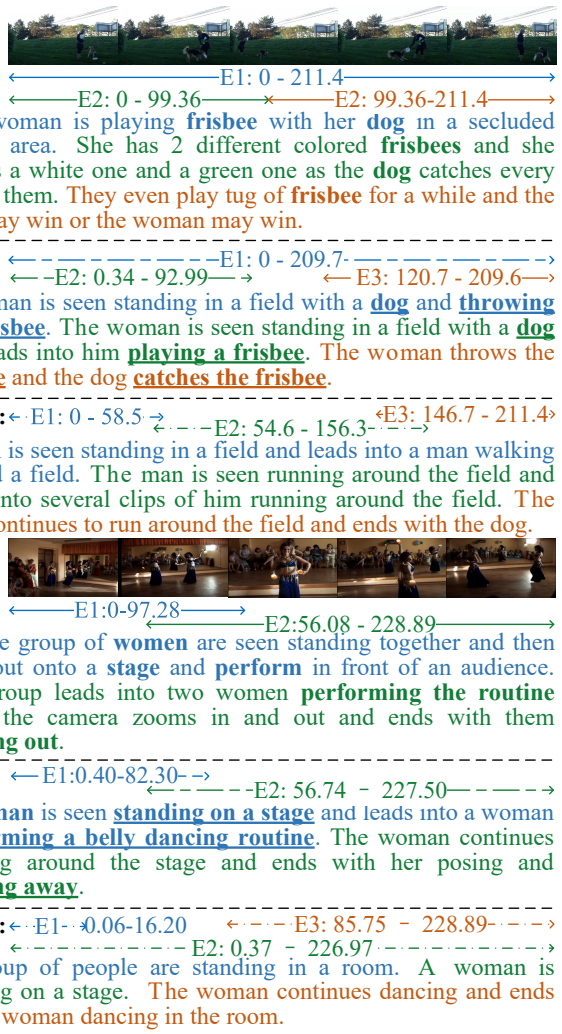

Figure 4: Examples of dense video captioning on ActivityNet Captions. Arrows and captions of the same color correspond to the same events (E), where the numbers inside the arrows signify the start and end times. Underlined words indicate correctly predicted information.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62372027, the National Key R&D Program of China under Grant No. 2023YFB4503700, and the National Natural Science Foundation of China under Grant No. 62372028.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–970.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023a. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19830–19843.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2023b. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 909–919.

Fujita, S.; Hirao, T.; Kamigaito, H.; Okumura, M.; and Nagata, M. 2020. SODA: Story Oriented Dense Video Captioning Evaluation Framework. In *European Conference on Computer Vision*, 517–531.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Idrees, H.; Zamir, A. R.; Jiang, Y.-G.; Gorban, A.; Laptev, I.; Sukthankar, R.; and Shah, M. 2017. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155: 1–23.

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, 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, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Mei, K.; and Patel, V. 2023. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9117–9125.

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.

Nag, S.; Zhu, X.; Deng, J.; Song, Y.-Z.; and Xiang, T. 2023. Difftad: Temporal action detection with proposal denoising diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10362–10374.

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.

Rahman, A.; Valanarasu, J. M. J.; Hacihaliloglu, I.; and Patel, V. M. 2023. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11536–11546.

Shang, S.; Shan, Z.; Liu, G.; Wang, L.; Wang, X.; Zhang, Z.; and Zhang, J. 2024. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8975–8983.

Shao, J.; Wang, X.; Quan, R.; Zheng, J.; Yang, J.; and Yang, Y. 2023. Action sensitivity learning for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13457–13469.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; and Tao, D. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18857–18866.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Suin, M.; and Rajagopalan, A. 2020. An efficient framework for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12039–12046.

Tan, J.; Tang, J.; Wang, L.; and Wu, G. 2021. Relaxed transformer decoders for direct action proposal generation. In

Proceedings of the IEEE/CVF international conference on computer vision, 13526–13535.

Tang, Y.; Wang, W.; Zhang, C.; Liu, J.; and Zhao, Y. 2024. Learnable Feature Augmentation Framework for Temporal Action Localization. *IEEE Transactions on Image Processing*.

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, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11): 2740–2755.

Wang, R.; Chen, Z.; Chen, C.; Ma, J.; Lu, H.; and Lin, X. 2024. Compositional text-to-image synthesis with attention map control of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5544–5552.

Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; and Luo, P. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6847–6857.

Wang, T.; Zheng, H.; Yu, M.; Tian, Q.; and Hu, H. 2020. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5): 1890–1900.

Xiong, Y.; Dai, B.; and Lin, D. 2018. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 468–483.

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.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhang, Q.; Song, Y.; and Jin, Q. 2022. Unifying event detection and captioning as sequence generation via pre-training. In *European Conference on Computer Vision*, 363–379. Springer.

Zhou, L.; Xu, C.; and Corso, J. J. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8739–8748.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.