DIBS: Enhancing Dense Video Captioning with Unlabeled Videos via Pseudo Boundary Enrichment and Online Refinement

Hao Wu[1,2] ♣, Huabin Liu[2,3] ♣, Yu Qiao[2], Xiao Sun[2] ✉
[1] University of Science and Technology of China [2] Shanghai Artificial Intelligence Laboratory
[3] Shanghai Jiao Tong University
wuhao@mail.ustc.edu.cn, huabinliu@sjtu.edu.cn, {qiaoyu, sunxiao}@pjlab.org.cn

Abstract

*We present **D**ive **I**nto the **B**oundarie**S** (**DIBS**), a novel pretraining framework for dense video captioning (DVC), that elaborates on improving the quality of the generated event captions and their associated pseudo event boundaries from unlabeled videos. By leveraging the capabilities of diverse large language models (LLMs), we generate rich DVC-oriented caption candidates and optimize the corresponding pseudo boundaries under several meticulously designed objectives, considering diversity, event-centricity, temporal ordering, and coherence. Moreover, we further introduce a novel online boundary refinement strategy that iteratively improves the quality of pseudo boundaries during training. Comprehensive experiments have been conducted to examine the effectiveness of the proposed technique components. By leveraging a substantial amount of unlabeled video data, such as HowTo100M [16], we achieve a remarkable advancement on standard DVC datasets like YouCook2 [31] and ActivityNet [13]. We outperform the previous state-of-the-art Vid2Seq [27] across a majority of metrics, achieving this with just 0.4% of the unlabeled video data used for pre-training by Vid2Seq.*

1. Introduction

Dense video captioning (DVC), a challenging task in video understanding, involves the temporal localization and captioning of all events within an untrimmed video [12]. Compared to standard video captioning that generates a single caption for a short video clip [4, 26, 30], the complexity of dense captioning significantly increases as it requires localizing multiple events in long-term video sequences and much more detailed captioning.

Particularly, the event boundary is paramount in dense video captioning. It provides precise event localization, ensuring the generated captions are accurate, coherent, and contextually relevant. Unfortunately, data containing precise event boundaries is rare and expensive to annotate. This scarcity poses a substantial performance bottleneck.

Pioneering efforts have been made to address data shortage challenges in DVC. Notably, several weakly supervised approaches [3, 6] have endeavored to approximate fully supervised performance without relying on all the boundary annotations provided by existing datasets, thereby circumventing the need for such annotations. However, these methods take a conservative approach to the problem by carefully re-designing the DVC training or testing framework, aiming to theoretically develop self-sufficient techniques and reduce the reliance on precise boundary annotations. They have not yet substantially incorporated larger-scale data for training purposes to drive improved performance. Instead, we take a direct approach to address the fundamental data scarcity issues. Specifically, we introduce an effective method for generating and enhancing pseudo boundaries and harnessing the capabilities of LLMs to produce higher-quality coherent captions. Given the enhanced captions, their corresponding boundaries are generated and further optimized using a carefully designed unified metric and optimization algorithm, ultimately achieving optimal quality. Moreover, we deploy this approach on a substantial volume of unannotated, large-scale video data, effectively bolstering the training data for the DVC task and thereby resulting in a notable performance improvement.

Similar to our motivation, Vid2Seq [27] also emphasizes the utilization of large-scale unlabeled video data for model training to boost performance. They first propose a single-stage DVC framework that collectively predicts all event captions and corresponding temporal boundaries by generating a single sequence of discrete tokens. Then, this framework is pre-trained on numerous unlabeled narrated videos, where the target event captions and boundaries are obtained directly from the text and timestamps of subtitles

♣ Co-first authors. Work done as interns at Shanghai AI Laboratory.
✉ Corresponding author.

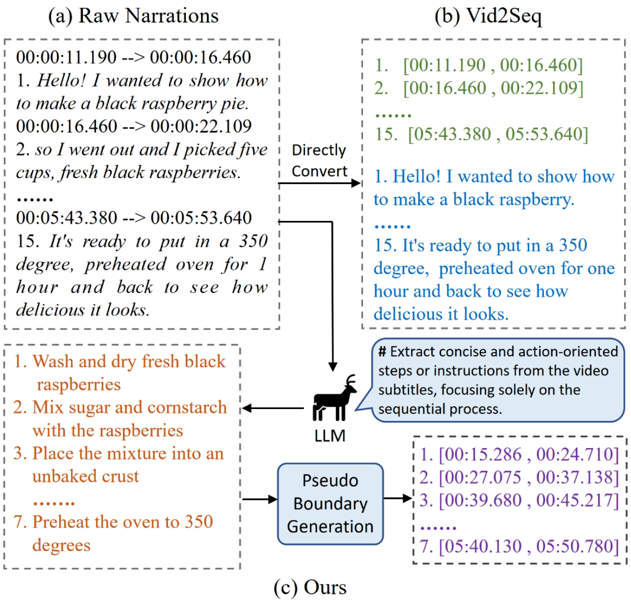
(a) Raw Narrations
00:00:11.190 --> 00:00:16.460
1. Hello! I wanted to show how to make a black raspberry pie.
00:00:16.460 --> 00:00:22.109
2. so I went out and I picked five cups, fresh black raspberries.
......
00:05:43.380 --> 00:05:53.640
15. It's ready to put in a 350 degree, preheated oven for 1 hour and back to see how delicious it looks.
Directly Convert
(b) Vid2Seq
1. [00:11.190 , 00:16.460]
2. [00:16.460 , 00:22.109]
......
15. [05:43.380 , 05:53.640]
1. Hello! I wanted to show how to make a black raspberry.
......
15. It's ready to put in a 350 degree, preheated oven for one hour and back to see how delicious it looks.
# Extract concise and action-oriented steps or instructions from the video subtitles, focusing solely on the sequential process.
LLM
1. Wash and dry fresh black raspberries
2. Mix sugar and cornstarch with the raspberries
3. Place the mixture into an unbaked crust
....
7. Preheat the oven to 350 degrees
Pseudo Boundary Generation
1. [00:15.286 , 00:24.710]
2. [00:27.075 , 00:37.138]
3. [00:39.680 , 00:45.217]
......
7. [05:40.130 , 05:50.780]
(c) Ours

Figure 1. Comparison between Vid2Seq [27] and ours. (a) Raw subtitles extracted from videos. (b) Vid2Seq [27] directly converts the raw text and timestamps of subtitles into pseudo event boundaries and captions for pretraining. (c) Our proposed framework utilizes LLMs to generate rich and accurate captions for events from raw narrations. Subsequently, corresponding pseudo boundaries can be generated using these captions through our devised optimization algorithm (cf. Sec. 3.2).

(as shown in Figure 1(b)). Nevertheless, these raw subtitles usually comprise dialogues, personal reflections, or superfluous background details, and are often temporally misaligned with the visual stream, which introduces significant noise into the event learning. Therefore, while Vid2Seq successfully integrates a larger scale of data for DVC, the effectiveness and utilization efficiency of its large-scale video data still need to be improved.

To this end, we present **DIBS**, a novel DVC framework that generates accurate event captions along with pseudo boundaries from unlabeled videos to enable effective pretraining for DVC. Specifically, DIBS harnesses diverse off-the-shelf LLMs, leveraging their proficiency in text processing to produce coherent and rich caption candidates for events from unlabeled videos. Moreover, it generates corresponding pseudo event boundaries for each caption, and further optimizes each boundary with a meticulously designed algorithm considering multiple metrics such as diversity, event-centricity, temporal ordering, and coherence. Furthermore, considering that the generated pseudo boundary is still imperfect and includes noises (e.g., background segments), we propose a novel training strategy with online boundary refinement. This strategy aims to iteratively refine and improve the pseudo boundaries during the training phase. Additionally, we seamlessly integrate our approach with state-of-the-art DVC training frameworks such as PDVC [25], achieving significantly improved results on extensive benchmarks.

2. Related Work

2.1. Dense Video Captioning

Dense video captioning entails event localization and captioning tasks in long-form videos. Early approaches to this problem typically employed a two-stage ”detect-then-describe” framework. Within this framework, previous methods [10, 12, 17, 28] focused extensively on improving event representation. For instance, HCN [28] observed that contextual modeling could significantly enhance event captioning performance, while [10, 11] incorporated the audio modality to produce a more robust representation.

Traditionally, the separation between event localization and captioning posed a gap in methods. However, recent approaches [5, 14, 27, 32, 33] strive for joint learning. MT [32] connected captioning loss and proposal boundaries via a differential masking mechanism for their mutual optimization. PDVC [25], inspired by DETR [2], framed the task as set prediction, enabling simultaneous optimization of both tasks. In contrast, SGR [5] proposed a top-down framework, generating paragraphs before assigning event descriptions to video segments. E2ESG [33] tackled dense video captioning as a unified sequence-to-sequence task using a multimodal Transformer, predicting event locations and captions concurrently.

Despite their advancements, current DVC algorithms still rely on comprehensive annotations for events, particularly precise event boundaries. This requirement restricts the utilization of large-scale datasets to enhance DVC performance. Recently, Vid2Seq [27] made a groundbreaking attempt by leveraging unlabeled narrated videos for DVC pre-training. However, it directly converts the text and timestamps from narrations into pseudo event captions and boundaries for training, which introduces considerable noise (e.g., backgrounds) into the learning targets. This operation limits the potential benefits of utilizing extensive narrated video datasets for DVC.

2.2. Weakly-supervised Dense Video Captioning

Research in weakly-supervised dense video captioning has gained traction due to its potential to bypass the tedious task of annotating precise event boundaries in lengthy videos. WSDEV [6] introduced a cyclical system tackling caption generation and sentence localization as dual tasks. WLT [20] extended this by incorporating audio inputs for improved event captioning. Recently, EC-SL [3] introduced a concept learner to enhance the sentence localizer. However, these methods primarily focus on reducing reliance on precise boundary annotations without incorporating unlabeled data for training or significantly improving performance compared to fully supervised approaches.

2.3. Video-Language Pretraining

Large-scale video-text pretraining has proven highly effective in diverse video applications like video retrieval, recognition, and non-dense video captioning. UniVL [15], for instance, employed multi-task pretraining on an instructional dataset for video retrieval. All-in-one [24] extended this approach with an end-to-end video-language model, incorporating multiple video datasets to support various downstream tasks like video question answering.

However, few studies have focused on extensive pretraining for dense video captioning due to the demanding nature of event annotation in these tasks. UEDVC [29] introduced a pretraining task on ActivityNet to enhance DVC performance on the same dataset. Vid2Seq [27] collected and used numerous narrated videos for pretraining in dense video captioning. However, it directly translates raw narrations and timestamps into pseudo event captions and boundaries for DVC pretraining, leading to significant noise and reduced effectiveness in utilizing large-scale narration video data. In contrast, our approach introduces a new pipeline to extract rich and accurate event captions and pseudo boundaries, preserving the wealth of information in large-scale videos for DVC.

3. Approach

Given an input video $\mathbf{V}$ comprising M frames $\mathbf{f}$, DVC aims to temporally localize and describe events within the video using natural language. In particular, DVC must predict timestamped boundaries $\mathbf{b}$ around key moments and generate descriptive captions $\mathbf{c}$ for each segmented event, resulting in a triplet $(\{\mathbf{f}_m|m \in M\}, \{\mathbf{b}_n|n \in N\}, \{\mathbf{c}_n|n \in N\})$, where N represents the number of events.

Therefore, large-scale pretraining for DVC requires an extensive collection of video data paired with textual descriptions like subtitles or speech-to-text transcriptions, represented as $(\mathbf{V}, \mathbf{C})$. However, this data lacks precise boundary information crucial for determining event quantity, positions, and durations. Additionally, the textual descriptions may not consistently align with the requirements of dense captioning, needing a coherent, sequentially narrated description based on pivotal events.

Previous methods, like Vid2Seq [27], tackled this issue by segmenting subtitles into boundaries and captions. However, these subtitles often contain dialogues, musings, reflections, or irrelevant details, leading to discrepancies between the intended content and the subtitles. This mismatch resulted in inaccuracies in automatically generated captions, misrepresenting the depicted events in the video.

Method	Vid2Seq [27]	Ours
N	#Sentences in subtitle	#Events reinterpreted by diverse LLMs
$\mathbf{c}$	Noisy raw sentences in subtitle	Rich, event-centric and sequential
$\mathbf{b}$	Timestamps of subtitle	Our pseudo boundaries generator in Sec. 3.2

Table 1. Comparison of event caption generation methods between Vid2Seq [27] and our approach in terms of '**b**', '**c**', and 'N'.

3.1. Prompting LLMs for DVC-Oriented Captions

To address the aforementioned issue more effectively, we aim to utilize the capabilities of off-the-shelf LLMs [21, 22] to enrich $\mathbf{C}$ for the DVC task. LLMs are recognized for their proficiency in text processing, showcasing a remarkable capacity to generate rich and contextually accurate captions, particularly when provided with carefully crafted prompts.

Diverse Off-the-shelf LLMs Our exploration into diverse LLMs includes both open-source models like LLAMA-2 [22] and InternLM [21], as well as closed-source API-based models such as ChatGPT and Claude. Leveraging diverse LLMs enables us to evaluate different backbones and datasets, tapping into the distinct capabilities of each model. Tailoring carefully selected prompts to the unique characteristics of each LLM, whether open-source or API-based, enables us to extract accurate and contextually relevant event captions. Integrating a diverse range of LLMs illustrates the adaptability and effectiveness of our methodology, highlighting its versatility across different LLMs and ensuring the extraction of high-quality event captions.

Leveraging Sparse Ground Truth Captions as Prompts for Prompt Generation Initially, we aim to extract the event events descriptions $\{\mathbf{c}_n|n \in N\}$ from the subtitles $\mathbf{C}$. To achieve this, we employ a circle-prompting strategy, initially providing a small set of ground truth event captions as hints and querying an LLM for prompts that can generate similar results. We subsequently conduct iterative testing of prompts and captions for better results and manually correct LLM errors in the loop to ensure that our prompts generate rich, accurate, concise, coherent, and event-centric captions. This iterative process seeks to achieve a balanced compromise between precision and conciseness. An example prompt we generate is specified as follows:

Task: Extract concise and action-oriented steps or instructions from the video subtitles, focusing solely on the sequential process. Each step should be presented as a single sentence with clear actions. Exclude any steps that are not directly related to the actions in the video. Generate the steps directly without repeating the original text.

Consequently, the original subtitles are transformed into logically structured and temporally coherent description of events. Table 1 and Figure 1 present a comparison between our method and Vid2Seq [27] in generating event captions, accompanied by an example. Utilizing various LLMs and prompts, we generate multiple candidate event captions. These candidates, introduced in the next section during the pseudo boundary generation process, are optimized with boundary generation.

3.2. Optimization of Pseudo Boundaries

After generating event captions $\{\mathbf{c}_n | n \in N\}$, the subsequent challenge is to ascertain the temporal boundaries $\{\mathbf{b}_n | n \in N\}$ of the corresponding events. In this section, we introduce how to obtain and optimize the corresponding boundaries for each event caption. The optimization comprises two main objectives: first, maximizing alignment between the event caption $\mathbf{c}_n$ and the video clip $\mathbf{f}_{[\mathbf{b}_n]}$; and second, ensuring that the temporal order relationships between the boundaries reflect those between event captions.

Vision-Language Similarity Matrix For semantic alignment between $\{\mathbf{c}_n | n \in N\}$ and $\{\mathbf{f}_m | m \in M\}$, we adopt a bottom-up optimization strategy. Initially, a pre-trained vision-language (VL) model, denoted as (M_V, M_L) for vision and language feature extractors, is employed to calculate similarities between individual frames in the video and each caption. This yields a similarity matrix $\mathbf{S}$ that signifies the associations between frames and captions.

$$\mathbf{S}_{m,n} = \frac{M_V(\mathbf{f}_m) \cdot M_L(\mathbf{c}_n)}{|M_V(\mathbf{f}_m)| \cdot |M_L(\mathbf{c}_n)|} \tag{1}$$

Note that when employing a video-language model, a short video clip centered around a frame is utilized to represent this frame. For implementation, image-language models such as CLIP [18] and video-language models like UniVL [15] are utilized. In the experiments, we aggregate scores from multiple vision-language models, averaging them to yield a more robust similarity matrix.

Discussion: the domain gap and noisy detection issue. Evidently, the quality of the similarity matrix significantly depends on the VL models' quality and the domain gap between the dataset used for VL model training and the DVC dataset. On the one hand, this underscores the importance of employing diverse LLMs to generate rich caption candidates. On the other hand, despite these concerted efforts, the similarity matrix still exhibits notable detection noise and, in some instances, false positives.

Caption-Aware Pseudo Boundary Generation with Soft Time Constraints Leveraging the similarity matrix $\mathbf{S}$, we convert the event localization problem into the identification of the optimal N frames in $\{\mathbf{f}_m | m \in M\}$ corresponding to captions $\{\mathbf{c}_n | n \in N\}$. This process should adhere to the temporal order specified in $\{\mathbf{c}_n | n \in N\}$, resulting in a drop dynamic time warping (Drop-DTW) [7] problem. We denote this baseline method as the **Drop-DTW Baseline**.

However, due to the noisy detection issue discussed above, the effectiveness of the direct Drop-DTW Baseline method is constrained. This ineffectiveness arises from several reasons: 1) **Multimodal Responses:** Higher response positions often exhibit multiple peaks, with their distribution in the video lacking concentration. 2) **Non-Uniform Event Coverage:** Events do not exhibit a preference for a uniformly distributed coverage across the majority of the entire video. 3) **Strict Temporal Ordering Constraints:** the hard global temporal ordering constraints in Drop-DTW not only slow down the optimization process significantly but also impede individual captions from effectively finding the best-matching video intervals; 4) **Duration Determination:** Additionally, determining the duration of events lacks a straightforward, intuitive method.

Algorithm 1 Pseudo Boundary Generation

```
Input: event captions $\{\mathbf{c}_n | n \in N\}$; similarity matrix $\mathbf{S}$; user-defined top-$k$ parameter $K$, total iterations $Q$.
Initialization: $\{\mathbf{b}_n^0 | n \in N\} \leftarrow$ Divide the entire video $\mathbf{V}$ into $N$ equal segments.
for each caption $\mathbf{c}_n$ in $\{\mathbf{c}_n | n \in N\}$ do
    for each iteration $t$ in $t \in Q$ do
        collect top-$\hat{k}$ frames $\{T_{\hat{k},n} | \hat{k} \in K\}$ around $\mathbf{b}_n^t$.
        for each $\hat{k}$ in $K$ do
            $D_{\hat{k},n} = \sum_{i=1}^{K} |T_{i,n} - T_{\hat{k},n}|$
        end for
        $\hat{k}^* \leftarrow \hat{k}$ with minimum $D_{\hat{k},n}$.
        new boundary center $\leftarrow T_{\hat{k}*,n}$
        new boundary size $\leftarrow 2 * \alpha * std_n$.
        update boundary $\leftarrow \mathbf{b}_n^t$ in Eq. 3
        record loss $\leftarrow L^t$ in Eq. 4.
    end for
    select iter with minimum loss $q = \underset{t}{\operatorname{argmin}} L^t, t \in Q$
    determine the pseudo boundary for caption $\mathbf{c}_n$: $\mathbf{b}_n \leftarrow \mathbf{b}_n^q$.
    determine the loss for captions : $L \leftarrow L^q$.
end for
Output: Generated pseudo boundaries $\{\mathbf{b}_n | n \in N\}$ and loss $L$ for input event captions $\{\mathbf{c}_n | n \in N\}$.
```

To address these issues, we propose a caption-aware pseudo boundary generation algorithm with soft time constraints, as illustrated in Algorithm 1. To encourage events cover most of the video length *evenly*, we initialize the boundaries $\mathbf{b}^0$ by dividing the video into N equal segments. Subsequently, we iteratively optimize the boundaries, as illustrated in Figure 2. For the nth caption, in its tth iteration, we first collect a *local* set of top-$\hat{k}$ frames around $\mathbf{b}_n^{t-1}$ (the boundary range in the previous iteration, starting from $\mathbf{b}_n^0$) with the highest similarity scores to $\mathbf{c}_n$. From these $\hat{k}$ positions, we select the new boundary center. For each of the $\hat{k}$ frames, we calculate the total temporal distance from this frame to the other $\hat{k} - 1$ frames. The frame with the minimum total distance indicates that these top $\hat{k}$ frames are more concentrated around it, thus setting it as the new boundary center, demoted as $T_{\hat{k}^*,n}$. The size of the boundary is determined by calculating the standard deviation with respect to the new boundary center as shown in Equation 2.

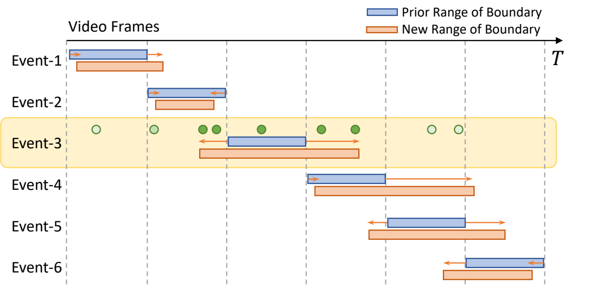

Figure 2. Prior range division and frame processing in pseudo boundary generation. Frames showcasing top-k global similarity to Event-3 are indicated by deep green points, contrasting with lighter green points denoting similarity outside the top-k range. The blue box delineates the evenly divided prior range, while the orange box signifies the adjusted range after an iteration. The pseudo boundary generation process is outlined in Algorithm 1.

$$std_n = (\frac{1}{k}\sum_{k}(T_{k,n} - T_{\hat{k}^*,n})^2)^{\frac{1}{2}} \tag{2}$$

To this end, we gain a coarse boundary $\mathbf{b}_{coarse,n}$ and we further determine the actual boundary $\mathbf{b}_n^t$ of the boundary by selecting the minimum and maximum values of k frames within the new coarse boundary as depicted in Equation 3:

$$\begin{aligned} \mathbf{b}_{coarse,n} &= [T_{\hat{k}^*,n} - \alpha * std_n, T_{\hat{k}^*,n} + \alpha * std_n] \\ \mathbf{b}_n^t &= [\min(T_{k,n}), \max(T_{k,n}) \text{ for } T_{k,n} \text{ in } \mathbf{b}_{coarse,n}] \end{aligned} \tag{3}$$

where α is a hyperparameter determining the size of the coarse boundary. After each iteration, we define a loss function to evaluate the quality of the generated boundary, as shown in Equation 4:

$$\begin{aligned} L^t &= \sum_n \sum_k \mathbf{S}_{k,n} \cdot Dis(T_{k,n}, \mathbf{b}_n^t), \\ Dis(f, \mathbf{b}) &= \begin{cases} -\min(f - \mathbf{b}_s, \mathbf{b}_e - f), & \text{if } \mathbf{b}_s < f < \mathbf{b}_e, \\ \max(\mathbf{b}_s - f, f - \mathbf{b}_e), & \text{if } f < \mathbf{b}_s \text{ or } f > \mathbf{b}_e. \end{cases} \end{aligned} \tag{4}$$

Here, the loss is the total distance weighted by the similarity score between the frame and caption. The function $Dis(\cdot)$ measures the distance between a frame f and the boundary $\mathbf{b}$, where $\mathbf{b}_s$ and $\mathbf{b}_e$ represent the start and end of the boundary respectively. The function $Dis(f, \mathbf{b})$ is positive when f is within the boundary range and negative otherwise. We then collect the local top-k frames with the highest similarity scores to the nth caption around the new boundary and proceed to the next iteration. Finally, we select the boundary with the minimum loss value as the final optimized boundary. Refer to Algorithm 1 for further details on generating the pseudo boundaries.

3.3. Training with Online Boundary Refinement

Although we can employ various fully supervised approaches [25, 29, 33] to train a DVC model using generated event captions and pseudo boundaries, these boundaries might be imperfect, encompassing inaccuracies like background elements. Utilizing them directly as ground truth during DVC pretraining could mislead the event learning process, especially in event localization. To counter this, we introduce an online strategy for refining pseudo boundaries, seamlessly integrating it with cutting-edge DVC training frameworks like PDVC [25].

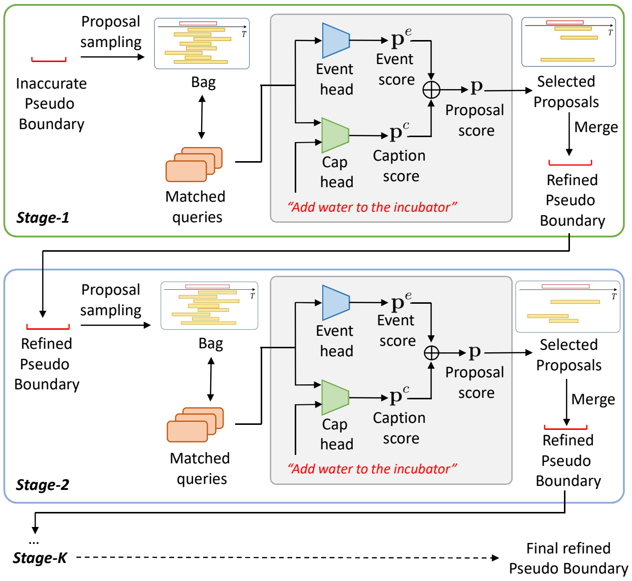

Figure 3. The process of online pseudo boundary refinement.

Background Review of PDVC PDVC [25] is an effective DVC framework employing parallel decoding and set prediction principles. It starts by utilizing a pre-trained video feature extractor and a transformer encoder to derive a sequence of frame-level features. Then, employing M learnable event queries $\{\mathbf{q}_i\}_{i=1}^M$, it employs a transformer decoder along with three prediction heads (localization head, caption head, and event counter) to simultaneously predict M boundaries, M captions, and event count. During inference, the model selects the top detected events by ranking captioning and localization scores, without using non-maximum suppression (NMS) [9].

Online Pseudo Boundary Refinement We refine pseudo boundaries by augmenting their quantity and conducting quality evaluations. As shown in Figure 3, consider a pseudo boundary $\mathbf{b} = (t, d)$ of caption $\mathbf{c}$, we employ a standard jitter augmentation on the boundary duration d and the boundary center t with jitter ratio r_1 and r_2 respectively, obtaining an augmented set of proposals, denoted as $\mathcal{B} = \{\hat{\mathbf{b}}_u\}_{u=1}^U$ (U proposals in total and $\mathbf{b}$ is contained in $\mathcal{B}$ as well), representing potentially superior event segments for caption $\mathbf{c}$. Then, following PDVC, we adopt the Hungarian matching method [2] between all proposals $\{\hat{\mathbf{b}}_u\}_{u=1}^U$ and query embeddings $\{\mathbf{q}_i\}_{i=1}^M$ to link each proposal with a specific query, yielding $\{\mathbf{Q}_u\}_{u=1}^U$ ($\mathbf{Q}_u = \mathbf{q}_i$, if the ith query is linked to the uth proposal). The linked query serves as the proxy feature for the proposal.

Utilizing the augmented set of proposals and their proxy features, we evaluate their quality online during training. PDVC offers an event classification head $\mathbf{h}^e$ and a caption scoring head $\mathbf{h}^c$, both using the query embeddings as input and producing confidence scores. Formally,

$$\mathbf{p}_u^e = \texttt{Sigmoid}(\mathbf{h}^e(\mathbf{Q}_u)) \tag{5}$$

$$\mathbf{p}_u^c = \texttt{Softmax}_{\{U\}}(\mathbf{h}^c(\mathbf{Q}_u, \mathbf{c})) \tag{6}$$

Note that the event score $\mathbf{p}_u^e$ is normalized with the Sigmoid activation while the caption score $\mathbf{p}_u^c$ is normalized with the Softmax activation across all proposals in $\mathcal{B}$. The final assessment score $\mathbf{p}_u$ for the proposal $\hat{\mathbf{b}}_u$ is calculated as $\mathbf{p}_u = \mathbf{p}_u^e + \mathbf{p}_u^c$. A higher proposal score means this proposal better represents the boundary of caption $\mathbf{c}$.

Finally, we select the top-K proposals in $\{\hat{\mathbf{b}}_u\}_{u=1}^U$ with the highest scores and compute a weighted average of their boundaries as the final refined boundary.

$$\mathbf{b}_{ref} = \frac{\sum_{k=1}^K \mathbf{p}_k \cdot \hat{\mathbf{b}}_k}{\sum_{k=1}^K \mathbf{p}_k} \tag{7}$$

The refined pseudo boundary $\mathbf{b}_{ref}$ replaces the original boundary $\mathbf{b}$ for subsequent training stages. This iterative process, shown in Figure 3, refines each pseudo boundary in multiple stages, resulting in increasingly accurate boundaries. Our implementation defaults to a 2-stage refinement process to derive the final boundary. Importantly, boundary refinement doesn't impact model inference, incurring no additional computational overhead during this phase.

4. Experiments

4.1. Experimental Setting

Datasets Our experiments encompass two prominent datasets commonly employed for dense video captioning: YouCook2 [31] and ActivityNet Captions [13]. For pretraining, we leverage a subset of the HowTo100M [16] dataset, specifically focusing on cooking videos, amounting to approximately 56,000 videos.

Implementation Details In our setup, we uniformly sample video frames at 1 FPS and adjust them to a fixed count denoted as F. For YouCook2, F is set to 200, and for ActivityNet Captions, it's 100. We utilize pretrained vision-language models like CLIP [18] and UniVL [15] to extract frame-level features across all datasets. The model undergoes a two-stage training process: initially, a 10-epoch pretraining on a subset of HowTo100M, followed by a 20-epoch fine-tuning phase on each target dataset. Notably, to address the domain gap between the HowTo100M dataset and the target dataset, we augment the training data by incorporating the target dataset using pseudo-boundaries during pretraining. Our model architecture mirrors PDVC [25], incorporating a transformer encoder, transformer decoder, and three prediction heads.

Evaluation Metrics We employ standard captioning metrics: METEOR [1] (M) for semantic similarity, and CIDEr [23] (C) for human judgment correlation. For event localization evaluation, we utilize average precision (Pre.) and recall (Rec.) at Intersection over Union(IoU) thresholds of 0.3, 0.5, 0.7 and 0.9, along with the overall F1 score for boundary prediction. These metrics are computed using the official ActivityNet challenge toolbox. Additionally, we incorporate the SODA_c [8] (S) metric, which provides a joint evaluation of caption quality and localization accuracy.

4.2. Comparison with State-of-the-art Methods

In Table 2, we compare our captioning performance against state-of-the-art approaches. We evaluate our model under two settings: weakly supervised and fully supervised paradigms. In the weakly supervised setting, our model is trained directly on the target dataset using ground truth captions and generated pseudo boundaries, without using any ground truth boundaries. It can be observed that this strategy achieves comparable performance compared to prior weakly supervised DVC methods, demonstrating the effectiveness of our pseudo boundary generation and refinement. Regarding the standard fully-supervised setting, our approach outperforms PDVC [25] with the same backbone architecture after pretraining. This illustrates the significant benefits of pretraining for DVC. Particularly, on YouCook2 and ActivityNet, our models with UniVL and CLIP backbones surpass previous methods across multiple metrics. Notably, Vid2Seq incorporates additional speech cues during inference, providing a richer multi-modal context. Besides, Vid2Seq uses much larger-scale pretraining data (15M videos) and captioning models (T5 [19]) compared to our DIBS. These factors likely explain some performance differences relative to Vid2Seq, particularly with the CLIP backbone. However, using the UniVL backbone, our model can surpass Vid2Seq on YouCook2 by a significant margin. On ActivityNet, our CLIP-based approach also achieves state-of-the-art results.

In Table 3, we see that our pretraining consistently enhances the event localization, enabling our DIBS to outperform previous methods on YouCook2 by a great margin. However, on ActivityNet, our approach lags behind PDVC and UEDVC [29] while outperforming Vid2Seq. This disparity may be due to the domain gap between instructional and activity videos, impacting the localization of less common events despite improvements in captioning.

Training	Method	Settings		YouCook2			ActivityNet		
		Pretrain	Backbone	M	C	S	M	C	S
Weakly-Supervised	WLT [20]	∅	TSN	-	-	-	4.93	13.79	-
	WS-DEC [6][†]	∅	C3D	-	-	-	6.30	18.77	-
	EC-SL [3][†]	∅	C3D	-	-	-	7.49	21.21	-
	DIBS (Ours)	∅	CLIP	4.63	22.05	4.03	7.33	15.86	4.67
	DIBS (Ours)	∅	UniVL	5.90	29.62	4.96	6.76	13.75	4.26
Fully-Supervised	PDVC [25]	∅	TSN	4.74	22.71	4.42	7.96	28.96	5.44
	PDVC [25][†]	∅	CLIP	5.47	28.37	5.00	8.31	30.11	5.63
	PDVC [25][†]	∅	UniVL	7.87	46.02	6.87	8.24	28.21	5.43
	UEDVC [29]	676k	TSN	2.18	8.37	3.34	-	-	5.49
	E2ESG [33]	∅	C3D	3.49	25.00	-	-	-	-
	Vid2Seq [27]	15M	CLIP	9.30	47.10	7.90	8.50	30.10	5.80
	DIBS (Ours)	56k	CLIP	7.51	44.44	6.39	**8.93**	**31.89**	**5.85**
	DIBS (Ours)	56k	UniVL	**9.41**	**59.35**	**7.97**	8.25	28.85	5.35

Table 2. Performance of caption generation on YouCook2 and ActivityNet. [†] Results are obtained from our implementation with official codebase.

Method	Settings		YouCook2		ActivityNet	
	Pretrain	Backbone	Rec.	Pre.	Rec.	Pre.
PDVC [25]	∅	TSN	-	-	55.42	58.07
PDVC [25][†]	∅	CLIP	21.76	31.92	50.82	55.73
PDVC [25][†]	∅	UniVL	29.66	42.04	53.21	59.46
UEDVC [29]	676k	TSN	-	-	**59.00**	**60.32**
E2ESG [33]	∅	C3D	3.49	25.00	-	-
Vid2Seq [27]	15M	CLIP	27.90	27.80	52.70	53.90
DIBS (Ours)	56k	CLIP	26.24	39.18	53.14	58.31
DIBS (Ours)	56k	UniVL	**30.80**	**45.13**	53.02	58.39

Table 3. Performance of event localization on YouCook2 and ActivityNet. [†] Results are obtained from our implementation with official codebase.

In summary, our proposed pretraining approach substantially boosts event localization on well-matched instructional datasets like YouCook2. Regarding general activity localization, additional techniques may be necessary to transfer specific knowledge from the instructional videos due to the clear domain gap. Therefore, taking careful consideration of domain similarity is critical when leveraging unlabeled videos for DVC pretraining.

4.3. Ablation Study

Effects of Pseudo Boundary We examine the impact of pseudo boundaries on YouCook2, comparing against a baseline where ground truth event boundaries are omitted. When utilizing pseudo boundaries, matching between proposals and captions primarily relies on the Generalized IoU (GIoU) cost. To provide a comprehensive comparison, we construct two additional settings, both incorporating caption information similar to our pseudo boundary. The first setting introduces an additional caption cost for matching, replacing the GIoU cost, while the second setting incorporates a caption-proposal similarity cost in place of the GIoU cost. In Table 4, we present a comprehensive comparison of model performance on event localization and caption generation using CLIP and UniVL as backbones. The results demonstrate a significant improvement in model performance when employing pseudo boundaries, whereas the other two settings exhibit inferior performance compared to the baseline at most metrics. This corroborates the effectiveness of generating pseudo boundaries using captions as a powerful method for leveraging caption information in event localization tasks.

Effects of Soft Time Constraints We assess the impact of soft time constraints on pseudo boundaries by excluding ground truth boundaries from both the YouCook2 and ActivityNet. Our method without soft time constraints performs global video iteration for each caption. Table 5 delineates model performance with and without soft time constraints, employing CLIP and UniVL backbones. Introducing soft time constraints notably enhances model performance in event localization and caption generation. This improvement underscores the efficacy of soft time constraints in generating higher-quality pseudo boundaries. The sequential nature of events in videos highlights the significance of maintaining temporal order, even in scenarios with potential event overlap, as observed in ActivityNet.

	Settings			Metrics				
	Feature	Boundary	Caption	M	C	S	Rec.	Pre.
1.	CLIP	✗	✗	2.32	9.07	3.40	18.47	25.01
2.	CLIP	✗	Cost	1.68	7.09	2.79	14.19	16.63
3.	CLIP	✗	Sim	2.18	8.37	3.34	**20.77**	24.70
4.	CLIP	✓	✗	**4.62**	**21.93**	**3.73**	15.48	**28.82**
1.	UniVL	✗	✗	2.66	9.97	3.21	15.62	25.18
2.	UniVL	✗	Cost	1.99	8.49	3.15	18.75	21.08
3.	UniVL	✗	Sim	2.27	9.16	3.16	15.83	24.88
4.	UniVL	✓	✗	**5.88**	**28.04**	**4.47**	**19.72**	**35.43**

Table 4. Performance comparison of event localization and caption generation on YouCook2 with and without pseudo boundaries.

Dataset	Features	STC	M	C	S	Rec.	Pre.
YouCook2	CLIP	✗	3.92	17.05	3.57	14.29	25.87
	CLIP	✓	**4.62**	**21.93**	**3.73**	**15.48**	**28.82**
	UniVL	✗	5.43	25.57	4.47	18.05	32.83
	UniVL	✓	**5.88**	**28.04**	**4.47**	**19.72**	**35.43**
ActivityNet	CLIP	✗	6.60	15.23	3.87	35.31	57.68
	CLIP	✓	**7.23**	**16.63**	**4.68**	**42.44**	**65.46**
	UniVL	✗	5.88	13.60	3.68	35.72	55.14
	UniVL	✓	**6.75**	**14.13**	**4.21**	**42.24**	**64.75**

Table 5. Comparison of model performance with and without soft time constraints. STC denotes soft time constraints.

Effects of Boundary Refinement To assess the impact of boundary refinement, we employ two distinct settings on the YouCook2 and ActivityNet Caption datasets. In the first setting, we examine performance by omitting the ground truth boundary, utilizing pseudo boundaries both with and without refinement. In the second setting, we perform pretrain-

Dataset	Pretrain	Refine	M	C	S	Rec.	Pre.
YouCook2	✗	✗	5.88	28.04	4.47	19.72	35.43
	✗	✓	**5.90**	**29.62**	**4.96**	**30.80**	**45.13**
	✓	✗	9.04	54.98	**8.13**	**31.55**	44.95
	✓	✓	**9.41**	**59.35**	7.97	30.80	**45.13**
ActivityNet	✗	✗	7.23	**16.63**	**4.68**	**42.44**	65.46
	✗	✓	**7.33**	15.86	4.67	42.21	**65.79**
	✓	✗	**8.94**	30.49	5.39	51.45	57.49
	✓	✓	8.93	**31.89**	**5.85**	**53.14**	**58.31**

Table 6. Comparison between models with and without pseudo boundary refinement.

Dataset	Pretrain	Features	M	C	S	Rec.	Pre.
YouCook2	✗	CLIP	5.47	28.37	5.00	21.76	31.92
	✓	CLIP	7.51	44.44	6.39	26.24	39.18
	✗	UniVL	7.87	46.02	6.87	29.66	42.04
	✓	UniVL	**9.41**	**59.35**	**7.97**	**30.80**	**45.13**
ActivityNet	✗	CLIP	8.31	30.11	5.63	50.82	55.73
	✓	CLIP	**8.93**	**31.89**	**5.85**	53.14	58.31
	✗	UniVL	8.24	28.21	5.43	**53.21**	**59.46**
	✓	UniVL	8.25	28.85	5.35	53.02	58.39

Table 7. Comparative analysis of model performance with and without pretraining on YouCook2 and ActivityNet Datasets.

ing on the HowTo100M subset with and without refinement, followed by fine-tuning the pretrained models on target datasets like YouCook2 and ActivityNet within our study. In Table 6, we present a comprehensive model performance comparison with and without refinement. Refinement enhances model performance on localization and caption metrics across both YouCook2 and ActivityNet datasets. However, while most metrics improve, a few experience a slight drop, and this discrepancy varies between the YouCook2 and ActivityNet datasets. We attribute this observation to the domain gap between the two datasets.

Effect of Pretraining To examine pretraining effects, we conducted a comprehensive comparison of model performance with and without pretraining using CLIP and UniVL backbones on YouCook2 and ActivityNet datasets. Table 7 details performance for event localization and caption generation. It indicates improved caption generation on both datasets after pretraining, yet stable improvements in event localization are not consistently observed, especially on ActivityNet. The enrichment of pseudo captions from large-scale unlabeled videos, generated by LLMs, likely contributes to enhanced caption performance. Despite employing diverse strategies to improve the quality of pseudo boundaries, the substantial distance from ground truth boundaries persists, possibly explaining the limited impact of pretraining on event localization. We also examined the impact of varying pretraining data amounts, please refer to our supplementary materials.

Dataset	Pretrain	FT Data	M	C	S	Rec.	Pre.	F1
YouCook2	✗	100%	7.87	46.02	6.87	29.66	42.04	34.78
	✓	25%	7.81	46.69	7.17	28.93	41.08	33.95
	✓	50%	8.60	55.73	7.84	30.14	42.73	35.34
	✓	75%	9.11	59.09	7.86	29.97	44.54	35.83
	✓	100%	**9.41**	**59.35**	**7.97**	**30.80**	**45.13**	**36.61**
ActivityNet	✗	100%	8.31	30.11	5.63	50.82	55.73	53.16
	✓	25%	8.66	28.24	5.01	51.08	56.24	53.54
	✓	50%	8.56	30.09	5.39	51.96	56.26	54.02
	✓	75%	8.83	30.80	5.55	52.90	57.19	54.96
	✓	100%	**8.93**	**31.89**	**5.85**	**53.14**	**58.31**	**55.61**

Table 8. Few-shot performance fine-tuning on partial data. "FT data" represents the percentage of data used for fine-tuning.

Few-shot Dense Video Captioning In Table 8, we depict the model's performance concerning varying proportions of fine-tuning data on YouCook2 and ActivityNet datasets. The results show a direct association with the amount of fine-tuning data used. Notably, even with a fraction of the complete training set post-pretraining, the model surpasses the performance of the setting without pretraining and using the entire training set. This distinction is particularly prominent on YouCook2, where achieving superior performance with only half of the target training set is feasible. Conversely, ActivityNet requires more training data to achieve comparable advancements.

Backbone Influence in Dense Video Captioning Table 7 compares CLIP and UniVL features, both with and without pretraining. UniVL demonstrates better performance on YouCook2, possibly due to its pretraining on instructional videos, narrowing the domain gap. In contrast, CLIP excels on ActivityNet, leveraging its strong generalization abilities. Our findings emphasize the importance of selecting backbones pretrained on tasks aligned with the dataset. UniVL suits instructional videos like YouCook2, while CLIP's broader generalization is advantageous for diverse datasets like ActivityNet, highlighting the need for a tailored approach based on dataset characteristics and backbone pretraining specifics in dense video captioning tasks.

5. Conclusion

We introduce **DIBS**, a novel pretraining framework for DVC that improves the quality of pseudo event boundaries and captions derived from large-scale unlabeled videos. Leveraging LLMs, we generate and optimize pseudo boundaries and captions simultaneously, emphasizing objectives like diversity and coherence. We also propose an online boundary refinement strategy to further improve the quality of pseudo boundaries. Extensive experiments validate the effectiveness of DIBS.

Acknowledgments This work is partially supported by the National Key R&D Program of China (NO.2022ZD0160104).

References

[1] Satanjeev Banerjee and Alon Lavie. 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, 2005. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 5

[3] Shaoxiang Chen and Yu-Gang Jiang. 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*, pages 8425–8435, 2021. 1, 2, 7

[4] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 358–373, 2018. 1

[5] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2021. 2

[6] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 2, 7

[7] Mikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan Jepson. Drop-dtw: Aligning common signal between sequences while dropping outliers. *Advances in Neural Information Processing Systems*, 34:13782–13793, 2021. 4

[8] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 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*, pages 517–531. Springer, 2020. 6

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 5

[10] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*, 2020. 2

[11] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020. 2

[12] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1, 2

[13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 1, 6

[14] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7492–7500, 2018. 2

[15] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3, 4, 6

[16] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1, 6

[17] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6588–6597, 2019. 2

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 6

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6

[20] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8908–8917, 2019. 2, 7

[21] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`, 2023. 3

[22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[23] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[24] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 3

[25] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning

with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021. 2, 5, 6, 7

[26] Ning Xu, An-An Liu, Yongkang Wong, Yongdong Zhang, Weizhi Nie, Yuting Su, and Mohan Kankanhalli. Dual-stream recurrent neural network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2482–2493, 2018. 1

[27] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 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*, pages 10714–10726, 2023. 1, 2, 3, 7

[28] Dali Yang and Chun Yuan. Hierarchical context encoding for events captioning in videos. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1288–1292. IEEE, 2018. 2

[29] Qi Zhang, Yuqing Song, and Qin Jin. Unifying event detection and captioning as sequence generation via pre-training. In *European Conference on Computer Vision*, pages 363–379. Springer, 2022. 3, 5, 6, 7

[30] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020. 1

[31] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. 1, 6

[32] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748, 2018. 2

[33] Wanrong Zhu, Bo Pang, Ashish V Thapliyal, William Yang Wang, and Radu Soricut. End-to-end dense video captioning as sequence generation. *arXiv preprint arXiv:2204.08121*, 2022. 2, 5, 7