SAIL: Similarity-Aware Guidance and Inter-Caption Augmentation-based Learning for Weakly-Supervised Dense Video Captioning

Anonymous CVPR submission

Paper ID 5504

Abstract

Weakly-Supervised Dense Video Captioning aims to localize and describe events in videos trained only on caption annotations, without temporal boundaries. Prior work introduced an implicit supervision paradigm based on Gaussian masking and complementary captioning. However, existing method focuses merely on generating non-overlapping masks without considering their semantic relationship to corresponding events, resulting in simplistic, uniformly distributed masks that fail to capture semantically meaningful regions. Moreover, relying solely on ground-truth captions leads to sub-optimal performance due to the inherent sparsity of existing datasets. In this work, we propose SAIL, which constructs semantically-aware masks through cross-modal alignment. Our similarity-aware training objective guides masks to emphasize video regions with high similarity to their corresponding event captions. Furthermore, to guide more accurate mask generation under sparse annotation settings, we introduce an LLM-based augmentation strategy that generates synthetic captions to provide additional alignment signals. These synthetic captions are incorporated through an inter-mask mechanism, providing auxiliary guidance for precise temporal localization without degrading the main objective. Experiments on ActivityNet Captions and YouCook2 demonstrate state-of-the-art performance on both captioning and localization metrics.

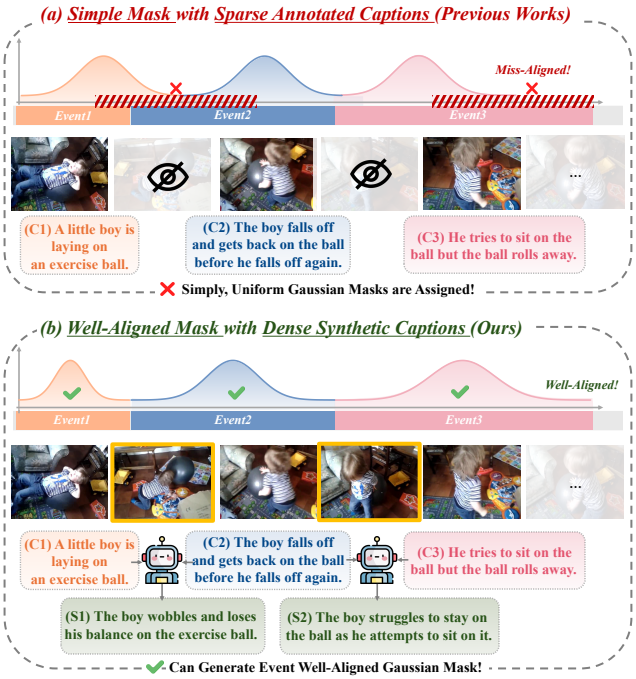
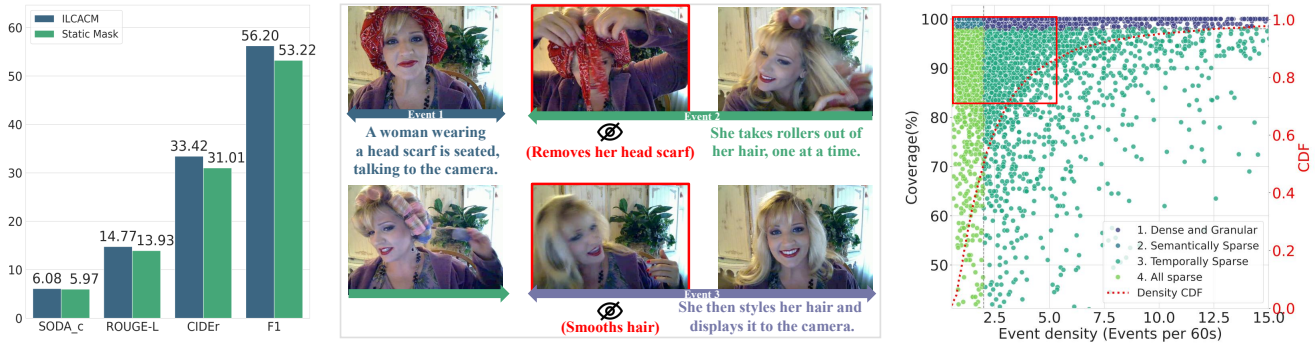


Figure 1. (a) Previous work generates masks that simply cover different temporal regions without considering semantic alignment with corresponding events. (b) Our proposed method leverages cross-modal similarity to guide masks toward event-relevant regions and addresses caption sparsity through LLM augmentation.

1. Introduction

In diverse multimodal applications [3, 17, 20], Dense Video Captioning (DVC) is one of the key research area in the video domain. Unlike conventional video captioning [5, 10, 34], DVC requires jointly localizing events and generating captions in long, untrimmed videos. Although diverse DVC methods have been proposed, research has mainly focused on fully-supervised settings [13–15, 41]. These methods depend on exhaustively labeled data with precise event boundaries and descriptive captions for each event, making annotation prohibitively expensive and labor-intensive for real-world applications. To address these issues, several Weakly-Supervised Dense Video Captioning (WSDVC) studies have emerged [7, 8, 11, 38], which do not rely on event boundaries and only use captions for training.

Recently, ILCACM [11], the current state-of-the-art WSDVC method, introduces a Gaussian masking strategy that applies differentiable masks to video features, enabling implicit event localization through complementary captioning. Although it shows promising results, we observe that this



(a) (b) (c)

Figure 2. (a) The fixed mask baseline applies Gaussian masks with equal width uniformly distributed according to the number of events. It performs comparably to the existing method [11]. (b) shows the dataset's annotation sparsity example, where potential events are missed. (c) The majority of video samples have events spanning the entire duration but contain only a small number of events (red box).

method overlooks the semantic alignment between masked features and their corresponding events, which we believe is a critical aspect for learning discriminative event representations. As shown in Figure 1, when constructing masks for target events, the existing approach merely focuses on generating mutually exclusive masks without ensuring semantic consistency with corresponding events, often resulting in simplistic and uniformly distributed masks. These *simplistic and uniformly* distributed masks lead the model to fail to learn meaningful event emphasis, resulting in low-quality captions and inaccurate localization. As shown in Figure 2 (a), we find that even a fixed, non-trainable mask baseline performs comparably to the existing method [11] on both captioning and localization tasks. This indicates that the existing approach merely learns to emphasize distinct temporal regions without capturing semantic event relevance.

To address this issue, our primary goal is to construct genuinely informative masks for target events. We define an informative mask as one that, rather than simply covering a distinct temporal region, highlights video segments whose features exhibit high cross-modal similarity with the corresponding event caption. To achieve this, we leverage the strong cross-modal alignment of Vision-Language model CLIP [32] and guide the model to generate masks based on cross-modal similarity signals rather than enforcing only non-overlap constraints. We train the model to align masked video features with corresponding ground-truth captions while distinguishing non-corresponding ones. Despite the strong overall performance of the cross-modal similarity guide, we observe some failure cases when videos have very sparse annotations (e.g., only 1-2 events). We hypothesize that these failures stem from the sparse characteristic of ground-truth annotations [24], which leads to insufficient signal for the model to effectively learn the alignment between masked features and their corresponding captions. For example, as illustrated in Figure 2(b), a sample video from ActivityNet [22] spans 235 seconds yet contains only three event annotations, leaving numerous potential events unannotated. Also, as shown in Figure 2(c), although the majority of videos have annotations covering their entire duration, the event density remains consistently low, demonstrating that most videos show sparse event distributions. In such sparse annotation regimes, we conjecture that over-reliance on only ground-truth caption guidance provides insufficient supervision for learning accurate mask-event alignment, especially in WSDVC.

To overcome these limitations, we introduce a Large-Language Model (LLM) based caption augmentation strategy that leverages recent advances in LLMs' contextual understanding and reasoning capabilities [1, 37]. Specifically, we construct prompts using consecutive pairs of ground-truth captions and task the LLM with generating plausible descriptions for the temporal segments connecting them. These synthetic captions are then interleaved with the original annotations, creating a dense supervision signal that captures a more diverse range of potential events. By training on both real and synthetic captions, our cross-modal objective learns more nuanced similarity guidance, enabling masks to capture finer-grained event details. To this end, we propose *SAIL*, **S**imilarity-**A**ware Guidance and **I**nter-Caption Augmentation-based **L**earning for WSDVC. Our contributions are as follows:

- We propose similarity-aware mask guide, which leverages the cross-modal alignment information to generate event-focused masks.
- We propose a novel method that generates synthetic captions via LLM and utilizes them as auxiliary supervision to address annotation sparsity, enabling the model to learn from dense alignment signals.
- We validate our method on ActivityNet and YouCook2, achieving state-of-the-art results in both the localization and captioning tasks.

2. Related Work

2.1. Dense Video Captioning

DVC is a challenging multi-task that requires the simultaneous execution of event localization and event captioning in untrimmed videos. In the early stages of DVC [12, 22, 35], many studies adopted a "localize-then-caption" strategy, where events are first localized and then a caption is generated for each event. However, this 2-stage approach overlooked the interaction between the localization and captioning tasks; also, they rely on complex, hand-crafted components and extensive hyper-parameter tuning. To overcome these limitations, DETR-based transformer architecture [13–15, 41] emerged to decode both event timestamps and captions in parallel. Yang et al. [41] proposed a method of pre-training on a large-scale narrated video dataset using ASR and then fine-tuning. Based on the above pre-trained model, [13–15, 41] enhances captioning ability by utilizing an external memory of captions. Despite extensive research into various DVC methods, most studies still operate in fully supervised settings. This is a notable limitation, given the extremely high annotation cost associated with localizing and captioning events in video data.

2.2. Weakly-Supervised Dense Video Captioning

WSDVC aims to perform the DVC task without this temporal supervision. Duan et al. [8] first addressed WSDVC by decomposing the problem into sentence localization and event captioning. Building upon this, subsequent research has largely revolved around cycle training systems [6]. Among these, Chen and Jiang [4] introduced a concept learner approach and Choi et al. [7] further enhanced performance by leveraging the MSR-VTT dataset [40] for pre-training. Recently, Ge et al. [11] has proposed a method that implicitly performs localization and captioning through complementary masked captioning. Although this complementary masked captioning has proven effective, its masking strategy is simplistic, as it merely ensures that the masks cover different regions. To address this issue, we propose a method to ensure that the masked region is well-aligned with its corresponding caption.

3. Method

In WSDVC, given a video $V = \{v_i\}_{i=1}^{N_v}$ with N_v frames that encompass N_s distinct events, the goal is to generate corresponding captions $\{C_i\}_{i=1}^{N_s}$ for each event and localize them temporally. As WSDVC is trained only with caption annotations, the model learns to infer temporal event boundaries by aligning video frames with their corresponding textual descriptions [7, 8, 11].

Our goal is to construct informative masks that emphasize the visual features most relevant to their corresponding event captions. To achieve this, we simultaneously enforce semantic alignment via a cross-modal objective and improve mask quality by supplementing supervision signals with synthetic captions, as illustrated in Figure 3.

3.1. Preliminaries

Baseline. ILCACM [11] introduces a Gaussian masking strategy that applies differentiable masks to video features, enabling implicit event localization through complementary captioning. Specifically, for each learnable event query, a transformer-based mask decoder outputs an event-specific feature embedding. This embedding is then passed through two separate fully-connected layers to predict two scalar values for the i-th event: its temporal center (c_i) and its width (w_i). These values are then used to construct a Gaussian mask M_i across the video's temporal axis t:

$$M_i(t) = \exp\left(-\frac{(t-c_i)^2}{2(w_i/\tau)^2}\right), \tag{1}$$

where τ is a temperature parameter that controls the sharpness of the Gaussian curve. To ensure that masks cover different regions, ILCACM employs a mask diversity loss based on cosine similarity between mask distributions.

Utilizing these Gaussian masks, the model generates two complementary features: the "positive" masked video features $(\boldsymbol{v} \cdot M_i)$ to produce a caption describing only the target event, while the "negative" masked video features $(\boldsymbol{v} \cdot (1 - M_i))$ are used to produce a caption describing all remaining events in the video, where $\boldsymbol{v}$ are frame-level embeddings from the spatial encoder and then these masked features are fed into the video encoder. The decoder takes as input these encoded visual features, along with prompts and target captions for both the positive and negative sets, and generates the corresponding captions in an auto-regressive manner. The resulting outputs are used to compute the positive captioning loss $\mathcal{L}_{\text{pos}}$ and negative captioning loss $\mathcal{L}_{\text{neg}}$. By constraining these two captions to form the complete video description (i.e., $C = C_i \cup \{C_j\}_{j=1, j\neq i}^{N_s}$), the mask prediction module learns to localize events implicitly.

3.2. Similarity-Aware Mask Guide

As established in the introduction, our goal is to generate event masks that are not merely temporally separated but are semantically aligned with their corresponding event captions. To achieve this, we design a mask generation process that is directly guided by cross-modal similarity.

Similarity-Aware Mask Training. We make the differentiable masks informative by designing a training objective that encourages them to highlight semantically salient regions. After a mask M_i is generated, it is multiplied element-wise with the video features $\boldsymbol{v}$ to produce positive masked video features $\boldsymbol{v}'_i = (\boldsymbol{v} \cdot M_i)$. These masked features serve as the visual representation of the emphasized region for event i. To ensure $\boldsymbol{v}'_i$ truly captures the correct

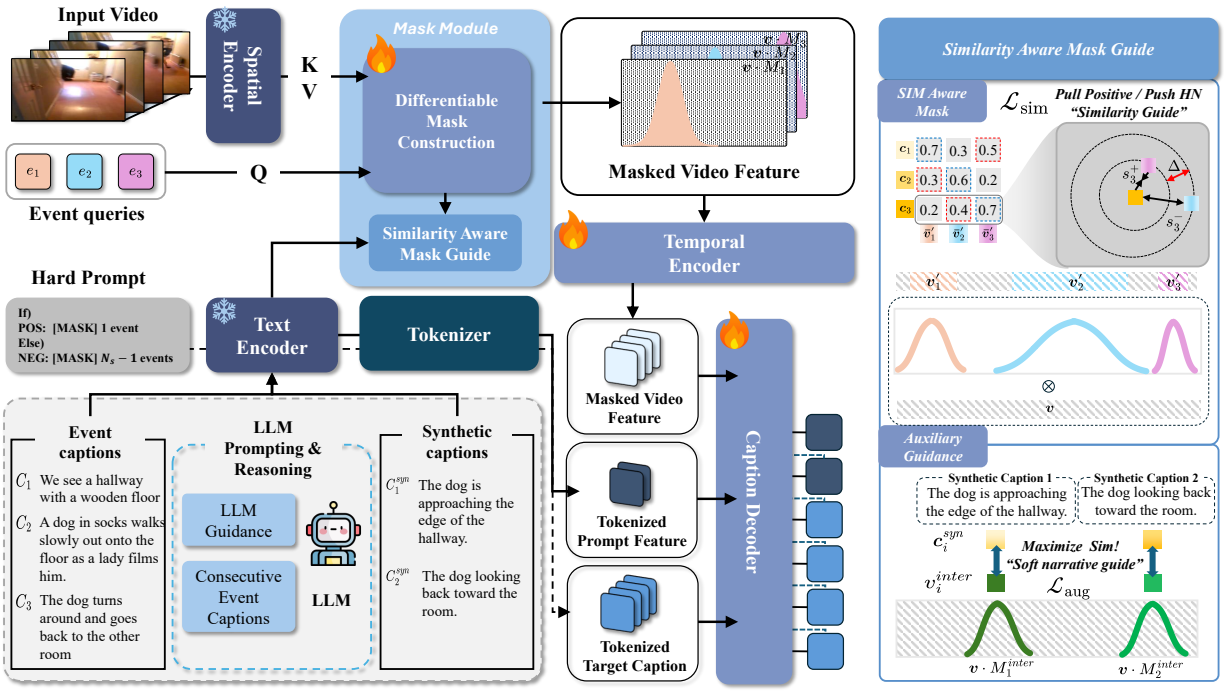


Figure 3. *SAIL* Pipeline. Our method exploits cross-modal similarity to guide mask optimization toward increased alignment with event captions and further enriches supervision through LLM-generated synthetic captions.

event content, we leverage event captions as semantic guidance. We maximize the cross-modal cosine similarity between the average pooled masked features $\bar{\boldsymbol{v}}'_i$ and their corresponding caption features $\boldsymbol{c}_i$ while minimizing similarity to other event captions in the same video through a margin ranking loss [19]:

$$\mathcal{L}_{\text{sim}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_s}\sum_{i=1}^{N_s}\max\left(0, \Delta - s^{+}_{b,i} + s^{-}_{b,i}\right), \quad (2)$$

where $s^{+}_{b,i} = \text{sim}(\bar{\boldsymbol{v}}'_{b,i}, \boldsymbol{c}_{b,i})$ is the positive pair similarity, $s^{-}_{b,i} = \max_{j\neq i}\text{sim}(\bar{\boldsymbol{v}}'_{b,i}, \boldsymbol{c}_{b,j})$ is the hard negative similarity within video v, B is the batch size, N_s is the number of events in video v, and Δ is a margin hyperparameter. By optimizing $\mathcal{L}_{\text{sim}}$, each mask learns to emphasize video regions that are semantically closest to its corresponding caption, effectively aligning the masks with event-specific visual content. For negative captioning, we follow the same approach as [11], constructing inverse masks and training them in the same manner as described above.

3.3. LLM-Based Caption Augmentation

While $\mathcal{L}_{\text{sim}}$ enables the masks to localize semantically salient event regions, we observe that the model often generates coarse event boundaries, especially when captions are temporally sparse in a given video. Empirically, we find that although event annotations may cover most of the video duration, they often consist of only a few coarse-grained descriptions with large temporal gaps between consecutive annotations (Figure 2 (c)). We hypothesize that such sparsity in temporal descriptions limits the model's ability to learn fine-grained event localization.

To better capture fine-grained event boundaries, we propose an LLM-based augmentation strategy that generates synthetic captions for intermediate events between consecutive ground-truth annotations. By leveraging the LLM's world knowledge and contextual reasoning [21, 30, 37], we guide the model to infer plausible transitional events that bridge these temporal gaps, thereby providing denser and more informative supervision across the video timeline.

Transition Caption Generation. First, we design a carefully structured prompt to guide the LLM in generating high-quality synthetic captions. Inspired by the effectiveness of role-based prompting in task-specific reasoning [37], we define the LLM's role as a "Video Context Inference Expert" specializing in temporal event analysis. The system prompt explicitly instructs the model to: (1) analyze both preceding and succeeding captions to understand narrative flow, (2) infer the most probable *transitional action* or *change of state* rather than merely rephrasing existing captions, and (3) maintain consistency in style, tone, and detail level with the original annotations. This structured format helps the LLM focus on the specific temporal gap while avoiding extraneous information. Then, we feed

consecutive captions C_i and C_{i+1} along with the structured query into the LLM to generate a synthetic description C_i^{syn} representing the transitional event. We apply this generation process to each pair of consecutive captions in a video, producing $N_s - 1$ synthetic captions per video. Further details about prompts are provided in the supplementary material.

Auxiliary Guidance via Synthetic Captions. The synthetic captions generated by LLM can provide a much denser, more fine-grained narrative guide. Although the synthetic captions contain plausible information, using them directly in the main contrastive loss $\mathcal{L}_{\text{sim}}$ (e.g., as hard negatives) could introduce noise and degrade performance (Table 6). Therefore, we propose to use these synthetic captions as a separate, auxiliary training signal. The goal is to guide the model to find visual evidence from these potential transitional events without strictly enforcing them. To achieve this, we first create a new set of "inter-masks" (M_i^{inter}), designed to represent the temporal segments between the predicted ground-truth events masks (M_i, M_{i+1}). For each adjacent pair of predicted event centers (c_i, c_{i+1}), we define a new inter-mask center c_i^{inter} as their average: $c_i^{inter} = \frac{c_i + c_{i+1}}{2}$. The width for these inter-masks is set to a fixed hyperparameter w^{inter}. These parameters (c_i^{inter}, w^{inter}) are fed into the same Gaussian function Equation (1) to generate a total of $N_s - 1$ new masks, M_i^{inter}. We then apply these inter-masks to the video features to get the augmented masked features $\boldsymbol{v}_i'^{inter} = (\boldsymbol{v} \cdot M_i^{inter})$, which are aggregated into representations. We apply average pooling to the augmented masked features, and $\mathcal{L}_{\text{aug}}$ encourages the resulting representations to align with their corresponding transition caption embeddings $\boldsymbol{c}_i^{syn}$:

$$\mathcal{L}_{\text{aug}} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_s - 1} \sum_{i=1}^{N_s - 1} \left(1 - \text{sim}(\bar{\boldsymbol{v}}_{b,i}'^{inter}, \boldsymbol{c}_{b,i}^{syn})\right). \tag{3}$$

This loss acts as a soft narrative guide, encouraging the model to capture the transitional events. This in turn, helps the main predictors for c_i and c_{i+1} to be more precise, allowing the model to represent more fine-grained event boundaries.

Final Objective. Our final training objective is a positive & negative captioning loss, cross-modal similarity loss, and the auxiliary augmentation loss:

$$\mathcal{L} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}} + \mathcal{L}_{\text{sim}} + \alpha_{\text{aug}} \mathcal{L}_{\text{aug}}. \tag{4}$$

α_{aug} is a hyper-parameters that controls the scale of $\mathcal{L}_{\text{aug}}$.

4. Experiments

4.1. Experimental Settings.

Datasets. Our experiments are conducted on two standard benchmarks for DVC. **ActivityNet Captions** [22] dataset is a larger collection of 20k untrimmed instructional videos, which average 120 seconds in length and are annotated with 3.7 localized events per video. **YouCook2** [43] dataset comprises approximately 2K untrimmed culinary videos, each with an average duration of 320 seconds and annotated with 7.7 localized language sentences.

Evaluation Metrics. We evaluate our model's performance on the two primary sub-tasks of DVC: caption generation and event localization. To evaluate caption quality, we adopt the official evaluation tool [22] to report standard metrics: METEOR [2], CIDEr [33], ROUGE-L [26] and BLEU-N [29]. To assess the narrative quality of the generated descriptions, we also employ the SODA_c metric [9]. For the event localization task, we measure the mean Average Precision (mAP), mean Average Recall (mAR), and the F1 score. In line with the standard protocol, all metrics for both tasks are averaged over a set of Intersection over Union (IoU) thresholds: $\{0.3, 0.5, 0.7, 0.9\}$.

Implementation Details. Following previous work [11], we use the Distilled-GPT2 model [31] as our caption decoder and the AdamW optimizer for training. For the ActivityNet Captions dataset [22], we set the initial learning rate to 1e-4 and train for 10 epochs for both the captioning and localization stages. For the YouCook2 dataset [43], we use an initial learning rate of 5e-5 and train for 5 and 15 epochs for the captioning and localization stages, respectively. For the synthetic caption generation, we utilize the Qwen3-8B [42]. The hyperparameters are fixed across all experiments with $\Delta = 0.1$, $w^{inter} = 0.6$ and $\alpha_{\text{aug}} = 0.25$.

4.2. Comparison with State-of-the-Art.

The results for the captioning and localization tasks on the ActivityNet Captions dataset are presented in Table 1 and Table 3, respectively. Our proposed model, *SAIL*, demonstrates the highest overall performance in both tasks. As shown in Table 1, our model achieves a CIDEr of 35.38, surpassing the previous state-of-the-art WSDVC model [11] and consistently outperforming it across all other captioning metrics. In the localization task, *SAIL* also shows the best performance, achieving the best F1 score of 57.00, which is driven by the highest recall (54.39) and precision (59.87) among all weakly-supervised methods. Remarkably, our weakly-supervised approach outperforms the fully-supervised method [14, 39] across the majority of metrics, despite not using temporal boundary annotations during training. These results validate that our semantically-aware masks successfully emphasize event-relevant regions, enabling the generation of high-quality captions that accurately describe target events. In Table 2, *SAIL* also achieves the highest captioning and localization scores on YouCook2 [43] compared to other WSDVC methods. These consistent improvements across datasets demonstrate our model's strong performance in both tasks.

Setting	Model	Features	Captioning							
			SODA_c	METEOR	CIDEr	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Fully-Supervised	DCE [22]	C3D	–	5.69	12.43	–	10.81	4.57	1.90	0.71
	DVC [25]	C3D	–	6.93	12.61	–	12.22	5.72	2.27	0.73
	PDVC [36]	C3D	5.26	7.50	25.87	–	–	–	–	1.65
	Vid2Seq [41]	CLIP	5.80	8.50	30.10	–	–	–	–	–
	CM^2 [14]	CLIP	6.18	8.55	33.01	–	–	–	–	2.38
	E^2DVC [39]	CLIP	6.13	8.57	33.63	–	–	–	–	2.43
Weakly-Supervised	WSDEC [8]	C3D	–	6.30	18.77	12.55	12.41	5.50	2.62	1.27
	ECG [38]	C3D	–	7.06	14.25	–	11.85	5.64	2.71	1.33
	EC-SL [4]	C3D	–	7.49	21.21	13.02	13.36	5.96	2.78	1.33
	PWS-DVC* [7]	C3D	–	7.28	20.59	12.71	–	–	–	1.35
	ILCACM [11]	CLIP	6.08	8.48	33.42	14.77	15.36	8.12	4.17	2.26
	SAIL (Ours)	CLIP	**6.29**	**8.63**	**35.38**	**15.29**	**15.86**	**8.30**	**4.27**	**2.30**

Table 1. Comparison with state-of-the-art methods on ActivityNet validation set. $*$ denotes results reported in [11]. Our method outperforms the previous state-of-the-art weakly-supervised approach [11] and remarkably surpasses several fully-supervised methods [14, 39, 41]. This demonstrates semantic alignment and caption augmentation can effectively replace temporal boundary supervision.

Setting	Model	Features	Captioning					Localization		
			SODA_c	METEOR	CIDEr	ROUGE-L	BLEU@N	R@AVG	P@AVG	F1
Weakly-Supervised	WSDEC* [8]	C3D	2.11	1.47	8.43	–	–	–	–	–
	PWS-DVC* [7]	C3D	3.14	2.48	9.81	–	–	–	–	–
	ILCACM† [11]	CLIP	3.60	3.41	13.49	4.75	2.59	17.76	18.01	17.88
	SAIL (Ours)	CLIP	**4.08**	**3.63**	**14.61**	**5.42**	**2.94**	**20.76**	**21.13**	**20.94**

Table 2. Comparison with other weakly-supervised methods on YouCook2 validation set. † means our re-implementation results, as the original paper does not report localization performance on this dataset. Our approach also achieves the best captioning and localization performance among all weakly-supervised methods on this dataset. BLEU@N denotes the average of BLEU-1 through BLEU-4.

Setting	Model	Features	Localization		
			R@Avg	P@Avg	F1
Fully-Supervised	SDVC [27]	C3D	55.58	57.57	56.56
	PDVC [36]	TSN	55.42	58.07	56.71
	Vid2Seq [41]	CLIP	52.70	53.90	53.29
	CM^2 [14]	CLIP	53.71	56.81	55.21
	E^2DVC [39]	CLIP	54.67	57.70	56.14
Weakly-Supervised	WSDEC* [8]	C3D	29.57	59.33	39.18
	PWS-DVC* [7]	C3D	40.85	55.82	47.09
	ILCACM [11]	C3D	53.09	59.20	55.98
	ILCACM [11]	CLIP	53.72	58.92	56.20
	SAIL (Ours)	CLIP	**54.39**	**59.87**	**57.00**

Table 3. Comparison of the event localization performance across various fully-supervised and weakly-supervised models on ActivityNet validation set. *SAIL* achieves state-of-the-art performance in most of the localization metrics.

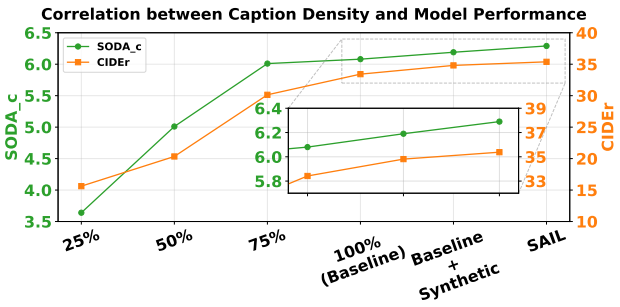


Figure 4. Impact of caption density on model performance. Performance decreases consistently as annotation density reduces from 100% to 25% (left), highlighting annotation sparsity as a critical challenge. Densifying supervision through LLM-generated synthetic captions (right) improves performance.

Correlation between Caption Density and Model Performance. To analyze the impact of event annotation density on WSDVC performance, we conduct experiments on the baseline method [11] using different proportions of ground-truth captions: 25%, 50%, 75%, and 100% (Baseline [11]). As shown in Figure 4, the model's performance drops consistently as the number of annotated events decreases. Notably, as WSDVC must learn event localization using only captions without temporal boundaries, sparse captions significantly degrade both the captioning quality (CIDEr) and the inter-sentence temporal alignment (SODA_c). The right side shows the performance when event annotations are densified through our LLM-based augmentation strategy. Beyond the original ground-truth descriptions, our synthetic caption generation improves performance even when simply incorporating them into the baseline method.

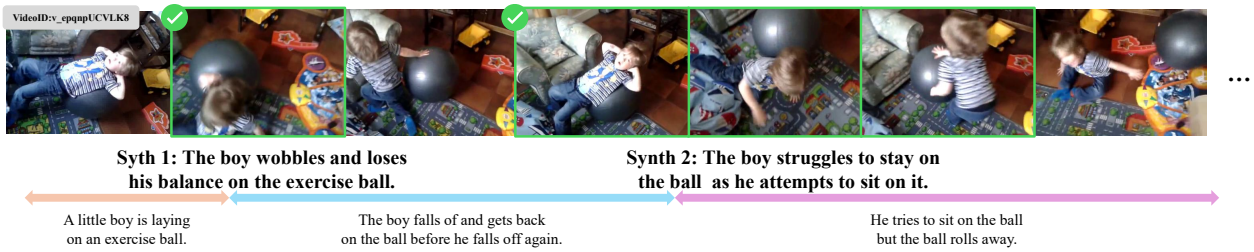


Figure 5. Synthetic caption's qualitative results. Our synthetic captions effectively capture potential intermediate events occurring between consecutive ground-truth annotations.

Sim	Synth	Captioning			Localization
		S_c	M	C	F1
✗	✗	6.08	8.48	33.42	56.20
✓	✗	6.27	8.58	35.18	56.89
✗	✓	**6.29**	8.52	34.92	56.79
✓	✓	**6.29**	**8.63**	**35.38**	**57.00**

Table 4. Ablation study on our key components on ActivityNet. **Sim** denotes similarity-aware mask, and **Synth** denotes Synthetic Captions with inter mask. S_c, M, and C denote SODA_c, METEOR, and CIDEr, respectively.

Method	Captioning			Localization
	S_c	C	R	F1
0% (Baseline)	6.08	33.42	14.77	56.20
25%	6.08	33.76	15.02	56.38
50%	6.27	34.19	15.00	56.39
75%	6.23	34.66	15.09	56.91
100% (Ours)	**6.29**	**35.38**	**15.29**	**57.00**

Table 5. Ablation study on the amount of synthetic captions used. The results suggest that the benefits of synthetic captions are evident even at a small scale. R denotes ROUGE-L.

Method	Captioning				
	S_c	M	C	R	B@4
ILCACM [11]	6.08	8.48	33.42	14.77	2.26
Synth + HN	6.24	8.44	33.84	15.10	2.02
Synth + Inter	**6.29**	**8.63**	**35.38**	**15.29**	**2.30**

Table 6. Ablation study on synthetic caption utilization strategies. Leveraging synthetic captions indirectly via an inter-mask mechanism yields the best overall performance. B@4 denotes BLEU-4.

4.3. Ablation Studies.

Component Ablation. In Table 4, we analyze the contribution of each component. We observe that applying a similarity-aware mask alone improves CIDEr by +1.76 and SODA_c by +0.21, indicating that similarity-based loss effectively produces masked features that are well-aligned with caption content. Similarly, employing synthetic captions with inter-event masks provides improvements for both captioning and localization, demonstrating the effectiveness of dense event supervision. Combining both components achieves the best performance, improving all metrics. This synergy enables our model to learn more discriminative event representations.

Data Scale Ablations. As shown in Table 5, incorporating even a small fraction (25%) of the synthetic captions leads to a performance gain over the baseline in both captioning and localization. Also, we observe that the performance in both tasks improves monotonically as the ratio of synthetic captions increases. Our full model (100%) achieves the highest scores across all metrics. Regardless of the ratio used, every configuration of our method consistently and substantially outperforms the baseline.

Synthetic Caption Utilization. We conduct experiments on different strategies for utilizing synthetic captions. First, we include synthetic captions and their corresponding inter-event masked video features when computing hard negatives for positive pairs in $\mathcal{L}_{\text{sim}}$. Second, as described in our method, we treat synthetic captions and inter-event masked features as auxiliary signals, computing $\mathcal{L}_{\text{sim}}$ only on the ground-truth annotation set while using synthetic captions as supplementary supervision through $\mathcal{L}_{\text{aug}}$. As shown in Table 6, simply incorporating synthetic captions (Synth + HN) already outperforms ILCACM [11], demonstrating the effectiveness of densifying supervision signals. Moreover, our auxiliary guidance approach, which provides a separate alignment signal, achieves the best results, validating our design choice of treating synthetic captions as soft guidance rather than hard constraints.

Mask-type Ablations. In Table 7, we evaluate different mask types used for video feature masking. Gaussian masks achieve the best performance at 35.38 CIDEr and 6.29 SODA_c, substantially outperforming Hard binary and Cauchy alternatives. Notably, all mask types show consistent improvements when combined with our proposed

Mask Type	Model	Captioning				Localization		
		SODA_c	METEOR	CIDEr	ROUGE-L	R@Avg	P@Avg	F1
Gaussian Mask	ILCACM [11]	6.08	8.48	33.42	14.77	53.72	58.92	56.20
	SAIL (Ours)	**6.29**	**8.63**	**35.38**	**15.29**	**54.39**	**59.87**	**57.00**
Hard Binary Mask	ILCACM† [11]	3.89	6.52	16.96	**11.24**	35.76	50.85	41.91
	SAIL (Ours)	**4.09**	**6.53**	**17.54**	11.07	**35.81**	**51.00**	**42.00**
Cauchy Mask	ILCACM† [11]	6.02	8.32	32.22	14.50	53.06	58.87	55.81
	SAIL (Ours)	**6.18**	**8.66**	**33.85**	**15.06**	**54.36**	**59.11**	**56.63**

Table 7. Ablation study of various mask designs. *SAIL* demonstrates consistent performance improvements over the baseline across all mask designs. † means our re-implementation localization results, as the original paper does not report localization performance.

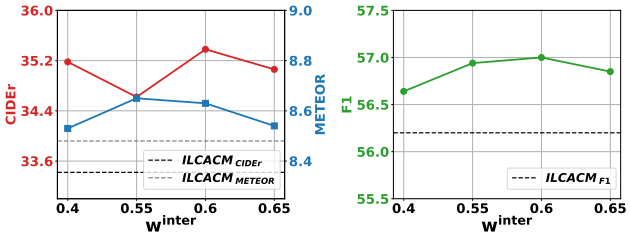


Figure 6. Impact of hyper-parameter w^{inter} for inter masks on captioning and localization performance.

Method	Train time	Inference time	GPU usage
ILCACM	1H 38M 27S	7M 11S	33.06 GiB
SAIL (Ours)	1H 41M 59S	7M 01S	33.11 GiB

Table 8. Computational cost comparison against the baseline [11].

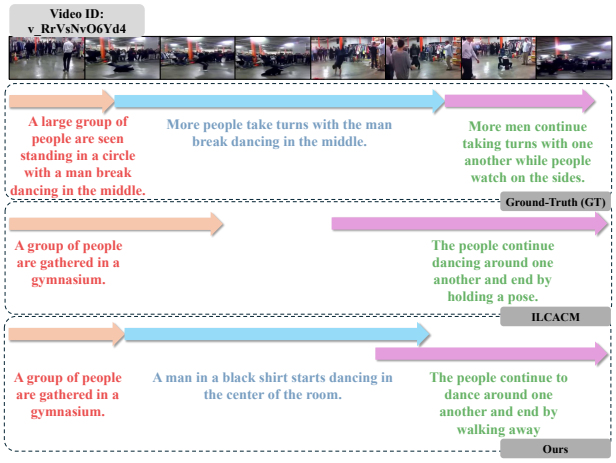


Figure 7. A qualitative result from the ActivityNet validation set.

framework compared to the baseline [11], demonstrating the effectiveness of our similarity-aware guidance and the use of synthetic captions with inter-event masks.

Performance-computation trade-off. We analyze the computational cost of our method in comparison to the baseline. All metrics are averaged over five independent runs. As shown in Table 8, the training time, inference time, and GPU memory consumption of *SAIL* remain nearly identical to those of the baseline [11]. Our approach only introduces a lightweight dot-product operation, incurring negligible overhead. Furthermore, the LLM-based caption augmentation is highly efficient; it is a one-time preprocessing step that operates solely on textual data, requiring only 0.28 hours to complete for the ActivityNet dataset.

Hyper-parameter Ablations. We conduct ablation experiments to evaluate captioning and localization scores under different values of w^{inter}, which controls the inter-event mask width, and maintain robust performance across different values(Figure 6). We choose $w^{\text{inter}} = 0.6$ for its balanced captioning and localization scores. Additional hyper-parameter experiments are in the supplementary material.

Qualitative Results. Figure 5 shows that our synthetic captions accurately describe the events situated between ground-truth annotations. Additionally, Figure 7 shows that our *SAIL* predicts event boundaries and captions that align well with the video content.

5. Conclusion

In this work, we propose *SAIL*, a novel weakly supervised approach to dense video captioning. Previous methods apply Gaussian masks to merely different temporal segments without accounting for semantic alignment between events and visual regions, leading to sub-optimal results. Our model addresses this limitation by leveraging cross-modal similarity to guide masks toward event-relevant regions. Crucially, we augment sparse GT annotations with LLM-augmented events, enabling our model to learn from a richer set of event-region alignments. Extensive experiments on ActivityNet and YouCook2 demonstrate the effectiveness of our proposed method. In future work, we plan to investigate methods for mitigating data scarcity across a diverse range of multimodal tasks [16, 18, 23, 28], beyond the DVC task.

Acknowledgments. This was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No.RS-2025-02215122, Development and Demonstration of Lightweight AI Model for Smart Homes).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] 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. 5

[3] SeungJu Cha, Kwanyoung Lee, Ye-Chan Kim, Hyunwoo Oh, and Dong-Jin Kim. Verbdiff: Text-only diffusion models with enhanced interaction awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8041–8050, 2025. 1

[4] 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. 3, 6

[5] Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. Video captioning with guidance of multimodal latent topics. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1838–1846, 2017. 1

[6] CE Chidume. Iterative approximation of fixed points of lipschitzian strictly pseudocontractive mappings. *Proceedings of the American Mathematical Society*, 99(2):283–288, 1987. 3

[7] Wangyu Choi, Jiasi Chen, and Jongwon Yoon. Pws-dvc: Enhancing weakly supervised dense video captioning with pretraining approach. *IEEE Access*, 11:128162–128174, 2023. 1, 3, 6

[8] 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, 3, 6

[9] 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. 5

[10] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017. 1

[11] Shiping Ge, Qiang Chen, Zhiwei Jiang, Yafeng Yin, Liu Qin, Ziyao Chen, and Qing Gu. Implicit location-caption alignment via complementary masking for weakly-supervised dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3113–3121, 2025. 1, 2, 3, 4, 5, 6, 7, 8

[12] 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. 3

[13] MinJu Jeon, Si-Woo Kim, Ye-Chan Kim, HyunGee Kim, and Dong-Jin Kim. Sali4vid: Saliency-aware video reweighting and adaptive caption retrieval for dense video captioning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25788–25801, 2025. 1, 3

[14] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904, 2024. 5, 6

[15] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Hicm^2: Hierarchical compact memory modeling for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4293–4301, 2025. 1, 3

[16] Si-Woo Kim, MinJu Jeon, Ye-Chan Kim, Soeun Lee, Taewhan Kim, and Dong-Jin Kim. Sync: Synthetic image caption dataset refinement with one-to-many mapping for zero-shot image captioning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2683–2692, 2025. 8

[17] Taewhan Kim, Soeun Lee, Si-Woo Kim, and Dong-Jin Kim. Vipcap: Retrieval text-based visual prompts for lightweight image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4320–4328, 2025. 1

[18] Ye-Chan Kim, SeungJu Cha, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. Sida: Synthetic image driven zero-shot domain adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 34–42, 2025. 8

[19] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *Advances in neural information processing systems*, 2014. 4

[20] Sungho Koh, SeungJu Cha, Hyunwoo Oh, Kwanyoung Lee, and Dong-Jin Kim. Scalediff: Higher-resolution image synthesis via efficient and model-agnostic diffusion. *arXiv preprint arXiv:2510.25818*, 2025. 1

[21] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 4

[22] 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. 2, 3, 5, 6

[23] Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. Ifcap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20715–20727, 2024. 8

[24] Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. Exploiting auxiliary caption for video grounding. In *Proceedings of the AAAI conference on artificial intelligence*, pages 18508–18516, 2024. 2

[25] 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. 6

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[27] 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. 6

[28] Hyunwoo Oh, SeungJu Cha, Kwanyoung Lee, Si-Woo Kim, and Dong-Jin Kim. Catchphrase: Exprompt-guided encoder adaptation for audio-to-image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9773–9782, 2025. 8

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[30] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019. 4

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 5

[32] 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. 2

[33] 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. 5

[34] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018. 1

[35] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[36] 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. 6

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 4

[38] Bofeng Wu, Guocheng Niu, Jun Yu, Xinyan Xiao, Jian Zhang, and Hua Wu. Weakly supervised dense video captioning via jointly usage of knowledge distillation and cross-modal matching. *arXiv preprint arXiv:2105.08252*, 2021. 1, 6

[39] Kangyi Wu, Pengna Li, Jingwen Fu, Yizhe Li, Yang Wu, Yuhan Liu, Jinjun Wang, and Sanping Zhou. Event-equalized dense video captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8417–8427, 2025. 5, 6

[40] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3

[41] 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, 3, 6

[42] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 5

[43] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5