Sali4Vid: Saliency-Aware Video Reweighting and Adaptive Caption Retrieval for Dense Video Captioning

MinJu Jeon Si-Woo Kim Ye-Chan Kim HyunGee Kim Dong-Jin Kim[†]
Hanyang University, South Korea.
{mnju5026, boreng0817, dpcksdl78, khjiiii2002, djdkim}@hanyang.ac.kr

Abstract

Dense video captioning aims to temporally localize events in video and generate captions for each event. While recent works propose end-to-end models, they suffer from two limitations: (1) applying timestamp supervision only to text while treating all video frames equally, and (2) retrieving captions from fixed-size video chunks, overlooking scene transitions. To address these, we propose *Sali4Vid*, a simple yet effective saliency-aware framework. We introduce *Saliency-aware Video Reweighting*, which converts timestamp annotations into sigmoid-based frame importance weights, and *Semantic-based Adaptive Caption Retrieval*, which segments videos by frame similarity to capture scene transitions and improve caption retrieval. *Sali4Vid* achieves state-of-the-art results on YouCook2 and ViTT, demonstrating the benefit of jointly improving video weighting and retrieval for dense video captioning.[1]

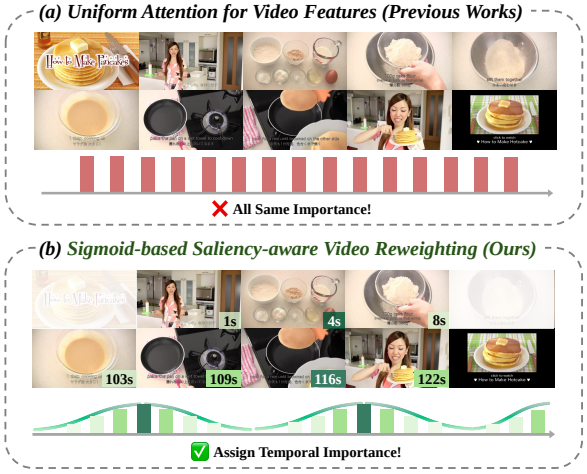

Figure 1: (a) Previous works incorporate timestamps only on the textual side, treating all video features with uniform features. (b) We propose *Sali4Vid*, a simple yet effective saliency-aware framework by leveraging sigmoid-based soft reweighting.

1 Introduction

The dense video captioning (DVC) task (Li et al., 2018; Wei et al., 2023; Duan et al., 2018; Zhou et al., 2018b; Krishna et al., 2017a; Mkhallati et al., 2023) aims to localize multiple events in untrimmed videos and generate descriptive captions for each. Unlike standard video captioning (VC) (Gao et al., 2017; Chen et al., 2017; Wang et al., 2018; Seo et al., 2022; Zhao et al., 2023; Lee et al., 2024; Kim et al., 2024a), which generates a single caption for a short and trimmed clip, DVC generates multiple temporally localized descriptions from long video streams, which is more challenging.

To effectively handle both localization and caption generation, prior works have proposed end-to-end modeling (Wang et al., 2021; Zhou et al., 2024). For instance, Vid2Seq (Yang et al., 2023) formulates DVC as a sequence-to-sequence task, adding time tokens to the text for timestamp supervision. More recently, CM^2 (Kim et al., 2024b) and HiCM^2 (Kim et al., 2025a) further extend this direction by retrieving auxiliary captions from an external datastore using video features as queries.

Despite these advances, existing methods still suffer from two key limitations. **First**, although fully-supervised timestamp annotations are available in training, previous work leverages them only on the textual side, while treating the video features as uniformly important across time as shown in Figure 1 (a). **Second**, recent caption retrieval methods (Kim et al., 2024b, 2025a) adopt fixed-size clip-level retrieval for auxiliary captions. However, this strategy overlooks semantic transitions and scene changes within the video, which can lead to misaligned or redundant caption retrieval. For example, as illustrated in Figure 2 (Left), unrelated actions like *preheat grill* and *remove squid* may be grouped into the same chunk, resulting in re-

[†]Corresponding author.
[1]Code: https://github.com/forminju/Sali4Vid

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25789–25802
November 4-9, 2025

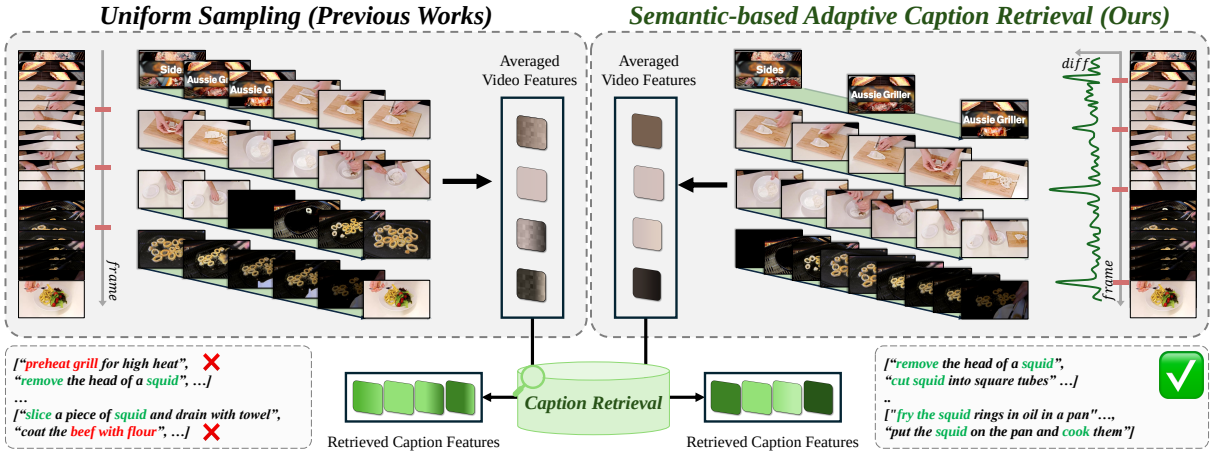

Figure 2: (Left) The previous caption retrieval approach overlooks the scene transition, leading to redundant or misaligned captions that may not adequately reflect meaningful changes in the video content. (Right) *Sali4Vid* adaptively segments frames based on similarity difference, enabling more contextually aligned and diverse caption retrieval for meaningful segments.

trieved captions that may not accurately describe each event.

To address these limitations, we propose *Sali4Vid*, a simple yet effective saliency-aware framework that explicitly applies temporal saliency cues to enhance video features during training and adaptively retrieves captions based on semantic transitions, providing more accurate event captions.

Specifically, we enhance video features by saliency reweighting based on timestamps, which emphasizes frames around annotated start and end points. This approach enables the model to directly utilize temporal supervision on the video side. Consequently, the model effectively focuses on salient visual regions, as demonstrated in Figure 1 (b). In addition, we calculate frame-to-frame similarity to find semantic transitions and segment the video adaptively, avoiding fixed-size clip-level retrieval that may group unrelated actions into the same chunk. This enables the retrieval of captions that are better aligned with each meaningful segment, as illustrated in Figure 2 (Right).

Our model contains two main key components: *Saliency-aware Video Reweighting* and *Semantic-based Adaptive Caption Retrieval*. **First**, during the training phase, the *Saliency-aware Video Reweighting* provides timestamp supervision to the visual side through sigmoid-based weights, allowing the model to continuously focus on salient video frames. **Second**, the *Semantic-based Adaptive Caption Retrieval* segments the video based on frame-to-frame similarity differences, finding meaningful semantic changes, and retrieves captions aligned with these semantically adaptive segments.

Empirically, our framework achieves state-of-the-art results with a CIDEr score of 75.80 on YouCook2 and 53.32 on ViTT, outperforming the previous state-of-the-art by +3.96 on YouCook2 and +2.58 on ViTT.

We summarize our contributions as follows:

- We propose *Sali4Vid*, a saliency-aware framework that enhances video features by applying sigmoid-based reweighting with timestamp supervision, focusing on more salient features in the training phase.
- We introduce a semantic-based adaptive caption retrieval strategy that segments videos based on frame-level similarity differences, enabling the retrieval of more contextually aligned captions for each semantic segment.
- We validate our method on YouCook2 and ViTT, achieving state-of-the-art results in both the localization and captioning tasks.

2 Related Work

2.1 Dense Video Captioning

Dense Video Captioning (DVC) aims to temporally localize events within untrimmed videos and generate captions for each event (Krishna et al., 2017b). Early approaches typically have adopted a two-stage "localize-and-describe" pipeline (Iashin and Rahtu, 2020a,b). However, this separation between localization and captioning often overlooks

the interaction between the two subtasks, leading to suboptimal performance. To address this, recent works have explored end-to-end frameworks that jointly model event localization and captioning. PDVC (Wang et al., 2021) reformulates DVC as a set prediction problem using a DETR-style transformer (Carion et al., 2020), enabling parallel prediction of temporal segments and captions without relying on intermediate proposals. More recently, Vid2Seq (Yang et al., 2023) formulates dense video captioning as a sequence-to-sequence task, generating both timestamp tokens and captions in a unified output while leveraging large-scale speech transcriptions. Building on this, Streaming V2S (Zhou et al., 2024) introduces streaming decoding with visual memory for online captioning, and DIBS (Wu et al., 2024) proposes scalable pretraining with pseudo-labeled segments. CM^2 (Kim et al., 2024b) and HICM^2 (Kim et al., 2025a) further extend this line of work by integrating retrieval-augmented generation using external caption memories. Unlike previous methods that apply timestamp supervision only to text, overlooking video-side temporal modeling, our *Sali4Vid* explicitly leverages timestamp annotations to reweight video features and adaptively retrieves segment-level captions based on semantic transitions.

3 Proposed Method

Recent work in dense video captioning (Kim et al., 2025a; Wu et al., 2024; Yang et al., 2023; Zhou et al., 2024) often utilizes timestamp annotations only on the text side, while treating all video frames as equally important. To address these limitations, we propose *Sali4Vid*, a framework that explicitly models frame-level importance through two complementary strategies, as illustrated in Figure 3.

First, *Sali4Vid* applies sigmoid-based time stamp-guided weighting to highlight salient frames, providing explicit temporal supervision on the visual side during training. *Second*, our model captures semantic transitions by measuring frame-to-frame similarity, enabling adaptive segmentation for retrieving captions that better align with meaningful video segments. Together, these strategies improve event localization and caption generation by focusing on important visual content and retrieving contextually relevant captions. We detail these components in Section 3.1 for saliency-aware video reweighting and Section 3.2 for semantic-based adaptive caption retrieval.

Preliminaries. We build on the structure of the Vid2Seq (Yang et al., 2023), fine-tuning a model pre-trained on 1.8 million videos. Given an input video, we extract frame-level features $\boldsymbol{x}^{spat} = \{\boldsymbol{x}_i^{spat}\}_{i=1}^{T}$ using CLIP ViT-L/14 (Radford et al., 2021; Dosovitskiy et al., 2020). Spatial features $\boldsymbol{x}^{spat}$ are processed by a temporal transformer to obtain context-aware video features $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{T}$.

The goal is to predict a set of event segments and corresponding captions $(t_n^s, t_n^e, C_n)_{n=1}^{N}$, where t_n^s and t_n^e denote the start and end timestamps, and C_n is the generated caption. Timestamps are normalized over the video duration d.

During training, we use ground-truth annotations (t_n^s, t_n^e, C_n) and speech transcripts features $\boldsymbol{y}$. During inference, the model predicts event boundaries and captions without ground-truth timestamps, relying on video features $\boldsymbol{x}$ and transcript features $\boldsymbol{y}$, following the Vid2Seq setup.

3.1 Saliency-aware Video Reweighting

Building on the observation that prior works (Yang et al., 2023; Kim et al., 2025a) treat video frames uniformly despite having timestamp annotations, we propose a *Saliency-Aware Video Reweighting* method that directly leverages these annotations to compute frame-level importance scores as continuous, fully-supervised weights. Unlike methods that rely solely on textual cues (Wu et al., 2024; Zhou et al., 2024) or require additional modules to infer saliency score (Ge et al., 2025), our approach makes direct use of ground-truth event boundaries to provide continuous, fine-grained frame weighting in the training phase.

Specifically, we assign continuous sigmoid-based weights to each frame, as illustrated in Figure 3. For each annotated event n with start and end times (t_n^s, t_n^e), we define the sigmoid-based importance weight for frame i as follows:

$$W_n^L(i) = Sigmoid\left(\alpha \cdot \left(\frac{i}{T} - \frac{t_n^s}{d}\right)\right), \quad (1)$$

$$W_n^R(i) = Sigmoid\left(\alpha \cdot \left(\frac{t_n^e}{d} - \frac{i}{T}\right)\right), \quad (2)$$

$$W_n(i) = W_n^L(i) \times W_n^R(i), \quad (3)$$

where α controls the sharpness of the sigmoid curve, and $\frac{t_n^s}{d}$, $\frac{t_n^e}{d}$ denote the normalized start and end of the n-th event.

If multiple events exist, the final importance weight is computed as:

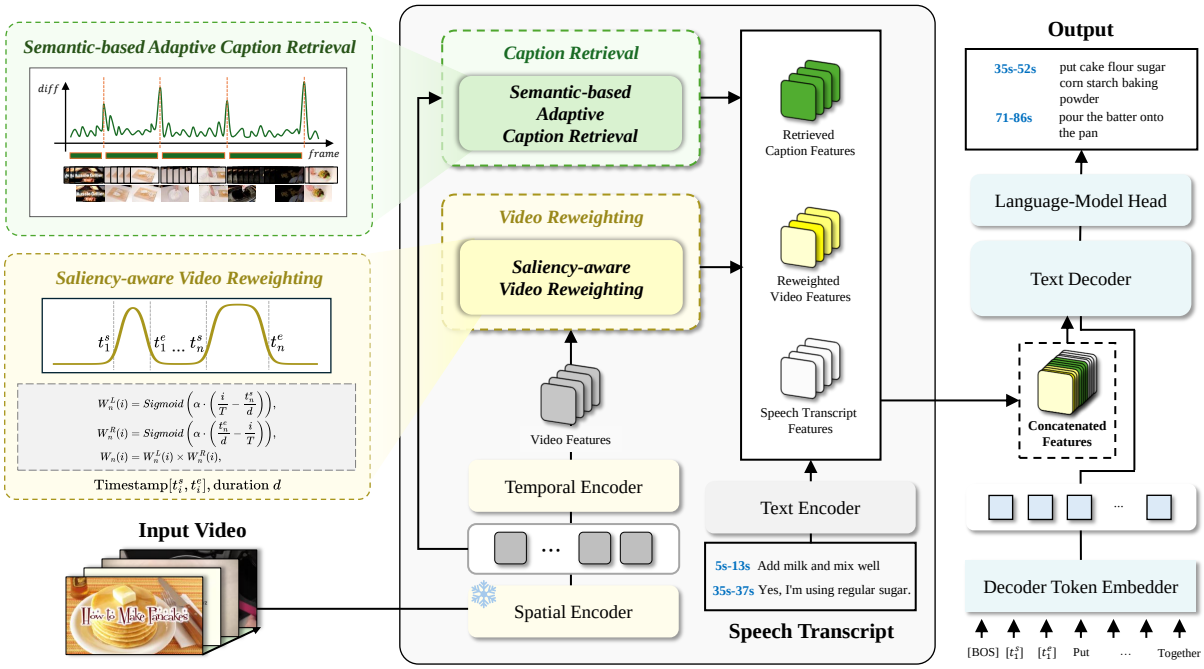

Figure 3: Overview of our *Sali4Vid* framework for dense video captioning. *Sali4Vid* enhances dense video captioning by combining *Saliency-Aware Video Reweighting* with *Semantic-based Adaptive Caption Retrieval*. Specifically, we utilize timestamp supervision to softly reweight video features in the training phase and retrieve relevant captions by clustering semantically similar video frames. The reweighted video features $\hat{\boldsymbol{x}}$, segment-level retrieved caption features $\tilde{\boldsymbol{r}}$, and speech features $\boldsymbol{y}$ are then concatenated and passed through the cross-attention layer of the text decoder, enabling the model to better localize events and generate accurate event captions.

$$W(i) = \max_n W_n(i). \tag{4}$$

The frame feature $\boldsymbol{x}_i$ is then reweighted by the importance score as follows:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i \cdot W(i), \tag{5}$$

resulting in reweighted video features $\hat{\boldsymbol{x}} = \{\hat{\boldsymbol{x}}_i\}_{i=1}^{T}$ used for decoding.

Our sigmoid-based reweighting strategy enables the model to effectively highlight salient event regions by assigning continuous importance scores to each frame, capturing both central areas and temporal boundaries. By providing smooth and fully supervised temporal guidance, it allows the model to focus on salient visual features, leading to improved caption generation and event localization, as illustrated in Figure 4.

3.2 Semantic-based Adaptive Caption Retrieval

Recent studies (Kim et al., 2024b, 2025a) suggest that auxiliary captions from an external datastore can provide useful semantic context for dense video captioning. However, these methods typically retrieve captions based on fixed-sized clip-level video features, overlooking the dynamic scene transitions within the video. When fixed-size clip-level video features are used as queries, multiple events may be mixed within a single clip. This makes it challenging to retrieve accurate captions that correctly match the target event. To address this limitation, we propose a *Semantic-based Adaptive Caption Retrieval* that adaptively finds semantic segments and retrieves relevant captions for each segment.

Frame Difference Calculation. We first compute frame-to-frame feature similarity differences to capture the semantic transitions. Given frame-level spatial features $\{\boldsymbol{x}_i^{spat}\}_{i=1}^{T}$, we calculate the cosine difference between consecutive frames as:

$$D(i) = 1 - \text{sim}\left(\boldsymbol{x}_i^{spat}, \boldsymbol{x}_{i+1}^{spat}\right), \tag{6}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. A large value of $D(i)$ signals a strong semantic change.

Adaptive Segment Construction. After calculating the frame-level semantic difference $D(i)$, we segment the video based on $D(i)$. A straightforward method is to set a boundary whenever $D(i)$ exceeds a fixed threshold τ_{fixed}. However, this frame-wise segmentation is sensitive to small changes and noise, which can result in over-

segmentation by splitting stable scenes into excessively short segments.

To address this issue, we adopt a *momentum-based accumulation strategy* that continuously aggregates frame differences and captures boundaries only when a sustained change is observed. This approach reduces sensitivity to small variations and improves boundary detection by considering the accumulated difference shift across frames, as inspired by (Kordopatis-Zilos et al., 2019).

We first define the adaptive threshold τ_{adap} using the mean μ and standard deviation σ of $\{D(i)\}_{i=1}^{T-1}$ as $\tau_{adap} = \mu + \beta \cdot \sigma$, where β is a scaling factor. Starting from the first frame indexed by $s_{\text{cur}} = 1$, we initialize the running segment feature with the first frame feature as $\boldsymbol{z}_{\text{cur}} = \boldsymbol{x}_1^{spat}$. We iteratively grow a segment by tracking $\boldsymbol{z}_{\text{cur}}$. For each subsequent frame $i+1$, we compute the semantic difference between the current segment feature and the incoming frame:

$$D'(i) = 1 - \boldsymbol{z}_{\text{cur}} \cdot \boldsymbol{x}_{i+1}^{spat}, \tag{7}$$

where $\boldsymbol{x}_{i+1}^{spat}$ is the spatial feature of frame $i+1$.

We apply the following decision rule: If $D'(i) > \tau_{adap}$, the current segment is ended at i, a new segment starts at $i+1$, and the segment feature is reset to $\boldsymbol{z}_{\text{cur}} = \boldsymbol{x}_{i+1}^{spat}$.

Otherwise, we include frame index $i+1$ in the current segment and update the segment feature by computing the moving average as:

$$\boldsymbol{z}_{\text{cur}} \leftarrow \frac{|\boldsymbol{S}_{\text{cur}}| \cdot \boldsymbol{z}_{\text{cur}} + \boldsymbol{x}_{i+1}^{spat}}{|\boldsymbol{S}_{\text{cur}}| + 1}, \tag{8}$$

where $|\boldsymbol{S}_{\text{cur}}|$ is the number of frames in the current segment. This process continues until all frames are processed, yielding segments $\{\boldsymbol{S_1}, \boldsymbol{S_2}, \ldots, \boldsymbol{S_m}\}$, where each segment is defined as the set of consecutive frame indices as $\boldsymbol{S_m} = \{s_m, s_m+1, \ldots, e_m\}$, with s_m and e_m denoting the start and end frame indices of the segment. Each corresponding segment is computed as the average of the frame features within the segment as $\boldsymbol{z}_{\boldsymbol{S_m}} = \frac{\sum_{j=s_m}^{e_m} \boldsymbol{x}_j^{spat}}{|\boldsymbol{S_m}|}$, which serves representation for retrieving semantically aligned auxiliary captions.

Segment-level Caption Retrieval. After segmenting the video into $\{\boldsymbol{S_m}\}_{m=1}^{M}$, where M is the total number of adaptively captured segments, we retrieve semantically aligned captions for each segment using its feature representation $\boldsymbol{z}_{\boldsymbol{S_m}}$, providing localized textual guidance to each segment.

Given the segment-level feature $\boldsymbol{z}_{\boldsymbol{S_m}}$, we compute similarity scores against the external caption datastore $R = \{\boldsymbol{r}_r\}_{r=1}^{N_R}$ and retrieve the Top-k semantically aligned captions as:

$$\mathcal{R}_{\boldsymbol{S_m}} = \underset{\boldsymbol{r}_r \in R}{\text{Top-}k} \left(\text{sim}(\boldsymbol{r}_r, \boldsymbol{z}_{\boldsymbol{S_m}})\right), \tag{9}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and $\mathcal{R}_{\boldsymbol{S_m}} \in \mathbb{R}^{k \times D}$ represents the Top-k retrieved caption embeddings for $\boldsymbol{S_m}$. We then aggregate these retrieved caption embeddings by average pooling to obtain a single caption guidance vector:

$$\tilde{\boldsymbol{r}}_{\boldsymbol{S_m}} = \frac{1}{k} \sum_{\boldsymbol{r} \in \mathcal{R}_{\boldsymbol{S_m}}} \boldsymbol{r}, \tag{10}$$

where $\tilde{\boldsymbol{r}}_{\boldsymbol{S_m}} \in \mathbb{R}^D$ is the averaged embedding representing the external semantic guidance for $\boldsymbol{S_m}$. We repeat this process for all segments, resulting in a set of segment-wise caption embeddings $\tilde{\boldsymbol{r}} = \{\tilde{\boldsymbol{r}}_{\boldsymbol{S_m}}\}_{m=1}^{M}$ that is used in decoding to provide textual guidance aligned with each segment.

3.3 Model Training and Inference

Our model integrates reweighted video features $\hat{\boldsymbol{x}}$, segment-level retrieved caption features $\tilde{\boldsymbol{r}}$, and speech transcripts features $\boldsymbol{y}$ to improve event localization and caption generation. We extract frame features $\{\boldsymbol{x}_i^{spat}\}_{i=1}^{T}$, and perform segment-wise retrieval of Top-k caption embeddings $\tilde{\boldsymbol{r}}_{\boldsymbol{S_m}}$ from an external datastore for each segment. The reweighted video features $\hat{\boldsymbol{x}}$ are obtained as described in Section 3.1. Speech transcripts are encoded using a transformer-based text encoder with time tokens to obtain $\boldsymbol{y}$.

We train the model using a cross-entropy loss conditioned on $\hat{\boldsymbol{x}}$, $\tilde{\boldsymbol{r}}$, $\boldsymbol{y}$ to predict the target sequence $\boldsymbol{o}$:

$$\mathcal{L}_\theta = \text{CE}(\boldsymbol{o} \mid \hat{\boldsymbol{x}}, \tilde{\boldsymbol{r}}, \boldsymbol{y}), \tag{11}$$

where θ denotes model parameters. During inference, the model generates event-aware captions based on the video features, transcript features, and segment-level retrieved caption features. Unlike training, where annotated timestamps are used to apply importance weights, we perform inference *without timestamp supervision and do not apply any weighting to the video features*.

Method	PT	YouCook2 (val)				ViTT (test)			
		CIDEr	METEOR	SODA_c	BLEU4	CIDEr	METEOR	SODA_c	BLEU4
PDVC ICCV21	✗	29.69	5.56	4.92	1.40	-	-	-	-
CM^2 CVPR24	✗	31.66	6.08	5.34	1.63	-	-	-	-
Streaming V2S CVPR24	✓	32.90	7.10	6.00	-	25.20	5.80	10.00	-
DIBS CVPR24	✓	44.44	7.51	6.39	-	-	-	-	-
Vid2Seq† CVPR23	✓	66.29	12.41	9.87	5.64	48.84	9.51	14.99	0.71
HiCM^2 AAAI25	✓	<u>71.84</u>	<u>12.80</u>	**10.73**	<u>6.11</u>	<u>51.29</u>	<u>9.66</u>	<u>15.07</u>	<u>0.86</u>
Ours	✓	**75.80**	**13.54**	<u>10.28</u>	**6.35**	**53.87**	**10.05**	**15.08**	**0.91**

Table 1: Comparison with state-of-the-art methods on YouCook2 validation and ViTT test sets. PT indicates whether the model is pretrained. Bold and underline denote the best and second-best scores, respectively. "-" indicates unavailable results. † denotes results reproduced from official implementations. Our method achieves state-of-the-art performance in most of the metrics.

Method	PT	YouCook2 (val)			ViTT (test)		
		F1	Recall	Precision	F1	Recall	Precision
PDVC	✗	26.81	22.89	32.37	-	-	-
CM^2	✗	28.43	24.76	33.38	-	-	-
Streaming V2S	✓	24.10	-	-	35.40	-	-
DIBS	✓	31.43	26.24	**39.81**	-	-	-
Vid2Seq†	✓	31.08	30.38	31.81	<u>46.21</u>	**45.89**	46.53
HiCM^2	✓	<u>32.51</u>	**32.51**	32.51	45.98	45.00	<u>47.00</u>
Ours	✓	**33.61**	<u>31.11</u>	<u>36.57</u>	**46.58**	44.31	**49.10**

Table 2: Localization results on YouCook2 validation and ViTT test sets. Bold denotes the best performance, and underline denotes the second-best performance. "-" result is unavailable.

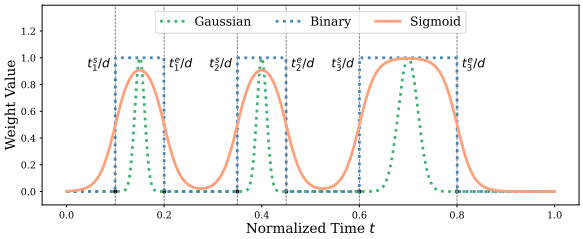

Figure 4: Comparison of different weights with multiple-timestamps. Unlike Gaussian or binary, our sigmoid-based weight provides continuous importance weights while preserving the start and end boundaries.

4 Experiment

4.1 Experimental Settings

Datasets. **YouCook2** (Zhou et al., 2018a) consists of 2,000 untrimmed videos. On average, 320 seconds and 7.7 localized sentences per video. **ViTT** (Huang et al., 2020) includes 8,000 untrimmed instructional videos averaging 250 seconds and annotated with 7.1 localized short tags.

Evaluation Metrics. We evaluate our method on two sub-tasks in DVC. By using the official evaluation tool (Wang et al., 2020), we use CIDEr (Vedantam et al., 2015), BLEU4 (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) metrics, which compare the generated captions to the ground truth across IoU thresholds of (0.3, 0.5, 0.7, 0.9). Additionally, to assess storytelling ability, we use the SODA_c metric (Fujita et al., 2020). For event localization, we calculate the average precision, average recall, and F1 score, averaging these metrics over IoU thresholds of (0.3, 0.5, 0.7, 0.9).

Implementation Details. Following previous works (Yang et al., 2023; Kim et al., 2025a), we build upon the Vid2Seq model that is pre-trained on 1.8M videos, which uses the T5-Base model (Raffel et al., 2020) as both the text encoder and decoder. Video frames are extracted at 1 FPS and sub-sampled or padded to a fixed length $T = 100$. The model is first trained for 10 epochs following Vid2Seq with a learning rate of 3e-4, then fine-tuned for 10 more epochs with our method using a learning rate of 1e-6, linearly warmed up over the first 10% of steps and decayed to 0 via a cosine schedule. Training is performed on a single A6000 GPU with batch size 8, taking approximately 1h 20m total (4m 20s/epoch). We set $\alpha = 10.0$ for sigmoid reweighting and $\beta = 1.0$ for adaptive segmentation. Hyperparameter details are provided in the supplementary. The caption retrieval datastore is constructed from the training captions only.

Comparison with State-of-the-Arts. Table 1 and Table 2 summarize the results on YouCook2 and ViTT. Our *Sali4Vid* achieves the best overall performance in both captioning and localization tasks. Specifically, in Table 1, our method achieves a CIDEr of **75.80** on YouCook2 and **53.87** on ViTT, outperforming prior state-of-the-art methods such as HiCM^2 (Kim et al., 2025a) and Vid2Seq (Yang et al., 2023) across most metrics. This improvement stems largely from our use of sigmoid-based

Method	Captioning				Localization		
	CIDEr	METEOR	SODA_c	BLEU4	F1	Recall	Precision
Baseline	66.29	12.41	9.87	5.64	31.08	30.38	31.81
Different Design of Saliency-Weights							
Hard binary mask	68.85	12.53	10.25	5.95	32.53	**33.06**	32.03
Gaussian weights	68.66	12.49	10.30	5.93	32.25	32.81	31.72
Sigmoid weights	**74.72**	**13.43**	**10.35**	**6.01**	**33.34**	31.24	**35.76**
Different Segment Feature Design for Caption Retrieval							
Mean-Pool	**75.80**	**13.54**	**10.28**	**6.35**	**33.61**	**31.11**	**36.57**
Max-Pool	75.02	13.48	10.22	6.23	33.01	30.37	36.17
Key-Frame	74.87	13.41	10.16	6.26	33.12	30.42	36.34

Table 3: Component-wise results on the YouCook2 validation set for both captioning and localization tasks. Sigmoid-based weights achieve the best performance among reweighting strategies, while mean-pooling caption retrieval shows the best results among the retrieval designs.

Mask Design	YouCook2			
	C	M	S_c	F1
Baseline	66.29	12.41	9.87	31.08
Start skew	67.45	12.44	10.29	32.22
End skew	67.54	12.47	**10.31**	32.22
Random skew	67.06	12.27	10.27	32.61
Center skew (Ours)	**75.80**	**13.54**	10.28	**33.61**

Table 4: Ablation study of various mask designs in sigmoid-based importance modeling with adaptive caption retrieval.

Retrieval Design	YouCook2			
	C	M	S_c	F1
Baseline	66.29	12.41	9.87	31.08
Fixed-size Clip-level	63.96	12.14	9.93	32.24
τ_{fixed} + w/o MMT	66.05	12.33	10.23	32.73
τ_{fixed} + MMT	66.87	12.26	10.32	32.49
τ_{adap} + w/o MMT	66.65	12.44	10.20	32.76
τ_{adap} + MMT (Ours)	**68.63**	**12.61**	**10.33**	**32.79**

Table 5: Ablation study on different designs for caption retrieval *without* video reweighting. MMT denotes the momentum-based accumulation strategy. For the fixed-size setting, we set the window size to 10. C, M, and S_c denote CIDEr, METEOR, and SODA_c, respectively.

video reweighting and semantic-based adaptive caption retrieval, which allows the model to generate more accurate and semantically aligned captions for each localized region.

In Table 2, *Sali4Vid* achieves the highest F1 scores of **33.61** on YouCook2 and **46.58** on ViTT. We also observe an improvement in precision over the baseline, +4.76 on YouCook2 and +2.57 on ViTT, demonstrating the effectiveness of applying supervision directly to video features during training via our saliency-aware reweighting strategy.

4.2 Ablation Study

We conduct ablation studies on the YouCook2 validation set to analyze the contributions of each proposed component, including saliency-aware video reweighting, semantic-based adaptive caption retrieval, and their combined impact on performance.

Saliency-aware Video Reweighting. In Table 3, we compare different importance weighting strategies. Sigmoid-based weighting achieves the best performance with a 74.72 CIDEr, significantly outperforming binary (68.85) and Gaussian (68.66) approaches and yielding a +8.43 CIDEr improvement over the baseline. We hypothesize that the sigmoid's smooth transitions around annotated event boundaries are particularly effective in our fully supervised setting. This property also substantially improves precision to 35.76 (vs. 32.03 for binary) by reducing false positives from overly sharp boundary decisions, as shown in Figure 4.

We further investigate sigmoid mask designs in Table 4, comparing our proposed center-skew design against variants that emphasize the start, end, or random regions of an event. The center-skew mask consistently outperformed others, demonstrating that emphasizing the central region while preserving boundary information is most aligned with the structure of instructional videos.

Semantic-based Adaptive Caption Retrieval. Ta-

k Retrieved Captions	YouCook2 C	M	S_c	F1
5	75.22	13.50	**10.39**	33.51
10	**75.80**	**13.54**	10.28	33.61
20	75.51	13.52	10.26	33.58
30	75.52	13.52	10.29	**33.65**

Table 6: Ablation study on different numbers of captions used for caption retrieval.

Data Stores	YouCook2 C	M	S_c	F1
COCO (2014)	75.51	**13.54**	10.44	33.59
CC3M (2021)	75.33	13.51	10.38	33.53
Hierarchical (2025a)	**76.77**	13.38	**10.57**	32.92
In-domain	75.80	**13.54**	10.28	**33.61**

Table 7: Ablation study on different datastores used for caption retrieval.

Reweight	Cap	YouCook2 C	M	S_c	F1
✗	✗	66.29	12.41	9.87	31.08
✗	✓	68.63	12.61	10.33	32.79
✓	✗	74.72	12.49	**10.35**	33.34
✓	✓	**75.80**	**13.54**	10.28	**33.61**

Table 8: Ablation study on our key components on YouCook2. **Reweight** denotes saliency-aware video reweighting, and **Cap** denotes semantic-based adaptive caption retrieval. Applying both components achieves the best performance.

Method	Positive (↑)	Negative (↓)	IoU@0.1 (↑)
Baseline (2023)	0.23	0.21	0.038
Ours	**0.25**	**0.20**	**0.049**

Table 9: Attention score and IoU comparison during inference. Positive and Negative denote the average attention within and outside ground-truth segments, while IoU@0.1 measures overlap between the top-10% attention regions and ground truth.

ble 3 shows that mean-pooling performs best among aggregation strategies, improving CIDEr to **75.80** and F1 to **33.61**. Table 5 further demonstrates the effectiveness of our momentum-based accumulation combined with adaptive thresholding without video reweighting, achieving the highest performance with a CIDEr of **68.63** and an F1 of **32.79**. In contrast, the fixed-size setting shows slightly lower performance and needs extensive window tuning across datasets. Our semantic-based adaptive retrieval alleviates this by using an adaptive thresholding strategy with a single scaling factor β.

We also analyze the number of retrieved features in Table 6, where retrieving 10 captions per segment performs best. Additionally, Table 7 shows that our approach is robust across datastores, while in-domain captions yield the highest F1 score, out-of-domain datastores like COCO (Lin et al., 2014) and CC3M (Changpinyo et al., 2021) produce comparable results, and hierarchical memory (Kim et al., 2025a) boosts both CIDEr and SODA_c.

Component Ablation. In Table 8, we analyze the contribution of each component. We observe that applying adaptive caption retrieval alone improves CIDEr by **+2.34** and F1 by **+1.71**, showing that adaptive caption retrieval is beneficial to both captioning and localization. Saliency-aware reweighting alone improves CIDEr by **+8.43** and F1 by **+2.26** compared to baseline, confirming the importance of focusing on informative frames. Combining both achieves the best performance, improving all metrics. These results show that the two components are complementary and most effective when applied together.

Efficiency of Caption Retrieval. Table 10 presents the efficiency-performance trade-off under different retrieval subset sizes. Notably, even a small subset achieves comparable performance to the full set, while reducing retrieval cost to around 2 ms per video. In addition, segmentation requires about 16 ms per video, which is not negligible but performed only once using a lightweight, model-free clustering algorithm based on frame similarity. We will revise our manuscript to include this segmentation time report, further demonstrating the overall efficiency of our retrieval strategy.

Analysis of Train-Test Mismatch Setting. During training, our saliency-aware video reweighting leverages timestamp supervision to guide the model, but not at test time. We regard this process as representation learning: with stronger supervision during training, the model learns more informative features and thus performs better at inference without extra components. To validate this, we compare attention maps from the last layer of the temporal encoder between Baseline (Yang et al., 2023) and our method in Table 9. The results show that our training strategy improves temporal focus, supporting accurate captioning during inference.

Qualitative Results. Figure 5 shows examples from the YouCook2 validation set. Our *Sali4Vid* predicts event boundaries and captions that align

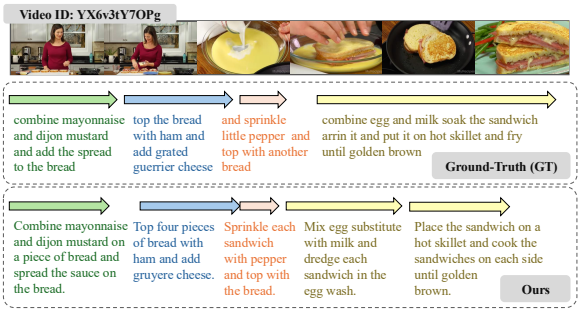

Figure 5: A qualitative result from YouCook2 validation set.

Subset	# Caps	Segment	Search	C	S_c	F1
0%	0	-	-	74.72	**10.35**	33.34
10%	0.96K	16.02	2.07	74.92	10.17	33.27
30%	2.8K	16.02	2.09	75.67	10.25	**33.62**
100%	9.6K	16.02	2.11	**75.80**	10.28	33.61

Table 10: Efficiency-performance trade-off under varying relevant caption subset sizes (ms/vid).

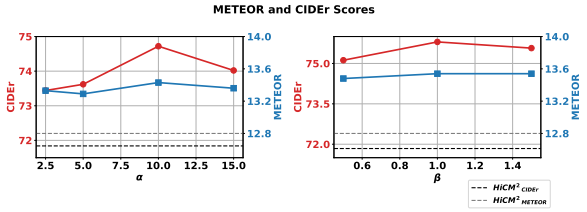

Figure 6: Impact of hyper-parameter α for video reweighting and β for semantic-based caption retrieval on captioning performance.

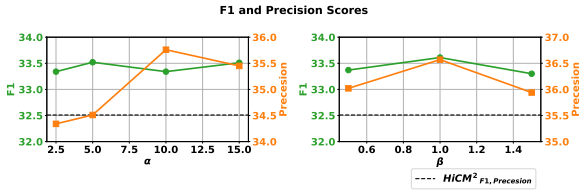

Figure 7: Impact of hyper-parameter α for video reweighting and β for semantic-based caption retrieval on localization performance.

well with the video content. For instance, it separates two closely occurring events (67–98s; yellow arrow) into distinct segments, illustrating its ability to capture detailed event transitions.

Hyperparameter Choice. We conducted ablation studies on two key parameters: the sharpness factor α for the sigmoid curve and the scaling factor β that controls τ_{adap}. Figure 6 presents CIDEr and METEOR scores across different values of α and β, while Figure 7 reports the corresponding F1 and Precision for localization. In all cases, our method maintains performance above the previous state-of-the-art (Kim et al., 2025a), demonstrating robustness to hyperparameter variation. We adopt $\alpha = 10$ and $\beta = 1.0$, which yield the best results.

5 Conclusion

We propose *Sali4Vid*, a framework that incorporates saliency-aware modeling via two complementary strategies. It consists of two key components: (1) *saliency-aware video reweighting*, which leverages timestamp annotations to compute continuous frame-level saliency weights, and (2) *semantic-based adaptive caption retrieval*, which captures meaningful scene transitions and retrieves more accurate captions aligned with these segments, suppressing irrelevant information. Extensive experiments on YouCook2 and ViTT demonstrate that *Sali4Vid* achieves state-of-the-art performance in both captioning and event localization. Moreover, this framework can be readily extended to a wide range of vision-language modeling tasks (Oh et al., 2024; Kim et al., 2025b; Cha et al., 2025; Kim et al., 2025d,c) beyond video captioning.

6 Limitation

Our model, *Sali4Vid*, achieves state-of-the-art performance on dense video captioning through annotation-based video reweighting and semantic difference-based adaptive caption retrieval. However, some limitations remain. The reweighting strategy relies on timestamp annotation, reducing applicability to weakly supervised or annotation-free settings. Moreover, as shown in Figure A.1 of the supplementary, the adaptive retrieval module may still yield noisy segments. In future work, we plan to explore supervision-free reweighting and enhance robustness of adaptive retrieval.

7 Acknowledge

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) RS-2025-25422680, Metacognitive AGI Framework and its Applications).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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

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

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Soda: Story oriented dense video captioning evaluation framework. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 517–531. Springer.

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

Shiping Ge, Qiang Chen, Zhiwei Jiang, Yafeng Yin, Liu Qin, Ziyao Chen, and Qing Gu. 2025. Implicit location-caption alignment via complementary masking for weakly-supervised dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3113–3121.

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*.

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

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

Dong-Jin Kim, Tae-Hyun Oh, Jinsoo Choi, and In So Kweon. 2024a. Semi-supervised image captioning by adversarially propagating labeled data. *IEEE Access*, 12:93580–93592.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. 2024b. 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.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. 2025a. Hicm^2: Hierarchical compact memory modeling for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4293–4301.

Si-Woo Kim, MinJu Jeon, Ye-Chan Kim, Soeun Lee, Taewhan Kim, and Dong-Jin Kim. 2025b. Sync: Synthetic image caption dataset refinement with one-to-many mapping for zero-shot image captioning. *arXiv preprint arXiv:2507.18616*.

Taewhan Kim, Soeun Lee, Si-Woo Kim, and Dong-Jin Kim. 2025c. Vipcap: Retrieval text-based visual prompts for lightweight image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4320–4328.

Ye-Chan Kim, SeungJu Cha, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2025d. Sida: Synthetic image driven zero-shot domain adaptation. *arXiv preprint arXiv:2507.18632*.

Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. 2019. Visil: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6351–6360.

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

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

Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2024. 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.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5085.

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. 2024. Preserving multimodal capabilities of pre-trained vlms for improving vision-linguistic compositionality. In *EMNLP 2024-2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 19060–19076. Association for Computational Linguistics (ACL).

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17959–17968.

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

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

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6847–6857.

Teng Wang, Huicheng Zheng, and Mingjing Yu. 2020. Dense-captioning events in videos: Sysu submission to activitynet challenge 2020. *arXiv preprint arXiv:2006.11693*.

Yiwei Wei, Shaozu Yuan, Meng Chen, Xin Shen, Longbiao Wang, Lei Shen, and Zhiling Yan. 2023. Mppnet: multi-perspective perception network for dense video captioning. *Neurocomputing*, 552:126523.

Hao Wu, Huabin Liu, Yu Qiao, and Xiao Sun. 2024. Dibs: Enhancing dense video captioning with unlabeled videos via pseudo boundary enrichment and online refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18699–18708.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

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

Appendix

In this Appendix, we provide additional details and qualitative results to support our findings. Specifically, §A offers an analysis of the semantic-based caption retrieval, and §B showcases further qualitative results of our model.

A More Analysis for Semantic-based Caption Retrieval

Figure A.1 provides examples of how our semantic-based adaptive thresholding identifies meaningful segment boundaries based on semantic differences across frames. In the first example (Video id: *igC0oJ48gxg*), our method successfully detects segment transitions that align well with visual changes in the cooking process, such as moving from chopping vegetables to frying. The detected peaks match the ground-truth timestamps, showing that the adaptive threshold (0.10) effectively captures event boundaries. In the second example (Video id: *_XxXWiOoYhY*), although some frames exhibit high semantic differences, our method correctly filters out false positives (red cross) that do not correspond to meaningful scene changes. The adaptive threshold (0.18) helps to focus on truly significant transitions, avoiding over-segmentation. These results demonstrate that our adaptive thresholding method dynamically adjusts to video content, effectively balancing sensitivity and precision in segment detection.

B More Qualitative Results.

We provide additional qualitative results in Figure A.2 to further illustrate the effectiveness of our method.

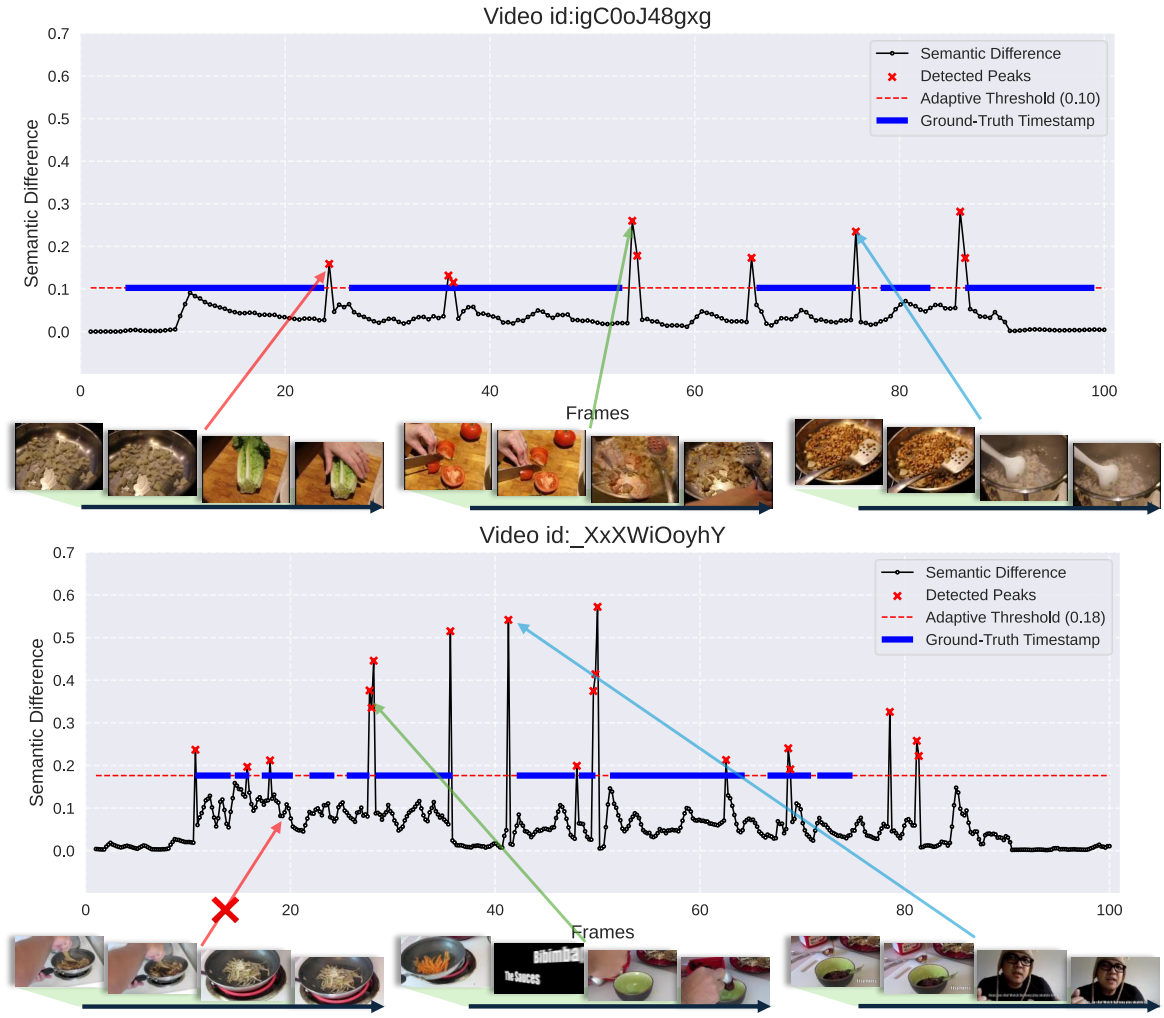

Figure A.1: Analysis for Semantic-Based Caption Retrieval.

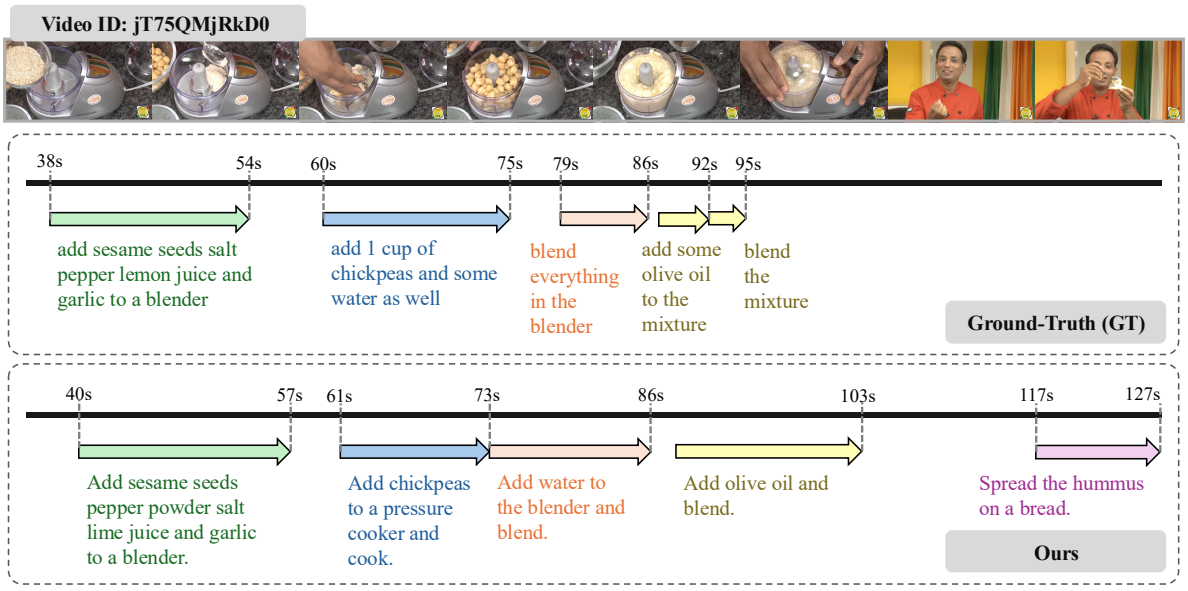

Figure A.2: More Qualitative Results for *Sali4Vid*.