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

Supplementary Material

A. Implementation Details

Backbone. We use CLIP ViT-L/14 [32] to extract visual features from video frames. For ActivityNet Captions, we uniformly sample 32 frames per video, while for YouCook2, we sample 100 frames. Text features are extracted using the CLIP text encoder.

LLM Settings. We employ the Qwen3-8B model [42] for synthetic caption generation with the following hyperparameters: maximum new tokens = 50, temperature = 0.7.

Event Queries. The number of event queries is set to 14 for ActivityNet Captions and 18 for YouCook2, respectively, to accommodate the different annotation densities of the two datasets.

Mask Generation Module. Following [11], our mask generation module consists of a transformer-based decoder that processes learnable event queries. Specifically, the transformer layers take the event queries as input and produce event-aware embeddings, which are then passed through separate fully-connected layers to predict the temporal center c_i and width w_i for each event. These values are used to construct Gaussian masks via Eq. (1).

B. Inference Details

At test time, our model employs the same inference procedure as [11] to generate temporally localized event captions. First, we obtain an initial set of event descriptions by feeding the complete video embedding into the caption decoder with a global context prompt ("[FULL]"). The model adaptively infers the number of events present in the video during this phase. Subsequently, we apply the mask generation module to predict temporal parameters (centers and widths) for each identified event, from which we derive Gaussian attention masks. Finally, we enhance caption quality through a refinement mechanism: each event's masked video representation is decoded using event-specific conditioning ("[MASK] 1 events:") to produce more precise and contextually appropriate descriptions.

B.1. Hyper-parameter Ablations.

We conduct additional hyper-parameter ablation experiments to evaluate captioning and localization performance under different values of α_{aug}. Figure 8 shows the captioning and localization scores under different values of α_{aug}. The hyperparameter α_{aug} serves to scale the auxiliary loss to a comparable magnitude with the main losses. The results demonstrate robust performance across different values, consistently outperforming the baseline [11] regardless of the specific weight choice. We select α_{aug} as 0.25.

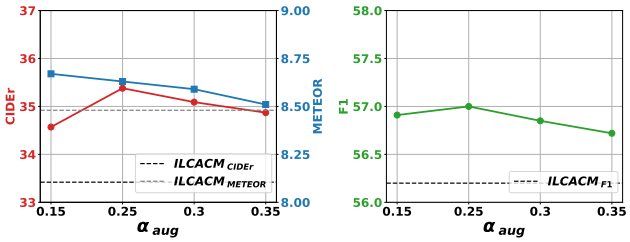

Figure 8. Impact of hyper-parameter α_{aug} for scaling $\mathcal{L}_{\text{aug}}$ on captioning and localization performance.

B.2. Qualitative Results.

In this section, we provide additional qualitative results to demonstrate the effectiveness of our approach. As shown in Figure 9, our method generates captions that are more closely aligned with ground-truth descriptions and achieves more accurate event localization compared to the baseline method [11].

B.3. Augmented Captions Quality.

We present additional qualitative results for LLM-generated captions in Figure 10. As shown, our LLM-generated captions effectively describe potential intermediate events such as "holding up a chart" and "playing the harmonica" that bridge consecutive ground-truth annotations. The generated captions are contextually coherent and semantically appropriate, demonstrating the LLM's capability to infer plausible transitional events from textual context.

B.4. Mask Quality.

As discussed earlier, the baseline method [11] focuses solely on ensuring that masks cover different temporal regions without considering semantic alignment with their corresponding events. If this is indeed the case, masks should exhibit similar width values to simply partition the video into distinct segments. To investigate this hypothesis, we calculate the standard deviation of mask widths across all videos in the training set and report the mean, minimum, and maximum standard deviation values in Table 9. As shown in the results, our similarity-aware mask construction generates masks with varying widths that accurately capture event-specific temporal characteristics.

Additionally, qualitative results in Figure 11 show that our approach produces event boundaries that align more

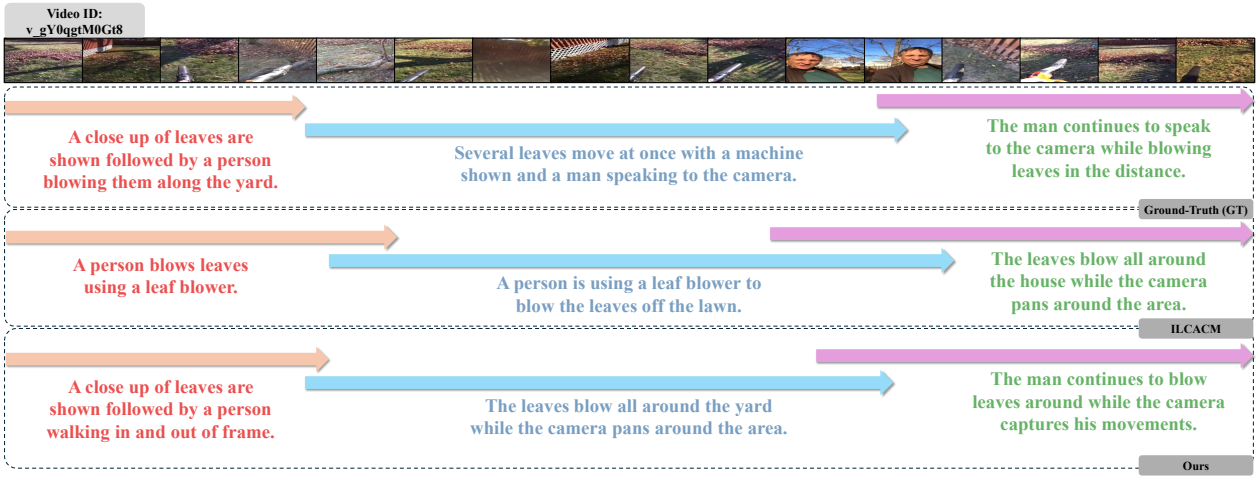

Figure 9. Another qualitative result from the ActivityNet validation set.

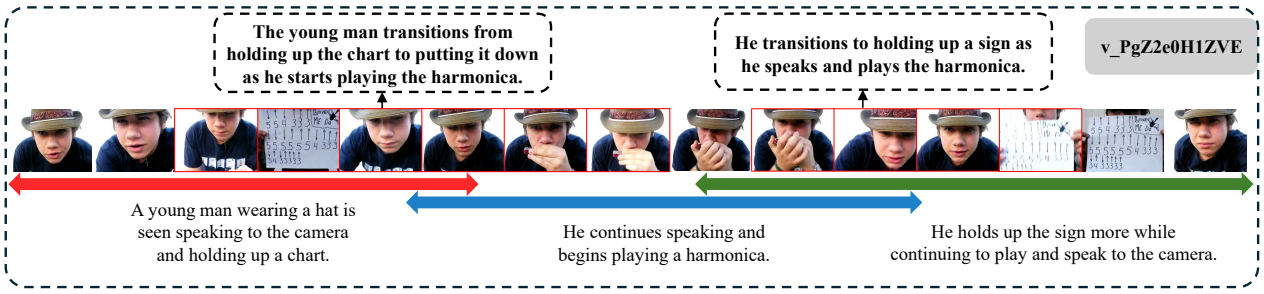

Figure 10. Qualitative results LLM augmentation captions. Our LLM-generated synthetic captions effectively describe the intermediate events occurring between consecutive ground-truth annotations.

Method	Mean	Min	Max
ILCACM [11]	0.3489	0.2690	0.3900
SAIL	**0.3535**	**0.2549**	**0.3914**

Table 9. Analysis of mask width diversity during training. We report the mean (Mean), minimum (Min), and maximum (Max) standard deviation of mask widths across all training videos.

closely with the non-uniform ground-truth timestamps compared to the baseline's uniform partitioning. The baseline method focuses primarily on distributing masks evenly across the video, resulting in masks with uniform widths regardless of actual event durations. For instance, the mask width for event 3 in ILCACM remains nearly constant throughout training epochs (0.3530 $\rightarrow$ 0.3530 $\rightarrow$ 0.3525 $\rightarrow$ 0.3507 $\rightarrow$ 0.3523), showing minimal adaptation to the event's actual temporal extent. In contrast, our method learns masks by considering semantic alignment with corresponding event captions, enabling adaptive width adjustment. As shown in Figure 11, the mask width for event 3 in our method progressively decreases during training (0.3529 $\rightarrow$ 0.3489 $\rightarrow$ 0.3467 $\rightarrow$ 0.3398 $\rightarrow$ 0.3281), demonstrating that it adapts to align with the event's shorter duration. These results validate that our similarity-aware guidance enables masks to learn event-specific temporal characteristics rather than converging to uniform partitioning, leading to more accurate event localization aligned with ground-truth boundaries.

C. Prompt Details

In Algorithm 1, we provide our instructions to the LLM for creating scene descriptions between consecutive events. We specify five instructions: (1) analyzing both preceding and succeeding captions, (2) inferring the most probable transitional action or change of state, (3) maintaining consistency with the original annotations, (4) ensuring the generated event is highly plausible, and (5) adhering to the specified output format. This structured approach helps the LLM concentrate on the specific temporal gap between events.

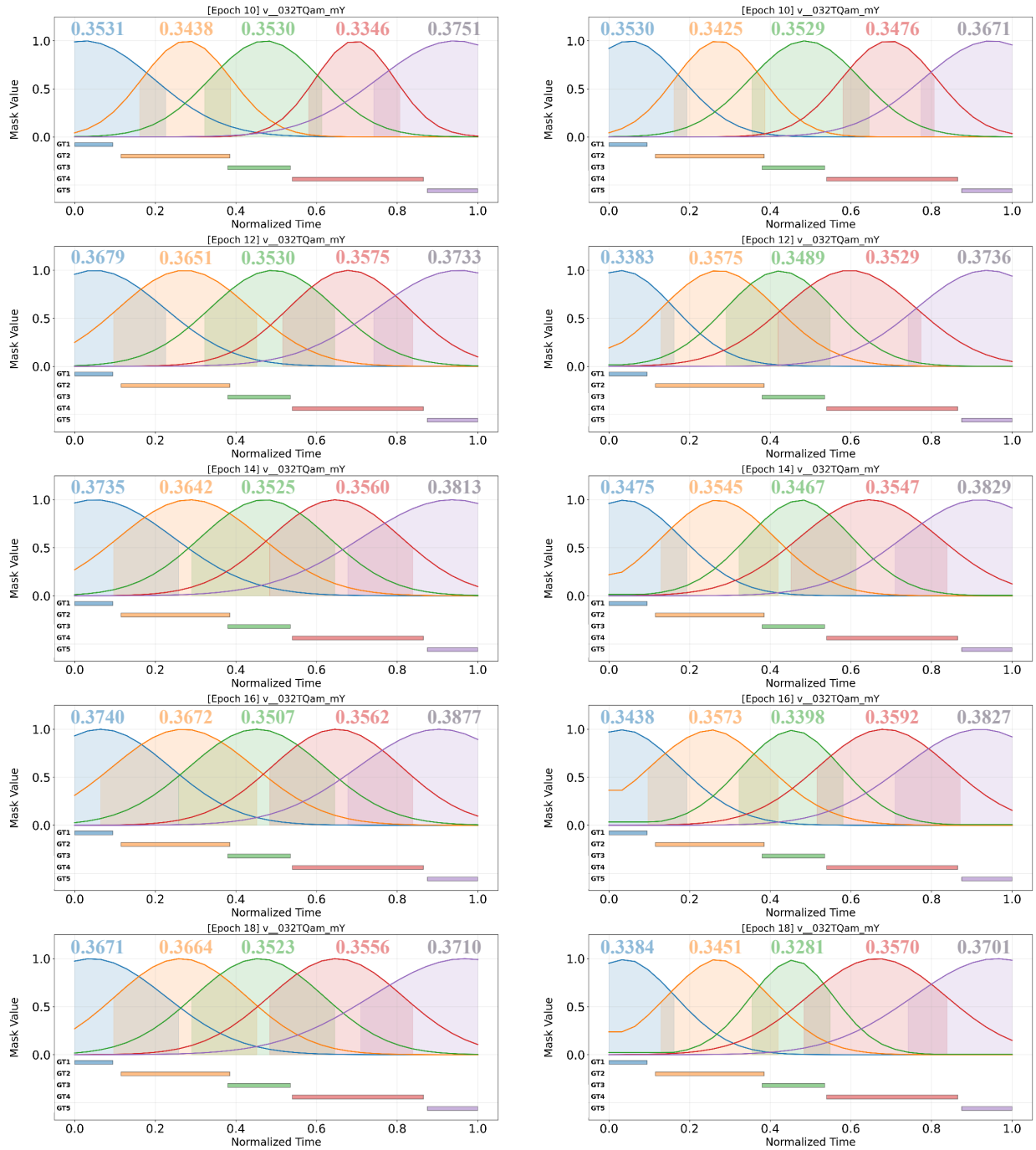

Figure 11. We visualize the mask outputs from the mask generation module during training and annotate each mask with its corresponding width value. ILCACM maintains almost constant mask widths throughout training, whereas *SAIL* shows progressive adaptation: masks for shorter events (blue, green) decrease in width, while masks for longer events (red) increase, aligning with actual event durations.

Algorithm 1: Our instruction prompts for LLM

SYSTEM PROMPT

You are a "Video Context Inference Expert," an AI specialized in analyzing sequences of video event captions. Your primary goal is to generate one new, plausible caption for the event that likely occurred *between* the two provided captions, creating a smooth and logical narrative flow.

1. **Context is Key**: Deeply analyze the preceding and succeeding captions. The generated caption must serve as a logical bridge, connecting the two events seamlessly.
2. **Infer the Unseen Action**: Do not simply rephrase or combine the given captions. Your task is to infer the most probable *transitional action* or *change of state*. Focus on the single most important action that connects the two moments.
3. **Maintain Consistency**: The style, tone, and level of detail of your generated captions should match the input captions. They should be concise, descriptive, and written from the perspective of an objective observer.
4. **Plausibility over Creativity**: The generated event must be highly plausible. Avoid introducing new elements that cannot be reasonably inferred.
5. **Output**: Provide ONLY the single generated caption text, without any labels or explanations.

USER PROMPT Analyze the following two consecutive video captions and generate a single, concise caption that describes the most plausible event happening between them.

Caption 1::{caption1}

Caption 2: :{caption2}

Based on the rules, what is the single event that connects these two captions?