This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Stay in your Lane: Role Specific Queries with Overlap Suppression Loss for Dense Video Captioning

Seung Hyup Baek[1*] Jimin Lee[2*] Hyeongkeun Lee[3] Jae Won Cho[1†]

[1]Konkuk University [2]Sejong University [3]KAIST

{edwardback, chojw}@konkuk.ac.kr, jjujimin@sju.ac.kr, lhk528@kaist.ac.kr

Abstract

Dense Video Captioning (DVC) is a challenging multimodal task that involves temporally localizing multiple events within a video and describing them with natural language. While query-based frameworks enable the simultaneous, end-to-end processing of localization and captioning, their reliance on shared queries often leads to significant multi-task interference between the two tasks, as well as temporal redundancy in localization. In this paper, we propose utilizing role-specific queries that separate localization and captioning into independent components, allowing each to exclusively learn its role. We then employ contrastive alignment to enforce semantic consistency between the corresponding outputs, ensuring coherent behavior across the separated queries. Furthermore, we design a novel suppression mechanism in which mutual temporal overlaps across queries are penalized to tackle temporal redundancy, supervising the model to learn distinct, non-overlapping event regions for more precise localization. Additionally, we introduce a lightweight module that captures core event concepts to further enhance semantic richness in captions through concept-level representations. We demonstrate the effectiveness of our method through extensive experiments on major DVC benchmarks YouCook2 and ActivityNet Captions. Code is available here.

1. Introduction

With the explosive growth of online video content, automatic video understanding [15, 16, 21, 22, 27, 37, 42, 46, 47, 53] has emerged as a central challenge in computer vision and natural language processing. Through these efforts, video captioning [6, 12, 30, 31, 40] has drawn significant attention for its ability to translate video content into natural language.

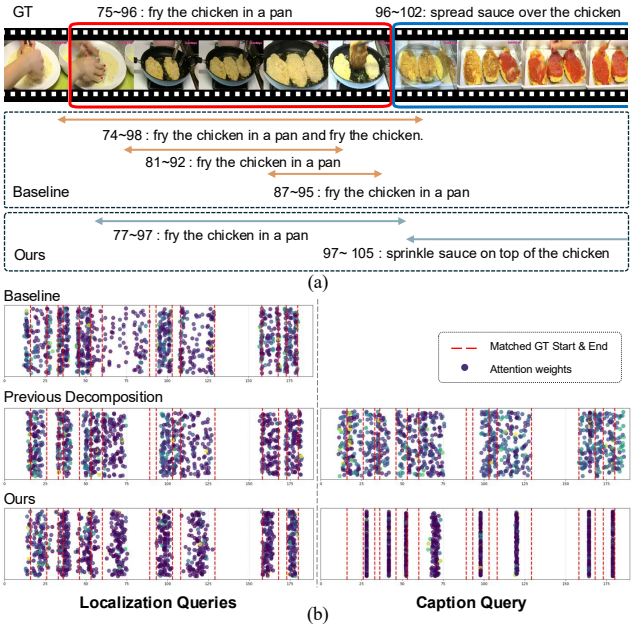

Figure 1. Comparison of our method with previous methods. (a) Captioning results. The baseline model captures overlapping events, causing redundant captions. In contrast, our model generates distinct non overlapping regions for a set caption. (b) Decoder query sampled attention weights. The baseline's single query attends indiscriminately. Previous "decomposition" queries show similar attention distributions. In contrast, our method employs two specialized queries: localization queries attend broadly for boundaries and a caption queries attend densely on key frames.

Most conventional video captioning approaches focus on short, trimmed clips. However, these methods struggle to generalize to real-world videos, which are often long, untrimmed, and contain multiple events. To address this limitation, Dense Video Captioning (DVC) [19] was introduced as a task. DVC requires not only the precise temporal localization of all salient events but also the generation of coherent textual descriptions for each localized segment.

Traditional DVC methods have adopted a two-stage

*Equal contribution
†Corresponding author

"localize-then-describe" approach [14, 19, 23, 41], which often suffered from not being end-to-end trainable and led to suboptimal performance from the lack of interaction between the two seminal tasks. Subsequently, PDVC [44] first leveraged the DETR [3] network and introduced a unified query-based framework where a fixed set of learnable queries jointly predict event segments and generate captions through parallel decoding. This design enabled end-to-end joint optimization of both tasks and significantly simplified the DVC pipeline. Building upon this paradigm, subsequent research [18, 24, 48, 49] has continued to improve performance within this DETR-based architecture.

However, utilizing a shared set of learnable queries for both localization and captioning simultaneously hinders the model's ability to capture the distinct characteristics of each task. Training a single query to handle both objectives has been shown to be a possible cause of task interference in general multi-task learning contexts, where the shared representation space leads to conflicting optimization directions [5, 8, 23, 32, 50]. We observe a similar issue in query-based DVC models. As shown at the top of Fig. 1(b), the baseline PDVC [44] exhibits ambiguous query behaviors, with attention weights that fail to focus either on precise event boundaries or on semantically coherent regions for captioning. Although DDVC [24] introduces query decomposition, we observe that, in the center of Fig. 1(b), the attention weights of its localization and caption queries appear highly similar, as they do not create separate queries but instead utilize an MLP, suggesting that proper task separation has not been effectively achieved. In addition, as shown in Fig. 1(a), we observe that queries tend to repeatedly capture similar temporal segments and scenes, generating redundant captions. This behavior degrades both localization and captioning performance.

To overcome these challenges, we argue that a more direct supervision is necessary to properly train the query learning process. To achieve this, we assert that queries must: 1) resolve task interference through clear task separation and 2) learn to avoid overlapping temporal regions. However, task separation can lead to misalignment between the two query types. Therefore, the queries must not only be separated and trained to be non-overlapping, but also be trained to be semantically coherent across the separated tasks through a cross-task supervision loss.

In this paper, we propose **R**ole Specific Query with **O**verlap **S**uppression **D**ense **V**ideo **C**aptioning (**ROS-DVC**), which tackles the aforementioned issues through three core techniques. We introduce role specific queries, where we separate localization and caption queries into independent query sets and initialize them each from separate learnable embedding spaces. In our role specific queries, as shown at the bottom of Fig. 1(b), the localization queries broadly attend to temporal context to predict boundaries, while the caption queries focus intensively around a key region to describe its semantics in detail. To ensure semantic alignment between these decoupled representations, we introduce the Cross-Task Contrastive Alignment (CTCA) Loss. This loss ensures the coupling between the two tasks by aligning localized events and captions together. To mitigate the query overlap problem, we introduce a novel Overlap Suppression Loss that explicitly penalizes query pairs that overlap excessively based on pairwise IoU between the queries. This encourages the localization queries to focus on distinct regions and ensure precise localization. To further enhance the semantic expressiveness of captions, we introduce a Concept Guider, a lightweight MLP-based concept head that captures the underlying concepts [25] of events. This simple yet effective module enriches the caption queries with concept-level representations, enabling the model to generate captions that better reflect the semantic understanding of each event.

Our main contributions can be summarized as follows:

- We resolve task interference in DVC by utilizing independent, role specific queries trained with our CTCA Loss.
- We design an Overlap Suppression Loss that discourages unnecessary query overlaps, enabling the model to learn more accurate temporal localizations.
- We introduce a Concept Guider that enhances the expressiveness and semantic accuracy of generated captions without utilizing an external memory bank.
- We achieve competitive performance on DVC benchmarks YouCook2 [54] and ActivityNet Captions [19].

2. Related Works

Dense Video Captioning (DVC) [19] is a multimodal task comprising two sub-tasks: event localization and captioning. Early works adopt numerous approaches to tackle this problem [1, 4, 14, 34, 36, 41, 43]. The pioneering work of DVC [19] has utilized a two-stage "localize-then-describe" approach [14, 41, 43], where it first identifies the temporal location of an event and then feed its boundaries into a separate captioning head to generate a description. However, Wang *et al*. [44] have criticized this line of research for its heavy reliance on anchor-based designs to improve localization and the inherent lack of mutual interaction between the two sub-tasks. To address these limitations, later studies began to explore end-to-end frameworks [17, 28, 34, 51, 55] that jointly optimize event localization and captioning within a unified architecture, allowing the two sub-tasks to interact more closely during training and inference. Due to the similarity of object detection's task structure and that of DVC, both of which involve the two core components of localization and semantic understanding, [44] first adopted the DETR architecture to improve efficiency by processing localization and captioning concurrently through a parallel decoding framework. Sub-

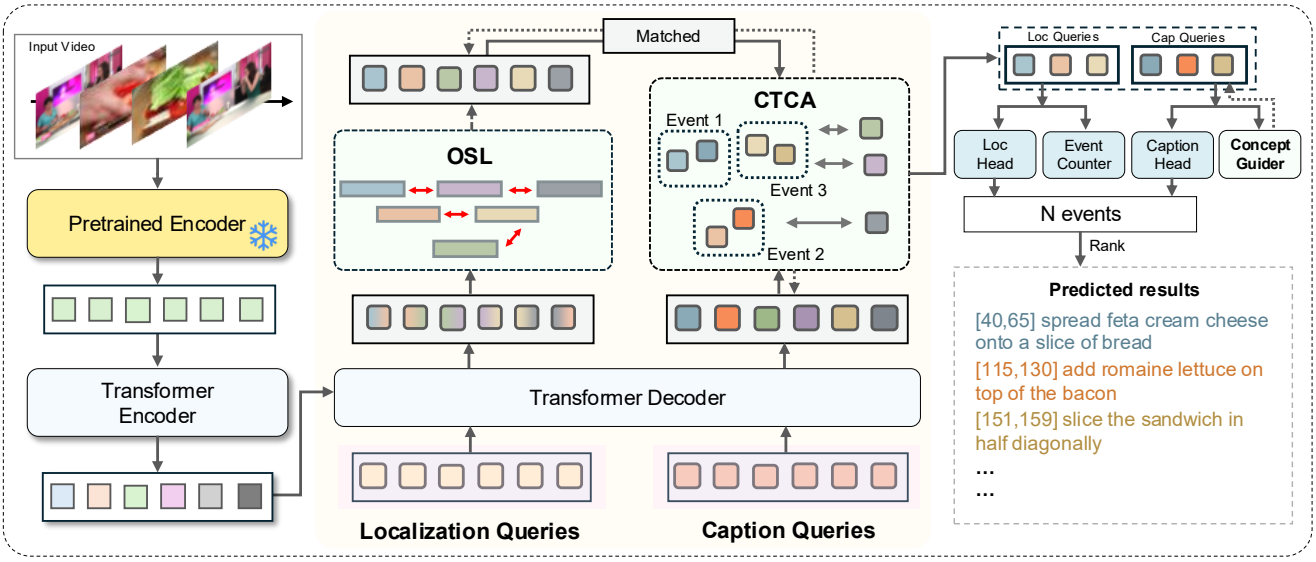

Figure 2. **An overview of our proposed ROS-DVC framework.** The input video is first fed into the pretrained encoder, and a transformer encoder processes it to generate frame-level features. In the decoding stage, two types of queries are independently initialized and retrieve their role-specific information from the frame-level features. The output localization queries are trained with the Overlap Suppression Loss to minimize mutual overlap and are matched with ground truths via the Hungarian algorithm. Subsequently, the CTCA loss is employed to semantically align the caption queries with their corresponding localization queries. Finally, these processed queries are fed into respective heads to obtain the predictions for event-number, localize timestamps, event captions, and event-level concepts.

sequent methods [18, 24, 48, 49] have improved upon the DETR-based end-to-end pipeline.

Among recent methods, a line of works has sought to improve the captioning ability of DVC models through a Retrieval-Augmented Generation method, where an external memory bank is used to better align textual and visual modalities for more diverse captions. CM^2 [18] first proposed the use of an external memory bank, and MCCL [49], pointed out the need for synergy between the captioning and localization heads and proposed a cyclic co-learning strategy. Another line of work focused on improving the representation of queries. E^2DVC [48] focused on query initialization to mitigate data bias and ensure a more uniform detection of events, while DDVC [24] employed query decomposition through an MLP to make the captioning and localization features more task-specific.

Most recently, Chen et al. [7] further explore query decomposition by introducing an additional learnable encoder prior to query initialization to enrich query representations. By enhancing query representations through encoder processing, their approach aims to provide more informative features for downstream localization and captioning. In contrast, our approach does not rely on additional encoders and instead emphasizes explicit query separation at initialization, focusing on improving the learning of task-specific queries themselves. We place our work in this line of research, where we tackle DVC through a general query-based framework without the use of an external memory bank. Our approach enhances the learning of queries themselves by introducing independent task-specific queries and novel loss functions that enable adaptive and disentangled optimization of localization and captioning.

Overlap Suppression. Localization tasks often suffer from overlapping predictions, which are typically mitigated through Non-Maximum Suppression (NMS). However, more complex tasks like crowd-scene detection adopt various additional regularization methods. Wang *et al*. [45] proposed the repulsion loss to push the predicted bounding box farther away from other predicted bounding boxes and other ground truths. Additionally, AutoPedestrian [38] introduced an automatic scheme that jointly searches for optimal data augmentation policies and loss function parameters, aiming to improve pedestrian detection under crowded conditions. Gao *et al*. [11] identified the problem of excessive false-positive predictions in the DETR architecture and proposed a learnable sample selection strategy applied to the matching cost and loss function. While DVC adopts the DETR architecture and does not use NMS, it still suffers from excessive overlapping predictions. Drawing inspiration from these approaches, we redesign the loss for our task, while employing a DETR-based architecture despite the task differences. To address temporal overlap, we propose our Overlap Suppression Loss that leverages the IoU between proposed events and their ground truths, penalizing overlapping proposals and promoting more precise event boundaries.

3. Methodology

Our method aims to enhance both event localization and captioning by supervising the learning process to capture the distinct characteristics of both detection and captioning. To this end, we introduce three key components: Role Specific Query 3.2, Overlap Suppression Loss 3.3, and Concept Guider 3.4, which address task-aware query optimization, suppression of redundant event proposals, and concept-level caption reasoning respectively. Built upon a parallel encoder–decoder architecture, our framework, the **ROS-DVC** method, employs separate localization and caption queries within a shared decoder to jointly refine temporal boundaries and event descriptions.

3.1. Overall Framework

Overall, the DVC model first encodes an input video using a pre-trained visual encoder for frame features and a transformer encoder for temporal context. In the decoding stage, a set of learnable queries interacts with the encoded frame features to generate N event-caption pairs, $\{(t_s^n, t_e^n, \text{Cap}^n)\}_{n=1}^{N}$, where N denotes the number of events detected in the video. During training, the Hungarian algorithm [20] is employed to find optimal one-to-one matchings between N generated pairs and the ground-truth events. These queries function as learnable representations and act as event proposal candidates. Each query is associated with a reference point, an initial guess for the event's features and location. Both the query features and reference points are then iteratively refined via cross-attention across decoding layers, progressively improving the precision of event locations and descriptions.

3.2. Role Specific Queries

3.2.1. Role Specific Query Initialization

Unlike previous works that use a single set of queries for both sub-tasks, we intend localization queries to capture broad temporal context for precise event boundary estimation, whereas caption queries focus on fine-grained details for accurate visual description. Hence, we explicitly separate the query space into two sets, each specialized for *localization* and *captioning*, respectively.

$$\{q_{\text{loc}}^j\}_{j=1}^{K}, \quad \{q_{\text{cap}}^j\}_{j=1}^{K}. \tag{1}$$

Here, $\{q_{\text{loc}}^j\}$ denotes K localization queries, and $\{q_{\text{cap}}^j\}$ denotes K caption queries. Unlike [24], which generates caption queries from localization queries through an MLP, creating dependencies and limiting the ability to learn more diverse attention distributions, our approach initializes each query set from its own separate embedding space. This allows each query to be independently optimized for its own role specific objective while reducing dependencies.

Although the queries are initialized from separate embedding spaces, they maintain complementarity by referencing the same visual location, as they are injected together into the cross-attention layers. This design ensures that the queries can learn rich semantic representations independently while still being grounded in the relevant visual context defined by the 'localization' query. This allows the two tasks to learn at different temporal abstraction levels without conflict and converge to their respective roles.

3.2.2. Cross-Task Contrastive Alignment

Following our Role Specific Query Initialization, the model employs localization queries to precisely estimate event boundaries, and N events are selected via the Hungarian algorithm. Next, each of these localization queries must be paired with a semantically matched caption query. However, because caption queries are specialized for capturing semantic meaning and are also initialized from a different embedding space, their semantic consistency with its corresponding localization queries might not be guaranteed.

To address this issue, we introduce a Cross-Task Contrastive Alignment (CTCA) Loss that bridges the gap between the two obtained queries. The output query features are denoted as $\tilde{q}_{\text{loc}}$, $\tilde{q}_{\text{cap}}$. In CTCA, the Hungarian algorithm first identifies a set of indices $\mathcal{M}$ corresponding to the ground-truth events. For each $j \in \mathcal{M}$, the j-th caption query ($\tilde{q}_{\text{cap}}^j$) and the j-th localization query ($\tilde{q}_{\text{loc}}^j$) are treated as positive pairs. The j-th caption query is then contrasted with all other localization queries ($\tilde{q}_{\text{loc}}^{j'}$, where $j' \neq j$) as negatives. This formulation encourages localization queries to be pulled closer to their corresponding caption representations (reinforcing event semantics) while being pushed away from unrelated ones. This, in turn, enables the localization query to effectively learn semantic cues from the caption query. The CTCA Loss is defined as follows:

$$\mathcal{L}_{\text{CTCA}} = -\sum_{j \in \mathcal{M}} \log \frac{\exp(\text{sim}(\tilde{q}_{\text{cap}}^j, \tilde{q}_{\text{loc}}^j)/\tau)}{\sum_{j'} \exp(\text{sim}(\tilde{q}_{\text{cap}}^j, \tilde{q}_{\text{loc}}^{j'})/\tau)}, \tag{2}$$

Where $\text{sim}(\cdot)$ denotes cosine similarity and τ is the temperature parameter. Through this asymmetric contrastive objective, the localization queries acquire semantic awareness from their corresponding captions. Consequently, the two query sets learn in distinct yet cooperative embedding spaces, one capturing event semantics and the other focusing on temporal structure, leading to more coherent and event-consistent dense video captions.

3.3. Overlap Suppression Loss (OSL)

To detect more precise event boundaries, we introduce the Overlap Suppression Loss (OSL), a loss that induces suppression of the distribution of predicted event boundaries. This loss is designed to solve two key challenges: first, to

prevent the model from producing excessively overlapping and redundant event predictions, and second, to discourage it from collapsing all boundaries into a non-informative solution. This helps reduce excessive overlap predictions

The foundation of our loss is the overlapping measure P_o, which we define as the temporal IoU [35] between any two distinct predicted boundaries, B_i and B_j:

$$P_o(i,j) = \text{IoU}(B_i, B_j), \quad i \neq j. \tag{3}$$

Minimizing P_o alone gives an aggressive penalty on all overlaps and could inadvertently suppress multiple correct predictions that are close to one another, thereby limiting the model's ability to identify genuine events. Therefore, relying on it alone can be insufficient.

To address this, we incorporate a ground-truth–aware weighting factor, α, which adaptively modulates the overlap penalty based on the predicted alignment with the ground truth. We first quantify this alignment by calculating the ground-truth overlap P_g, defined as the IoU between a predicted boundary B_i and the ground-truth boundary G_j:

$$P_g(i,j) = \text{IoU}(B_i, G_j). \tag{4}$$

We then define α as the combination of the alignment score P_g with its complement $(1 - P_g)$, where the balance between the two is controlled by a hyperparameter γ. Thus, α is defined as:

$$\alpha = \gamma \cdot P_g + (1-\gamma) \cdot (1 - P_g). \tag{5}$$

We let $\gamma \leq 0.5$. Given this constraint, a higher P_g leads to a smaller α, thereby decreasing the penalty when the prediction is well aligned with the ground truth. Conversely, when P_g decreases, α becomes larger, resulting in a stronger penalty, effectively discouraging the model from predicting overlapping predictions. The final OSL formulation is:

$$L_{\text{OSL}} = -\alpha \cdot \log(\beta - P_o). \tag{6}$$

With β being a hyperparameter that defines the maximum overlap. α allows the model to distinguish between relevant and redundant overlaps, as illustrated in Fig. 3. This mechanism achieves a critical balance between reducing redundancy and ensuring accurate localization, leading to a more precise set of event predictions.

3.4. Concept Guider

Previous research has shown that utilizing concepts can enhance semantic richness of generated captions in DVC [25, 26, 49, 52]. Hence, we introduce the Concept Guider, a lightweight MLP auxiliary head designed to output an event-level multi-hot vector $\hat{y}_c$, that assists the expressive power of the caption queries.

$$\hat{y}_c = sigmoid(MLP(\tilde{q}_{\text{cap}})) \tag{7}$$

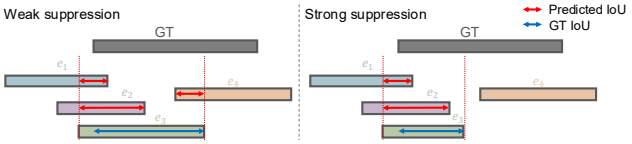

Figure 3. **Mechanism of the Overlap Suppression Loss.** This loss discourages overlap among queries. The suppression strength is inversely modulated by the query's IoU with the ground truth (GT). Queries with high GT IoU receive weak suppression, while queries with low GT IoU receive strong suppression.

This head takes the final embedding of the caption query $\tilde{q}_{\text{cap}}$ as input and is trained with an auxiliary objective that encourages the query representation to internalize the concepts associated with each event. Specifically, we extract the top N_c nouns and verbs from training captions, and define them as concepts. Then, for each event, a multi-hot concept label $Y^c \in \{0,1\}^{N_c}$ is constructed, where each element is set to 1 if the relevant concept appears in the ground-truth captions, and 0 otherwise. Through this auxiliary task, the caption query is guided to embed high-level semantics and implicit intentions within its representation, moving beyond simple visual features. Consequently, the Concept Guider enriches the query embedding, helping the main caption generator produce more specific and contextually aware sentences. This auxiliary head is trained using the Cross-Entropy Loss and is unused during inference.

3.5. Task Heads

Our method employs a role specific query based framework, in which the decoder's final representations are distributed across four specialized heads including the Concept Guider in section 3.4. Each head is responsible for a distinct component of dense video captioning.

Localization Head predicts temporal boundaries and assigns confidence scores for each localization query $\tilde{q}^i_{\text{loc}}$. The output is a set of tuples $\{t^s_i, t^e_i, c_i\}_{i=1}^{N_q}$, where t^s_i and t^e_i represent start and end times, and c_i is localization confidence.

Captioning Head is based on an LSTM [13]. At each time step t, the LSTM receives context features $z_{i,t}$, caption query $\tilde{q}^i_{\text{cap}}$, and sequence of previously generated words $\{w_{i,j}\}_{j=1}^{t-1}$. Then, it outputs the current word $w_{i,t}$. Through repetition, the complete description is generated as $S_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,T}\}$, where T is the sentence length.

Event Counter counts the number of events in the video. The localization query $\tilde{q}^i_{\text{loc}}$ is transformed using an MLP to r^{len}, which represents the predicted distribution over possible event counts. During inference, the estimated number of events is obtained as $N = \text{argmax}(r^{len})$. Subsequently, the N predicted sets $\{(t^s_n, t^e_n, S_n)\}_{n=1}^{N}$ are selected via the Hungarian algorithm [20] using a matching cost defined as $C = \mathcal{L}_{\text{cls}} + \alpha_{\text{giou}}\mathcal{L}_{\text{giou}} + \alpha_{\text{cap}}\mathcal{L}_{\text{cap}}$. The focal loss L_{cls} measures the difference between the predicted classification

scores and the ground truth labels, while the generalized IoU loss L_{giou} measures the discrepancy between the predicted boundary and ground truth temporal segments. The caption loss L_{cap} evaluates the quality of the generated captions through cross-entropy. This extended matching strategy allows caption queries to directly influence the pairing stage, ensuring that each localization & caption pair is both temporally accurate and semantically consistent.

3.6. Loss Function

For the standard DVC losses, we use $\mathcal{L}_{\text{giou}}$, $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{cap}}$, and $\mathcal{L}_{\text{ec}}$, where $\mathcal{L}_{\text{ec}}$ is the cross-entropy loss between the predicted event count distribution and the ground-truth count. In addition, we add $\mathcal{L}_{\text{CTCA}}$, $\mathcal{L}_{\text{OSL}}$, and $\mathcal{L}_{\text{CG}}$, with $\mathcal{L}_{\text{CG}}$ defined as the cross-entropy loss between the event-level concept labels and the predicted concept distribution.

Therefore, our final loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{cap}}\mathcal{L}_{\text{cap}} + \lambda_{\text{ec}}\mathcal{L}_{\text{ec}} \\ &+ \lambda_{\text{CTCA}}\mathcal{L}_{\text{CTCA}} + \lambda_{\text{OSL}}\mathcal{L}_{\text{OSL}} + \lambda_{\text{CG}}\mathcal{L}_{\text{CG}}\end{aligned} \tag{8}$$

where each λ is a hyperparameter that balances the contribution of its respective loss term.

4. Experiments

4.1. Experimental settings

Datasets. We evaluate our proposed method on two widely used benchmark datasets for Dense Video Captioning (DVC): YouCook2 [54] and ActivityNet Captions [19]. YouCook2 consists of 2k untrimmed videos of cooking procedures from 89 distinct recipes. The videos have an average length of about 320 seconds and contain an average of 7.7 annotated segments, each with a corresponding sentence description. For our experiments, we adhere to the official split of 1,333 videos for training, 457 for validation, and 210 for testing. ActivityNet Captions contains 20k untrimmed videos covering a wide range of human activities, with an average duration of approximately 120 seconds. Each video is annotated with an average of 3.65 temporally localized sentences, resulting in a total of over 100k annotations. We follow the official data split, using 10,009, 4,925, and 5,044 videos for training, validation, and testing, respectively. We only utilize videos that remain accessible on YouTube, which accounts for approximately 7% fewer videos than the original dataset splits.

Implementation Details. For both datasets, video frames are sampled at 1 FPS. The resulting frame sequences are then subsampled or padded to a fixed length F, which is set to 200 for YouCook2 and 100 for ActivityNet. We employ CLIP ViT-L/14 [9, 33] to extract visual features from the video frames, following the standard protocol used in previous works [18]. Following prior DVC frameworks, we adopt a 2-layer deformable transformer decoder with multi-scale deformable attention across four feature levels. The number of localization and captioning queries are set to 50 each for YouCook2 and 10 each for ActivityNet Captions. For our proposed modules, we set the hyperparameters $\gamma = 0.25$, $\beta = 1.0$, and $N_C = 30$.

Table 1. **YouCook2 Caption metrics. Bold** represents best and underline represents second best performance. We do not directly compare to DDVC for fair comparison as they utilize GPT2 as their captioning head.

Method	#PT	CIDEr↑	METEOR↑	BLEU4↑	SODA_c↑
Vid2Seq	✓	47.10	9.30	–	7.90
DIBS	✓	44.40	7.51	–	6.39
DDVC	✗	38.75	6.92	1.92	6.68
PDVC	✗	29.69	5.56	1.40	4.92
CM^2	✗	31.66	6.08	1.63	5.34
MCCL	✗	36.09	**6.53**	2.04	5.21
E^2DVC	✗	34.26	6.11	1.68	5.39
Ours	✗	**39.18**	6.14	**2.10**	**7.06**

Table 2. **ActivityNet Captions metrics. Bold** represents best and underline represents second best performance. We do not directly compare to DDVC for fair comparison as they utilize GPT2 as their captioning head.

Method	#PT	CIDEr↑	METEOR↑	BLEU4↑	SODA_c↑
Vid2Seq	✓	30.10	8.50	–	5.80
DIBS	✓	31.89	8.93	–	5.85
DDVC	✗	35.48	8.62	2.44	6.55
PDVC	✗	29.97	8.06	2.21	5.92
CM^2	✗	33.01	8.55	2.38	6.18
MCCL	✗	34.92	**9.05**	**2.68**	6.16
E^2DVC	✗	33.63	8.57	2.43	6.13
Ours	✗	**35.04**	8.45	2.36	**6.45**

Evaluation metrics. We evaluate our method on two subtasks within DVC: captioning performance and event localization. Following standard practice, all metrics are averaged over IoU thresholds of $\{0.3, 0.5, 0.7, 0.9\}$ to ensure a robust evaluation across various degrees of temporal overlap. For captioning performance, we use the official ActivityNet Challenge evaluation tool to compute scores for CIDEr [39], METEOR [2], BLEU-4 [29], and SODA_c [10]. For event localization, we evaluate performance using average precision, average recall, and F1-score, which is the harmonic mean of precision and recall.

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

Captioning Performance. As shown in Table 1 and Table 2, our method achieves strong captioning results on

both datasets. On YouCook2, our model surpasses all non-pretrained methods, achieving the best CIDEr and SODA_c scores. Compared to MCCL which utilizes an external memory bank to enhance captioning diversity, our model is able to improve the CIDEr score by 3.09, BLEU4 by 0.06, and SODA_c by 1.85. Even for METEOR, our method only lags behind by 0.39. In addition, compared to the latest non-memory bank method E^{2}DVC, our method surpasses it in CIDEr by 4.92. While DDVC [24] employs a GPT-2 language model for caption generation, making it not directly comparable to ours, we still achieve comparable or better scores without utilizing an LLM based model, even surpassing it in CIDEr by 0.43. This demonstrates that our role-specific queries and Overlap Suppression Loss enable the caption head to produce coherent and semantically rich sentences purely from visual cues.

On ActivityNet, our model performs favorably, and even achieving the best CIDEr and SODA_c scores, surpassing MCCL in CIDEr by 0.12 and SODA_c by 0.32. These consistent gains confirm that our method contributes to precise temporal grounding and caption generation.

Event Localization Performance. We show localization results on YouCook2 and ActivityNet in Table 3. On YouCook2, our model surpasses all non-pretrained methods, achieving the best Recall, Precision, and F1 scores. Compared to the most recent non-memory bank method, E^{2}DVC, our method surpasses on Recall by 4.98, Precision by 0.51 and F1 by 3.39. On ActivityNet, our model's Recall surpasses E^{2}DVC by 0.68. We also find that, unlike previous methods where Recall is consistently lower than Precision, our model yields nearly matched Recall and Precision. Since Recall is computed as predicted events over the number of ground-truth events, and Precision as predicted events over the number of model's estimated events, this indicates that our event counter produces the number of events much closer to the ground truth.

4.3. Ablation Experiments

We conduct ablation studies to evaluate the impact of each component and design choice in our framework. All experiments are performed on the YouCook2 dataset with PDVC [44] set as our baselines unless otherwise specified.

Effect of Each Component. As shown in Table 4, we ablate each component of our model, Role Specific Query Initialization (RSQI), Cross-Task Contrastive Alignment (CTCA), Overlap Suppression Loss (OSL), and Concept Guider (CG). Using only RSQI allows each query to better specialize for its corresponding task, which in turn leads to noticeable improvements in both CIDEr and F1. When only OSL is applied, the provided query guidance significantly boosts the F1 score. Adding CG alone enhances the expressiveness of the generated captions, thereby improving the CIDEr score. Finally, our model with all components achieves the best performance across all metrics.

Table 3. **Localization metrics on YouCook2 and ActivityNet. Bold** represents best and underline represents second best performance. We do not directly compare to DDVC for fair comparison as they utilize GPT2 as their captioning head.

Method	#PT	YouCook2			ActivityNet		
		Rec.↑	Pre.↑	F1↑	Rec.↑	Pre.↑	F1↑
Vid2Seq	✓	27.90	27.80	27.84	52.70	53.90	53.29
DIBS	✓	26.24	39.18	31.43	53.14	58.31	55.61
DDVC	✗	30.81	37.25	33.73	54.77	57.54	56.12
PDVC	✗	22.89	32.37	26.81	53.27	56.38	54.78
CM^2	✗	24.76	33.38	28.43	53.71	56.81	55.21
MCCL	✗	–	–	–	53.19	57.36	55.23
E^{2}DVC	✗	24.36	34.75	28.64	54.67	**57.70**	**56.14**
Ours	✗	**29.34**	**35.26**	**32.03**	**55.35**	55.65	55.50

Table 4. **The ablation results of different components.** RSQI denotes *Role Specific Query Initialization*, CTCA denotes *Cross-Task Contrastive Loss*, OSL denotes *Overlap Suppression Loss*, and CG denotes *Concept Guider*. Note that CTCA cannot be applied when RSQI is not used.

RSQI	CTCA	OSL	CG	C↑	M↑	B@4↑	S↑	F1↑
✗	✗	✗	✗	29.69	5.56	1.40	5.39	26.81
✓	✗	✗	✗	32.33	5.66	1.66	5.43	27.00
✗	✗	✓	✗	33.60	5.41	1.46	6.79	31.22
✗	✗	✗	✓	31.40	5.58	1.58	5.62	27.69
✓	✓	✗	✗	34.48	5.82	1.88	5.58	27.59
✓	✗	✓	✗	32.72	5.60	1.49	6.73	31.98
✓	✗	✗	✓	33.36	5.63	1.63	6.03	28.94
✗	✗	✓	✓	33.11	5.32	1.58	6.91	31.63
✓	✗	✓	✓	34.92	5.66	1.62	6.83	30.58
✓	✓	✓	✗	36.63	5.80	1.87	6.95	30.62
✓	✓	✗	✓	36.93	5.73	2.02	6.05	28.59
✓	✓	✓	✓	**39.18**	**6.14**	**2.10**	**7.06**	**32.03**

Table 5. **Ablation results of varying γ on OSL.** γ is a hyperparameter that adjusts the influence of IoU with respect to the ground truth.

γ	CIDEr↑	SODA_c↑	Rec.↑	Pre.↑	F1↑
0.15	37.00	6.69	29.65	**35.44**	32.28
0.25	**39.18**	7.06	29.34	35.26	32.03
0.35	37.49	**7.16**	**30.80**	35.28	**32.88**
0.45	35.76	6.94	30.40	34.43	32.28

Effect of Overlap Suppression Loss. To ablate the effect of OSL, we set γ from 0.15 to 0.45. γ is a hyperparameter that adjusts the influence of IoU with the ground truth. As shown in Table 5, setting γ to 0.25, which is the middle of the range, we achieve a good balance among different met-

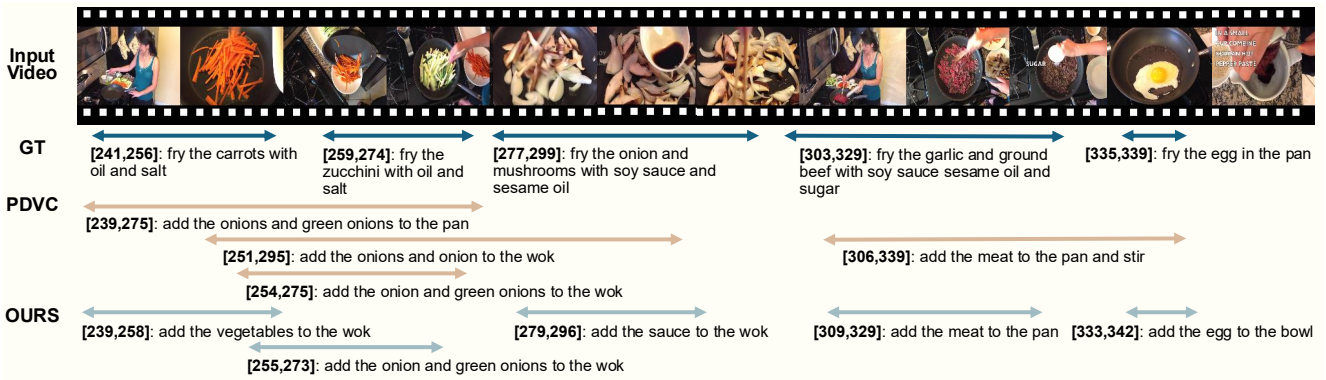

Figure 4. **Qualitative results of dense video captioning on YouCook2.** We compare the localization and captioning results with the ground truth, the baseline (PDVC), and ours. Each arrow represents localization boundaries and its corresponding caption is provided below. We find that while PDVC captures redundant events with the same captions, our model avoids this phenomenon.

Table 6. **Ablation results with and without α.** α controls whether the IoU is compared with the ground truth in OSL.

α	CIDEr↑	SODA_c↑	Rec.↑	Pre.↑	F1↑
W/o α	34.81	**7.20**	**31.01**	33.82	**32.35**
W/ α	**39.18**	7.06	29.34	**35.26**	32.03

Table 7. **Ablation results varying numbers of queries.** N_q denotes the number of queries.

N_q	CIDEr↑	METEOR↑	BLEU4↑	$SODA_c$↑	F1↑
10	28.72	4.78	1.17	6.79	29.64
20	35.21	5.69	1.60	7.16	32.55
30	34.29	5.65	1.56	7.03	31.96
40	34.59	5.77	1.58	**7.19**	**33.56**
50	**39.18**	**6.14**	**2.10**	7.06	32.02
60	36.71	5.90	1.77	6.53	30.64
70	36.24	5.88	1.83	6.81	32.01
80	34.44	5.75	1.65	6.66	31.32
90	36.83	5.96	1.94	6.28	29.34
100	34.82	5.66	1.79	6.47	27.83

rics, yielding the highest CIDEr score of 39.16 and competitive results in other metrics. This suggests that $\gamma = 0.25$ effectively manages the trade-off between Recall and Precision. Table 6 shows the performance of OSL with and without α. Without α, we just penalize the IoU between two distinct predicted boundaries. The model achieves higher Recall without the α term, whereas adding the α term improves Precision and CIDEr. Since DVC should achieve strong performance in both captioning and localization to ensure accurate and coherent video understanding, OSL with α at $\gamma = 0.25$ demonstrates an effective balance, yielding strong results across both metrics.

Query number. We ablate the number of queries to identify the optimal query number for our method in Table 7. PDVC sets the number of queries to 100 for YC2, while ours uses $2 \times N_q$ queries due to the two query sets. When the number of queries is small, the model tends to miss events, resulting in lower Recall and limited caption coverage. As the number increases, performance improves and peaks at around 50 queries, where captioning quality and event localization are well balanced. Using too many queries leads to redundant or overlapping proposals that slightly reduce temporal precision. We find $N_q = 50$ to be optimal across all metrics.

4.4. Qualitative Results

We show qualitative results in Fig. 4. The baseline PDVC captures redundant overlaps that generate the same captions. In contrast, our method effectively suppresses unnecessary overlapping events, demonstrating that the Role Specific Query and OSL help prevent both repetitive captions and redundant overlaps, ultimately enhancing localization and captioning quality.

5. Conclusion

In this work, we tackle Dense Video Captioning through a query-based framework that focuses on improving how learnable queries are trained for localization and captioning. With our model ROS-DVC, we introduce three key components, Role Specific Query, Overlap Suppression Loss, and Concept Guider, to enable task-aware query optimization, suppress redundant detections, and enhance concept-level reasoning. Experimental results on YouCook2 and ActivityNet Captions demonstrate that our approach effectively mitigates task interference and redundancy, achieving more precise localization and semantically coherent captions. Overall, ROS-DVC provides a simple yet effective solution for query-based DVC.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korean government(MSIT)(No. RS-2026-25476101).

References

[1] Nayyer Aafaq, Ajmal Mian, Naveed Akhtar, Wei Liu, and Mubarak Shah. Dense video captioning with early linguistic information fusion. *IEEE Transactions on Multimedia*, 25: 2309–2322, 2022. 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. 6

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

[4] Zhi Chang, Dexin Zhao, Huilin Chen, Jingdan Li, and Pengfei Liu. Event-centric multi-modal fusion method for dense video captioning. *Neural Networks*, 146:120–129, 2022. 2

[5] Runhang Chen, Xiao-Yuan Jing, Fei Wu, Wei Zheng, and Yaru Hao. Task-specific parameter decoupling for class incremental learning. *Information Sciences*, 651:119731, 2023. 2

[6] 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

[7] Wei Chen, Jianwei Niu, Xuefeng Liu, and Xinghao Wu. Decoupling dense video captioning via task-specific prompts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3124–3132, 2025. 3

[8] Chuntao Ding, Zhichao Lu, Shangguang Wang, Ran Cheng, and Vishnu Naresh Boddeti. Mitigating task interference in multi-task learning via explicit task routing with non-learnable primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7756–7765, 2023. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[10] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. Soda: Story oriented dense video captioning evaluation framework. In *European Conference on Computer Vision*, pages 517–531. Springer, 2020. 6

[11] Feng Gao, Jiaxu Leng, Ji Gan, and Xinbo Gao. Selecting learnable training samples is all detrs need in crowded pedestrian detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2714–2722, 2023. 3

[12] 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

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

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

[15] Youngjoon Jang, Youngtaek Oh, Jae-Won Cho, Dong-Jin Kim, Joon Son Chung, and In So Kweon. Signing outside the studio: Benchmarking background robustness for continuous sign language recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 1

[16] Youngjoon Jang, Youngtaek Oh, Jae Won Cho, Myungchul Kim, Dong-Jin Kim, In So Kweon, and Joon Son Chung. Self-sufficient framework for continuous sign language recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[17] 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. 2

[18] 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 (CVPR)*, pages 13894–13904, 2024. 2, 3, 6

[19] 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 (ICCV)*, 2017. 1, 2, 6

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4, 5

[21] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer, 2024. 1

[22] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1

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

[24] Zhiyue Liu, Xinru Zhang, and Jinyuan Liu. Task-specific information decomposition for end-to-end dense video captioning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 16524–16536, Vienna, Austria, 2025. Association for Computational Linguistics. 2, 3, 4, 7

[25] Yifan Lu, Ziqi Zhang, Chunfeng Yuan, Peng Li, Yan Wang, Bing Li, and Weiming Hu. Set prediction guided by semantic concepts for diverse video captioning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3909–3917, 2024. 2, 5

[26] Yifan Lu, Ziqi Zhang, Chunfeng Yuan, Peng Li, Yan Wang, Bing Li, and Weiming Hu. Set prediction guided by semantic concepts for diverse video captioning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3909–3917, 2024. 5

[27] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 1

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

[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. 6

[30] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8347–8356, 2019. 1

[31] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633, 2019. 1

[32] Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. Towards consistent multi-task learning: Unlocking the potential of task-specific parameters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10067–10076, 2025. 2

[33] 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. 6

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

[35] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[36] Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D Yoo. Semantic grouping network for video captioning. In *proceedings of the AAAI Conference on Artificial Intelligence*, pages 2514–2522, 2021. 2

[37] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 1

[38] Yi Tang, Baopu Li, Min Liu, Boyu Chen, Yaonan Wang, and Wanli Ouyang. Autopedestrian: An automatic data augmentation and loss function search scheme for pedestrian detection. *IEEE transactions on image processing*, 30:8483–8496, 2021. 3

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

[40] 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

[41] 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*, pages 7190–7198, 2018. 2

[42] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023. 1

[43] Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1890–1900, 2020. 2

[44] 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 (ICCV)*, pages 6847–6857, 2021. 2, 7

[45] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7774–7783, 2018. 3

[46] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 1

[47] Jongbhin Woo, Hyeonggon Ryu, Youngjoon Jang, Jae Won Cho, and Joon Son Chung. Let me finish my sentence: Video temporal grounding with holistic text understanding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8199–8208, 2024. 1

[48] 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*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8417–8427, 2025. 2, 3

[49] Zhuyang Xie, Yan Yang, Yankai Yu, Jie Wang, Yongquan Jiang, and Xiao Wu. Exploring temporal event cues for dense video captioning in cyclic co-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8771–8779, 2025. 2, 3, 5

[50] Ruiran Yan, Rui Fan, and Defu Lian. Multi-task recommendation with task information decoupling. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2786–2795, 2024. 2

[51] 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. 2

[52] Bang Yang, Meng Cao, and Yuexian Zou. Concept-aware video captioning: Describing videos with effective prior information. *IEEE Transactions on Image Processing*, 32: 5366–5378, 2023. 5

[53] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10727–10738, 2023. 1

[54] 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. 2, 6

[55] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024. 2