Multimodal Context Fusion Based Dense Video Captioning Algorithm

Meiqi Li and Ziwei Zhou*

Abstract—**The core task of dense video description is to identify all events occurring in an unedited video and generate textual descriptions for these events. This has applications in fields such as assisting visually impaired individuals, generating news headlines, and enhancing human-computer interaction. However, existing dense video description models often overlook the role of textual information (e.g., road signs, subtitles) in video comprehension, as well as the contextual relationships between events, which are crucial for accurate description generation. To address these issues, this paper proposes a multimodal dense video description approach based on event-context fusion. The model utilizes a C3D network to extract visual features from the video and integrates OCR technology to extract textual information, thereby enhancing the semantic understanding of the video content. During feature extraction, sliding window and temporal alignment techniques are applied to ensure the temporal consistency of visual, audio, and textual features. A multimodal context fusion encoder is used to capture the temporal and semantic relationships between events and to deeply integrate multimodal features. The SCN decoder then generates descriptions word by word, improving both semantic consistency and fluency. The model is trained and evaluated on the MSVD and MSR-VTT datasets, and its performance is compared with several popular models. Experimental results show significant improvements in CIDEr evaluation scores, achieving 98.8 and 53.7 on the two datasets, respectively. Additionally, ablation studies are conducted to comprehensively assess the effectiveness and stability of each component of the model.**

Index Terms—**Dense Video Description, Transformer, Mult-imodal feature fusion, Event context, SCN Decoder**

I. Introduction

In recent years, with the rapid growth of video data, the demand for cross-modal data processing has significantly increased, and dense video captioning technology has gradually attracted widespread attention. Dense video captioning technology constructs algorithms to perform event localization and proposal generation for videos containing multiple events, and presents the content in natural language. This technology effectively bridges the semantic gap between low-level visual data and high-level abstract data in the field of artificial intelligence. Unlike traditional single-event video description tasks, dense video captioning does not require video editing but can directly generate coherent paragraph descriptions based on raw videos. This technology has broad application prospects in areas such as automatic narration, human-computer interaction, video summarization, video retrieval, and providing daily life support for visually impaired individuals. Therefore, how to enable computers to better understand video content and organize and describe it using language, similar to humans, has become an important research topic.

With the rise of deep learning, computers can leverage their powerful representational capabilities to automatically optimize and extract feature information from videos using neural network models. This has enabled machines to possess human-like abilities to understand video content, significantly advancing the development of video understanding. Since 2017, when Krishna et al. [1] first proposed the dense video captioning method DCE (Dense Captioning Events), which addressed the issues of multi-scale and overlapping event proposals caused by video extension. DCE generates proposals through the event proposal module and independently describes each event, ensuring that each video segment contains only one describable event. Since then, scholars both domestically and internationally have carried out extensive research on dense video captioning tasks, proposing various optimized and improved dense captioning methods. For example, in 2019, Rahman et al. [2] first introduced audio features into the dense video captioning task and proposed methods such as multimodal scene fusion, penalty-based hybrid fusion, and multimodal Tucker decomposition, aimed at combining visual and audio features from videos. With these methods, the model can better integrate information from different modalities and then feed it into a GRU-based description decoder to generate textual descriptions of the video. However, this method only combines visual and audio features, neglecting textual features, which limits the model's comprehensive understanding of video semantics. Additionally, the model uses weakly-supervised learning, leading to unsatisfactory description results, as it fails to fully leverage multimodal features to significantly improve the quality of the descriptions. In 2020, Iashin et al. [3] proposed a dense video captioning method based on open video datasets, which extracts multimodal features of videos through different pre-trained models: I3D convolutional networks [4] for visual features, VGGish networks [5] for audio features, and automatic speech recognition systems to obtain spoken text in the video. Then, independent

Manuscript received December 4, 2024; revised February 21, 2025.

This work was supported by the Natural Science Foundation of China (No. 61575090), the Natural Science Foundation of China Youth Fund (No. 61803189), Natural Science Foundation of Liaoning Province(2019-ZD-0031 and 2020FWDF13).

Meiqi Li is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (phone:86-16642298860, e-mail: 1473582731@qq.com).

Ziwei Zhou* is an Associate Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (Corresponding author to provide phone: 86-139-4125-5680; e-mail: 381431970@qq.com).

Transformer models are trained for each of these three modal features, and their output features are fused by concatenation to generate the final video description. Although this method effectively integrates multimodal information, the feature fusion approach is overly simplistic and inefficient, failing to fully exploit the synergistic effect of multimodal features. Furthermore, the event detection in the model is based only on visual features, which contradicts the original intention of multimodal methods.

Dense video captioning not only requires a general description of the entire video but also needs to capture a series of fine-grained events within the video and generate accurate and coherent textual descriptions for each event. Events in videos typically have temporal order and causal relationships, meaning that dense video captioning tasks must consider not only the individual features of each event but also fully utilize contextual information to understand the relationships between events. Most existing dense video captioning methods primarily rely on visual features from the video, neglecting the role of audio and textual features. Audio features provide additional cues, such as dialogues, background sounds, and sound effects, which help in more accurately understanding the context of events. Textual features (e.g., subtitles, text in images, labels, or tags) are also important contextual clues. These textual features can provide clear indicators that help quickly comprehend the video content and its events.

To address the above issues, this paper proposes a multimodal dense video captioning method based on event contextual features, with the following main contributions:

1. Text Information Extraction Using OCR: This paper proposes using OCR to extract textual information from the video (i.e., video textual features) and integrate it into multimodal features to enhance the model's ability to understand the video content. By combining information from the visual, audio, and textual modalities, the model can leverage the complementarity of these modalities to achieve a more comprehensive understanding of the events in the video, thereby improving overall event comprehension.

2. Multimodal Context Fusion Encoder: A multimodal context fusion encoder is designed to integrate temporal semantic consistency and multimodal information, capturing both local and global context. This improves the coherence, fluency, and diversity of the descriptions. Additionally, it effectively models the dependencies between events, making the generated descriptions more logical and richer.

3. SCN Decoder for Fine-Grained Description Generation: The SCN [6] semantic decoder is used in the decoding phase. The SCN decoder enhances the semantic consistency, detail-capturing ability, long-term dependency handling, and fluency of natural language generation through fine-grained, word-by-word generation. It also effectively integrates multimodal information, addressing issues such as semantic incoherence and imprecise descriptions that are common in traditional video captioning methods.

II. Relates Jobs

A. Video captioning

Early video description techniques primarily relied on predefined templates [7], which dominated the field of video description research before 2014. These methods initially fed videos into detection networks, where classifiers identified key elements in the videos, such as objects, actions, and scenes, and mapped these elements to basic components of sentences, such as subjects, predicates, and objects. Subsequently, based on these key elements, relationships between words were constructed, and appropriate conjunctions were inserted to generate complete natural language descriptions. Although this method was simple and could quickly produce descriptions, it had significant limitations. On one hand, embedded models were typically trained for specific topics, and when the topic changed, retraining was necessary, limiting their generalization capabilities. On the other hand, the setting of templates and vocabulary depended on human input, and the description generation process was merely the filling in of keywords, lacking flexibility and failing to match the expression of human natural language.

With the continuous advancement of research, video description methods based on templates have seen significant improvements in the precision and efficiency of video feature extraction. The rapid development of deep learning has led to the introduction of more related technologies, which are used to extract richer feature vectors and to process multimodal data such as optical flow, RGB images, and text, thereby more accurately identifying video content and themes. Currently, models based on the encoder-decoder architecture have become the mainstream in the field of video description. These models, inspired by machine translation, adopt a sequence-to-sequence transformation approach. Specifically, they first vectorize the features of the video and then use trained models to transform these features into descriptive natural language. The entire process is roughly as follows: In the video feature processing stage, to consider the motion information brought by temporal changes, optical flow modules are typically introduced or the network structure is enhanced with a temporal dimension to handle sequential information. Subsequently, Convolutional Neural Networks (CNN) are used to extract video features, which are then encoded into vectors and passed to the decoder. The decoder usually employs a recurrent neural network or a Transformer [8] structure, focusing on the contextual relationships of video features during language generation. Based on the already generated vocabulary, it further guides the generation of descriptions, ensuring that the output language is more natural and fluent.

B. Transformer-based video captioning

In video captioning tasks, the self-attention mechanism of Transformer offers significant advantages, particularly in capturing the complex spatiotemporal relationships between video frames. Unlike traditional video captioning methods based on Recurrent Neural Networks (RNNs), Transformer can simultaneously process both visual information and temporal dynamics of the video. This allows it to effectively capture temporal dependencies between frames, resulting in more coherent and natural captions. While RNNs often struggle with issues like vanishing or exploding gradients when processing long sequences, Transformer alleviates these challenges through global dependency modeling, leading to more stable and fluent video caption generation.

In recent years, the application of the Transformer architecture in the field of video understanding has garnered increasing attention, particularly in the area of multimodal learning where significant advancements have been made. Models such as VideoBERT [9] and ViLBERT [10] have integrated visual and linguistic information, utilizing end-to-end training to enable the simultaneous learning of visual features from videos and linguistic features from text. VideoBERT employs a joint pre-training method combining Transformer [8] and BERT [11], while ViLBERT processes video and language information through a dual-stream architecture, achieving remarkable results in tasks such as video description generation and video question answering.

To more effectively model spatiotemporal information in videos, improvement schemes such as TimeSformer [12] have proposed the novel idea of separating the temporal dimension from the spatial dimension. TimeSformer uses independent self-attention mechanisms to model spatial and temporal features separately, thereby enhancing the efficiency and effectiveness of video processing. Similarly, the Video Transformer [13], [14] model, by adopting the Transformer architecture, has realized efficient modeling and description generation of video content. These innovations have demonstrated powerful performance in capturing spatiotemporal dependencies, modeling long-term dependencies between video frames, and fusing multimodal information, enabling the model to provide more accurate, smooth, and context-rich natural language descriptions in video description generation tasks, especially when dealing with long videos, complex scenes, and multimodal tasks.

In addition to improvements in model architecture, researchers have also delved into cross-modal alignment. In 2021, Chandra Sekhar et al. [15] introduced a multimodal attention mechanism that combines visual and linguistic information for video description generation, effectively addressing the alignment issues between visual information and language descriptions. By integrating textual information, this approach enhances the accuracy and naturalness of the generated descriptions, thus producing more precise and coherent descriptions. Furthermore, in 2022, Liu et al. [16] proposed a multimodal Transformer model that fuses visual and audio information, further enhancing the performance of video description generation. This method not only increases the diversity of description generation but also improves the model's ability to model relationships between different modalities, significantly enhancing the overall quality of video descriptions.

To address challenges such as long videos and event overlap, many Transformer-based methods have undergone significant improvements. During the event proposal stage, researchers have developed specialized event proposal modules designed to detect key events within a video. These modules are often combined with advanced temporal dependency modeling techniques, helping to resolve issues like event overlap and boundary ambiguity, which can hinder accurate event identification [17]. In addition, several methods have focused on enhancing the contextual relationships between event proposals and the generated captions. By ensuring that the captions reflect the logical and temporal connections between events, these approaches improve the overall coherence and relevance of the generated descriptions [18], [19]. Moreover, mutual guidance and reinforcement mechanisms, which facilitate the interaction between event proposal and event description modules, have been widely adopted. These mechanisms help refine the accuracy of event detection and caption generation, ensuring that the events are both correctly identified and appropriately described [20], [21]. To further reduce redundancy and improve the diversity of the generated captions, some models have refined their event generation strategies, allowing for more precise and varied descriptions of the events [22]. Collectively, these innovations have significantly advanced the performance of dense video captioning tasks, leading to more accurate, coherent, and contextually rich descriptions that better reflect real-world scenarios.

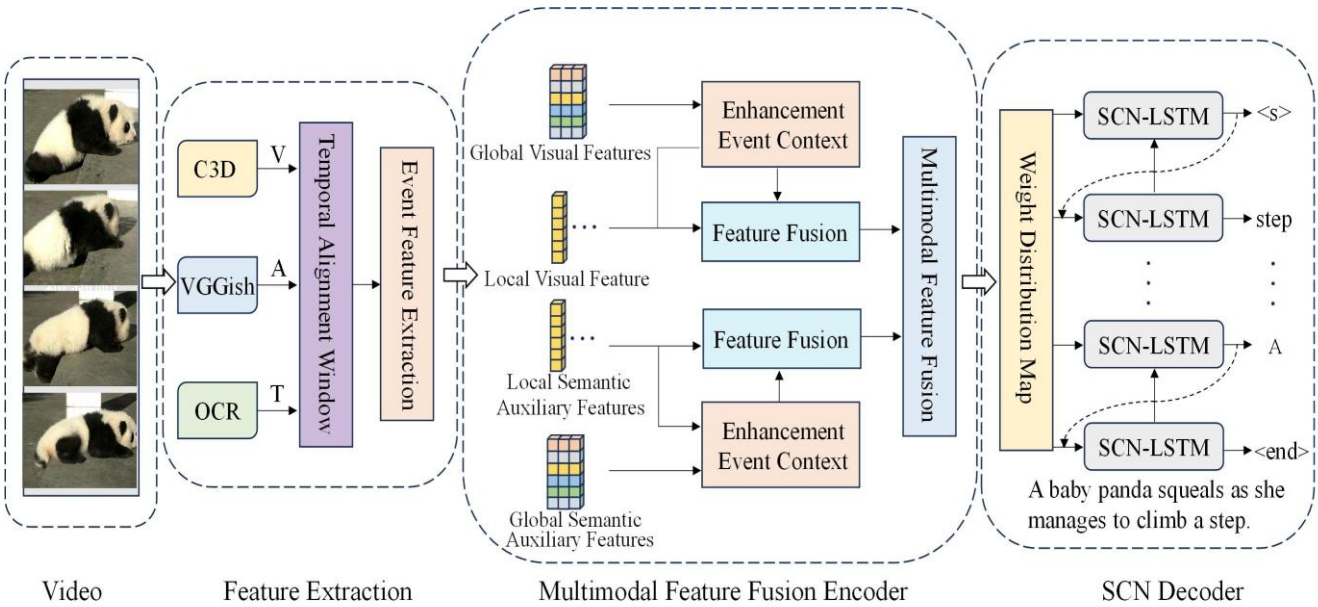

Fig. 1. The overall structure of the model in this article

III. Model Design

This study proposes a video captioning model, as shown in Figure 1, which consists of three main components: multi-feature extraction with temporal alignment, a multimodal context fusion encoder, and an SCN decoder. Existing models often overlook the importance of textual information (e.g., road signs, subtitles) in enhancing event understanding, and fail to capture cross-event relationships in caption generation. To address these issues, our model integrates visual, audio, and textual features using multimodal context modeling and cross-modal interactions. Visual features are extracted with C3D, audio features with VGGish, and textual features through OCR. Temporal alignment ensures consistency across modalities. The context fusion encoder uses self-attention and AOA attention [23] to align and combine features, capturing dependencies between events and enhancing contextual understanding. Finally, the SCN decoder generates captions word by word, improving semantic consistency, logic, and fluency, overcoming the limitations of traditional models in handling event relationships and semantic expression.

A. Feature Extraction and Temporal Alignment

To synchronize the processing of multimodal features and ensure the alignment of visual, audio, and textual features on the same time scale, this paper proposes an alignment method based on sliding windows and temporal interpolation. This approach enables the model to capture corresponding multimodal information at each time point, providing a solid foundation for subsequent feature fusion and caption generation. Specifically, visual features (including spatio-temporal information), audio features (such as speech and background sounds), and textual features (e.g., subtitles, road signs, or speech-to-text) are extracted and temporally aligned to ensure consistency across modalities.

1) *Visual Feature Extraction and Alignment:* In the process of visual feature extraction, to capture the spatio-temporal characteristics and dynamic information of the video, this paper extracts RGB video frames from the video at a sampling rate of 25 frames per second and generates optical flow frames using PWC-Net. RGB frames are used to capture the static information of the scene and objects, while optical flow frames describe the dynamic motion characteristics of the objects. To ensure input consistency and computational efficiency, the shorter side of all frames is resized to 256 pixels, and a central crop is taken to obtain a 224×224 image region.

These processed RGB and optical flow frames are then fed into the pre-trained C3D network to extract their spatio-temporal features. The C3D model analyzes approximately 2.56 seconds of video content (64 frames) and generates feature vectors that contain dynamic information, represented as:

$$V_t = C3D(Frame_t) \tag{1}$$

Where $V_t \in R^{n_v \times d_v}$ represents the visual features at time t, with n_v and d_v denoting the temporal steps and the dimensionality of the visual features, respectively. Through this process, visual features enriched with spatio-temporal and motion information are obtained.

2) *Audio Feature Extraction and Alignment:* To obtain clear and consistent audio features, the audio signal is first preprocessed through the following steps: (1) The audio signal is resampled to 16kHz mono to ensure consistent sampling rate; (2) Noise suppression techniques are applied to remove background noise and enhance signal quality; (3) A band-pass filter is used to retain the frequency range from 300 Hz to 3400 Hz, filtering out irrelevant low- and high-frequency noise; (4) Dynamic range compression and normalization techniques are applied to balance volume differences and ensure signal consistency.

The processed audio signal is then divided into continuous segments and transformed into a spectrogram using Short-Time Fourier Transform (STFT) to capture frequency information. The spectrogram is subsequently converted into a 96×96 Mel spectrogram and fed into the pre-trained VGGish network to extract high-level audio features, represented as:

$$A_{raw} = VGGish(MelSpectrogram(STFT(Audio)) \tag{2}$$

Since the temporal resolution of audio features is usually denser than that of visual features, this paper downsamples the audio features to match the temporal scale of the visual features using average pooling within a sliding window. Specifically, considering that the video is sampled at 25 frames per second, where each frame represents approximately 40 milliseconds, and the sampling rate of audio features is typically higher, a pooling window size of 40 milliseconds is used to ensure temporal alignment between the audio and visual features.

In the pooling operation, 4 audio feature points (corresponding to a 40-millisecond time window) are selected as one pooling window. Through average pooling, the audio features are downsampled to match the temporal resolution of the visual features. The pooling process is represented as:

$$A_t = Polling(A_{raw}(t_1 : t_n)) \tag{3}$$

Where $A_t \in R^{n_v \times d_a}$ represents the audio features at time t, with n_v and d_a denoting the temporal step and dimensionality of the audio features, respectively. Through temporal alignment, the audio features are synchronized with the visual features, providing rich information for multimodal fusion.

3) *Text Feature Extraction and Alignment:* Text information in videos, such as subtitles, road signs, billboards, etc., typically appears in a scattered form across specific frames. To extract this text information and ensure its alignment with visual and audio features, this paper utilizes Optical Character Recognition (OCR) and a pre-trained BERT model for text feature extraction and encoding. The processing steps are as follows:

The text content is detected and extracted from the video frames using OCR. The extracted text is then fed into a pre-trained BERT model to generate the corresponding text embedding vector, represented as:

$$T_t = BERT(OCR(Frame_t)) \tag{4}$$

If no text is detected within a given time window, a zero vector is used for padding to maintain temporal alignment. This ensures that time periods with and without text are clearly distinguished, preventing the inclusion of irrelevant

features.Through this process, the text features are aligned with the visual and audio features in terms of time.

Where $T_t \in R^{n_v \times d_t}$ represents the text features at time t, and n_v and d_t denote the time step and the dimension of the text features, respectively. Through the sliding window and time interpolation process, the text features are synchronized with the visual and audio features, providing semantic supplementation for multimodal fusion.

B. Multi-modal Context Fusion Encoder

In the task of dense multimodal video captioning, videos not only contain rich multimodal feature information (such as visual, audio, and textual data) but also harbor complex contextual relationships. These pieces of information are crucial for generating accurate and coherent descriptions. Multimodal features help the model understand the content of the current video segment, while contextual relationships reveal the connections between different events, further enhancing the understanding of the overall scene. To address this, we propose a multimodal context fusion encoder, aimed at achieving deep fusion of multimodal information through cross-modal feature interactions and alignment, as well as capturing temporal relationships and semantic interactions between events.

The model is based on an improved Transformer architecture, consisting of four components. The first part is event feature extraction, which extracts local event features. The second part is the self-attention mechanism, which encodes the three features separately. The third part is the AOA attention mechanism, which integrates event context. The fourth part is multimodal information fusion, which applies adaptive weighted fusion of visual, audio, and textual features to ensure that the weight of each modality is automatically adjusted based on its contribution to the current task.

First, the self-attention mechanism is used to encode the three types of features separately. In the multimodal encoder, visual features serve as the core support, while textual and audio features act as semantic auxiliary features to complement and enhance the precision of video descriptions. The encoder then employs the AOA attention mechanism, which centers around the current event and guides the contextual information into the current feature, generating a contextual feature representation with global semantics. This approach not only strengthens the expression of each event but also optimizes the logical connections between events, helping the model more accurately depict independent events when generating descriptions, while improving the coherence and consistency of the captions. The structure of the encoder is shown in Figure 2.

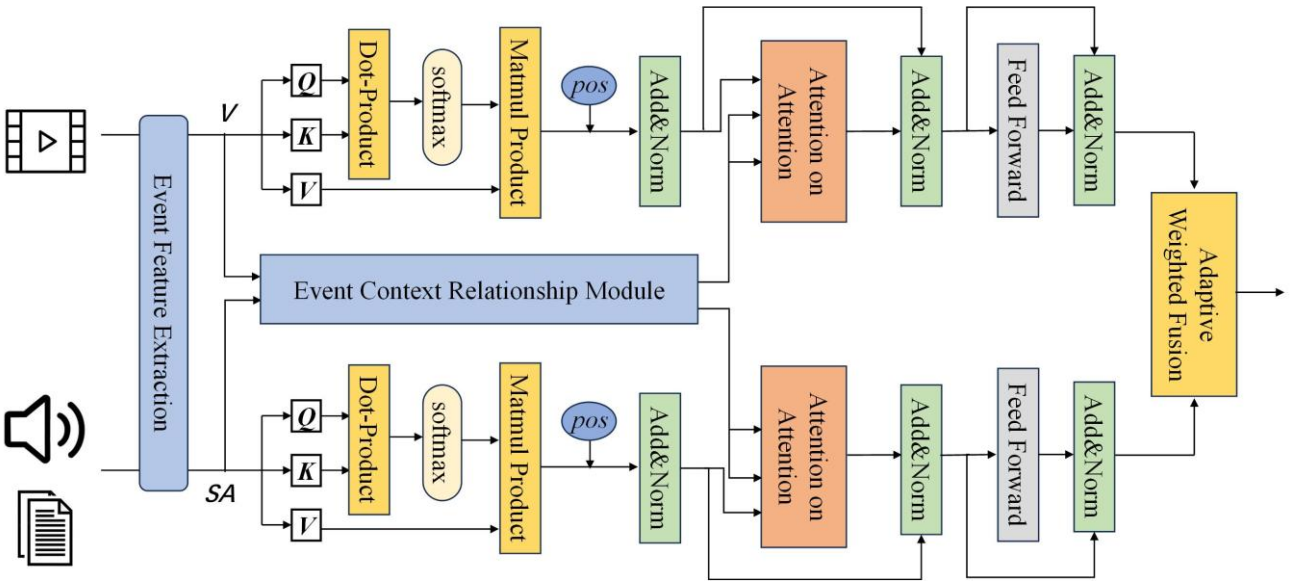

Fig. 2. The structure diagram of the Multi-modal Context Fusion Encoder

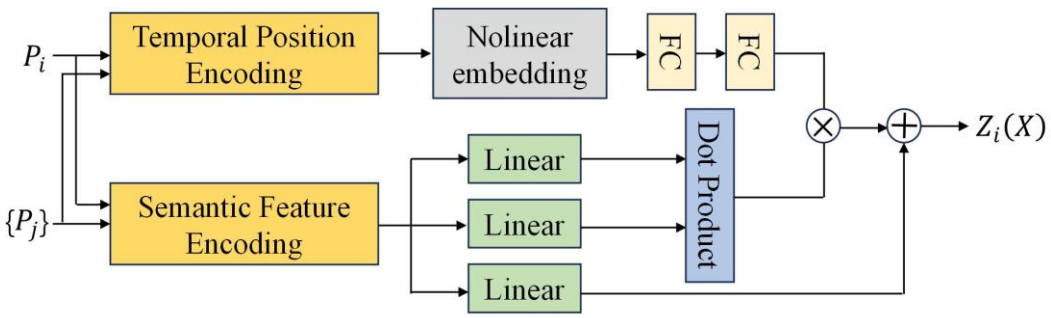

Fig. 3. The structure diagram of the Event Context Relationship Module

The core function of the event context relationship module is to capture the temporal relationships and semantic similarities between different events [24]. To achieve this, the module introduces a dual-branch network based on the attention mechanism, which separately assesses the temporal relationships and semantic associations between events. By performing a weighted fusion of these two relationship scores, the module effectively integrates global event features, generating more accurate and enriched context-enhanced event representations that improve the overall event representation capability. Figure 3 shows the structure of the event context relationship module.

The temporal relationship primarily focuses on the time order and duration of events. For the current event P_i and other events $\{P_j\}$ in the video, position encoding is applied based on their relative distance and time length:

$$P_{ij} = \left[\frac{c_i - c_j}{l_i}, \log\left(\frac{l_j}{l_i} \right) \right] \tag{5}$$

Where c_i and c_j epresent the center positions of events P_i and $\{P_j\}$, respectively, and l_i and l_j denote the corresponding durations of the events. The term $c_i - c_j$ is used to represent the temporal order between events, which helps distinguish their sequence, as many event annotations in video datasets contain temporal order words such as "then," "continue," and "end."

Additionally, to eliminate the influence of the overall video duration on the model, the duration of the events is normalized to ensure the features are independent of the time scale. This normalization process helps the model better adapt to video data with varying lengths.

After position encoding, the low-dimensional vector P_{ij} may not have sufficient discriminative power, so a nonlinear function is applied to embed P_{ij} into a high-dimensional space, represented as $\hat{P}_{ij}$, which is then further fed into two fully connected (FC) layers to predict the event score. The specific formula is as follows:

$$\hat{P}_{ij} = ReLU\left(W_P P_{ij} + b_P \right) \tag{6}$$

$$h_1 = ReLU\left(W_1 \hat{P}_{ij} + b_1 \right) \tag{7}$$

$$h_2 = ReLU\left(W_2 \hat{P}_{ij} + b_2 \right) \tag{8}$$

$$a_{p_{ij}} = W_S h_2 + b_S \tag{9}$$

Where W_P, W_1, W_2, and W_S are weight matrices, while b_P, b_1, b_2 and b_S are bias terms, each corresponding to position encoding mapping, high-dimensional space mapping, fully connected layer transformations, and score prediction, respectively.

Unlike the temporal relationship, the semantic relationship focuses on capturing the similarity between event contents. Since video content can be quite lengthy, the semantic features of each event must account for both short-term and long-term information. To achieve this, an LSTM is used to encode all event features into recurrent representations, which are then concatenated with the frame features after mean pooling to obtain the event's semantic feature S_i Subsequently, the scaled dot-product method is used to capture the linear correlations within the semantic space, and the semantic relationship score is computed as follows:

$$a_{s_{ij}} = \frac{\left(W_Q S_i \right)^T \left(W_K S_j \right)}{\sqrt{d_E}} \tag{10}$$

Where W_Q and W_K are learned linear mapping layers that project the semantic features S_i and S_j into a shared embedding space of dimension d_E. This method allows the model to efficiently capture the semantic similarity between event proposals.

By calculating the temporal and semantic relationship scores between events, the context relationship features for the current event can be obtained, as expressed by the following formula:

$$score = softmax\left(a_{p_{ij}}(X) \cdot a_{s_{ij}}(X) \right) \tag{11}$$

$$Z_i(X) = \sum_{j=1}^{N_p} score(X) \cdot \left(W_V \cdot S_j(X) \right) \tag{12}$$

Where N_p represents the number of events, W_V is the linear embedding layer, and X can represent visual, audio, or textual features. Using this approach, the context relationship features for the current event can be obtained for different modalities, including visual context features Z_V, audio context features Z_A, and textual context features Z_T.

In multimodal tasks, effective fusion of local and global features is crucial for improving model performance. This paper employs the AOA attention mechanism to enhance the feature representation of local events by leveraging the context relationship features of the current event. The context relationship features of the current event are combined with event features and passed into the AOA attention mechanism layer, where the context relationship features (global visual features) serve as keys (K) and values (V), and the event features (local features) serve as the query (Q). The attention mechanism then integrates the context information with the event information to enhance the semantic relevance of the local event features. The structure of the AOA attention mechanism is shown in Figure 4. Specifically, by calculating the correlation information based on the query and attention results, an attention gate is generated to control the strength of the information flow. Finally, through the multi-head attention mechanism, this information is fused to enhance the representational capacity of the local event features.

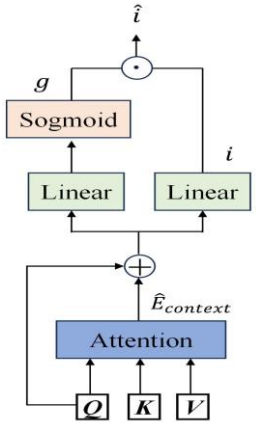

Fig. 3. Attention on Attention

First, the standard attention mechanism is used to compute the similarity between the current event feature $E_{current}$ and the context event feature $E_{context}$, yielding a context-based weighted average result, as shown in the following formula:

$$\hat{E}_{context} = Attention(E_{current}, E_{context}, E_{context}) \quad (13)$$

Next, we introduce an information vector $i_{local-global}$ to represent the associative information between the current event and the context events. This information vector is extracted by applying a linear transformation to the current event feature and the attention result, capturing the contextual information:

$$i_{local-global} = W_q^i E_{current} + W_v^i \hat{E}_{context} + b_i \quad (14)$$

Where W_q^i and W_v^i are weight matrices, and b_i is the bias term. Then, an attention gate $g_{local-global}$ is computed to control the contribution of each part of the information vector $i_{local-global}$. This gate dynamically adjusts the strength of the information flow by evaluating the correlation between the current event and the context event:

$$g_{local-global} = \sigma\left(W_q^g E_{current} + W_v^g \hat{E}_{context} + b_g\right) \quad (15)$$

Where σ is the activation function, W_q^g and W_v^g are weight matrices, and b_g is the bias term. The attention gate $g_{local-global}$ and the information vector $i_{local-global}$ are combined through element-wise multiplication to obtain the refined fused information $\hat{i}_{local-global}$, as follows:

$$\hat{i}_{local-global} = g_{local-global} \odot i_{local-global} \quad (16)$$

Where $\odot$ represents element-wise multiplication, ensuring that the attention gate dynamically adjusts the information based on the contribution of each part.

Finally, the generated fused information is further processed through a multi-head attention mechanism to obtain the context-enhanced feature of the current event. This step captures multidimensional interactions between event features by combining multiple attention heads, as shown in the following formula:

$$\hat{E}(X) = MultHead(E_{current}, \hat{E}_{context}, \hat{i}_{local-global}) \quad (17)$$

Where *MultiHead* denotes the multi-head attention mechanism, which can parallelly capture different interaction patterns between event features through multiple heads. X can represent visual, audio, or textual features.

We represent the enhanced visual, audio, and textual event context features as $\hat{E}(V)$, $\hat{E}(A)$, and $\hat{E}(T)$, respectively. To effectively fuse these three modalities, we employ a modality-adaptive weighted fusion method. This method adjusts the weight of each modality based on its contribution to the current task, ensuring that the information from different modalities is appropriately emphasized during fusion. The formula is as follows:

$$E_{fused} = \alpha_V \hat{E}(V) + \alpha_A \hat{E}(A) + \alpha \hat{E}(T) \quad (18)$$

Where E_{fused} represents the fused multimodal feature, and α_V, α_A, and α_T are learnable weight parameters that denote the importance of each modality in the final fused feature. These weights can be adaptively adjusted during the training process to optimize the fusion results.

C. SCN Decoder

In the task of video description, traditional recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) have been widely used to capture the temporal dependencies in sequential data. However, these networks often fail to fully exploit the interactions and semantic information between different modalities when processing multimodal information. To address this issue, we employ a Semantic Composition Network (SCN) as the decoder and introduce multimodal fusion features to further enhance the model's ability to express video content.

At each time step t, the SCN-LSTM decoder updates its state through the LSTM unit. We use the multimodal feature E_{fused} to initialize the LSTM's hidden state and cell state:

$$h_0 = LSTM(E_{fused}) \quad (19)$$

Where h_0 is the initial hidden state of the LSTM, and E_{fused} is the fused multimodal feature, which includes visual, audio, and textual information. At each time step, the LSTM updates its state based on the previous hidden state h_{t-1} and the current input x_t. The specific calculations involve updating the input gate i_t, forget gate f_t, output gate o_t, and cell state c_t, as follows:

$$i_t = \sigma\left(W_i x_i + U_i h_{t-1} + b_i\right) \quad (20)$$

$$f_t = \sigma\left(W_f x_f + U_f h_{t-1} + b_f\right) \quad (21)$$

$$o_t = \sigma\left(W_o x_o + U_o h_{t-1} + b_o\right) \quad (22)$$

$$\tilde{c}_t = \tanh\left(W_c x_c + U_c h_{t-1} + b_c\right) \quad (23)$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \quad (24)$$

$$h_t = o_t \odot \tanh\left(c_t\right) \quad (25)$$

To improve the decoder's understanding of the video's semantics, the SCN-LSTM model introduces a semantic concept vector s. This semantic concept vector is extracted from the annotated text of the video and represents semantic information in the video, such as objects, actions, etc. It helps the model better understand the specific semantic content of the video and guides the decoding process. At each time step, the semantic concept vector s is used to weight the parameters of the LSTM, making the decoding process more aligned with the video's semantic structure. Through this weighting mechanism, the semantic concept vector s adjusts the set of LSTM parameters, thereby influencing the decoder's behavior at different time steps and ensuring that the generated description is more consistent with the video's semantic content and context. The formula is as follows:

$$LSTM_s = \sum_{k=1}^{K} \alpha_k \cdot LSTM_k \quad (26)$$

Where α_k is the weight parameter influenced by the semantic concept vector s, used to adjust the influence of different LSTM units.

During the decoding process, the LSTM generates the video description based on the multimodal features E_{fused} at the current time step and the weighted parameters of the

semantic concept vector s. By capturing temporal dependencies and guiding with semantic information, the generated description can accurately and coherently express the video's content.

IV. Experiments and Analysis

A. Datasets and Evaluation Indicators

There are many publicly available datasets for video captioning, but many of them tend to focus on specific topics or scenarios, such as movies, sports, or food. To ensure the model's generalization ability and its applicability to a broader range of real-world scenarios, this paper selects two commonly used video captioning datasets that cover various scenes and contexts: the Microsoft Video Description (MSVD) [25] and Microsoft Research Video to Text (MSR-VTT) [26]. These two datasets offer high diversity and a large number of captioning samples, helping the model learn richer semantic and visual associations. Specifically, the MSVD dataset contains short videos from various everyday life scenes, with each video being annotated with an average of 41 individual sentences. On the other hand, the MSR-VTT dataset provides more diverse long videos with richer descriptions, where each video is annotated with about 20 sentences. The diversity and large number of annotations in these datasets make them ideal choices for training and evaluating multimodal video captioning models. Table I lists the detailed information of these two datasets.

This paper selects four widely recognized evaluation metrics to measure the performance of the proposed method: BLEU [27], ROUGE [28], METEOR [29], and CIDEr [30]. These metrics are widely used in the field of video description generation and can effectively assess the quality of generated descriptions. Specifically, the BLEU metric measures the accuracy of descriptions by comparing the number of common words between the generated text and the reference text, focusing on precise word matching. The ROUGE metric, similar to BLEU, focuses on assessing the lexical overlap at the sentence level but places more emphasis on recall, that is, evaluating the completeness of the original information covered by the generated text. The METEOR metric takes into account both precision and recall, evaluating the quality of generated descriptions through grammatical structure and semantic similarity, providing a more comprehensive evaluation perspective. The CIDEr metric is designed based on human evaluation consensus and is tailored for image and video description tasks. It can accurately reflect the consistency between generated descriptions and human evaluations, thereby providing a more comprehensive and human-perception-oriented evaluation dimension for description quality.

B. Experimental environment

The experiments in this paper were conducted on a Linux operating system (Ubuntu 20.04 version), using the PyTorch deep learning framework for model training. The hardware configuration included an Intel Core i9-13900KF CPU paired with a Geforce RTX 4060 Ti GPU (16GB VRAM) and 32GB of RAM, which was sufficient to support the training of large-scale video data. During the training process, the Adam optimizer was used with an initial learning rate set at 0.0001, combined with L2 regularization (weight decay rate of 0.0005) to prevent overfitting. The optimizer's momentum was set to 0.9 to accelerate convergence. The batch size was set to 64 with 5000 iterations. To further prevent overfitting, Dropout regularization was applied with a rate of 0.2, reducing model complexity and enhancing generalization by randomly dropping some connections in the neural network. The Transformer encoder was configured with 6 layers, each containing 8 self-attention heads, effectively capturing long-range dependencies within video frames. The LSTM decoder's hidden layer size was set to 512, optimizing the video-to-text generation task. Beam Search with a beam size of 5 was employed to improve the quality of text generation.

C. Comparative Experimental Analysis

To objectively evaluate the video description generation model proposed in this paper, we conduct comparative experiments with current state-of-the-art video description generation models, including POS-CG [31], ORG-TAL [32], SAAT [33], NACF [34], and SGN [35], and perform evaluations on the same datasets. To comprehensively assess the model's performance, we use four common evaluation metrics: BLEU, CIDEr, METEOR, and ROUGE-L, to quantitatively analyze the generation results of each model. The experimental results are summarized in Tables II and III, which show the scores of each model on different metrics.

As shown in Table II, on the MSVD dataset, the proposed model outperforms other models in all metrics. The most significant improvement is observed in the CIDEr score, which increased by 2.5%. CIDEr, specifically designed for descriptive tasks, aligns most closely with human judgment, thus confirming the effectiveness of the proposed model. This demonstrates that incorporating the event context enhancement module to capture the temporal and semantic relationships between events can significantly improve the model's performance.From Table III, it can be seen that on the MSR-VTT dataset, all metrics show improvement: BLEU-4 increased by 1.3%, METEOR by 1.0%, CIDEr by 2.3%, and ROUGE by 1.7%. Once again, the significant improvement in the CIDEr score indicates that the introduction of the multimodal context fusion encoder enables a deep integration of visual, audio, and textual features, leading to better description performance.

TABLE I
Dataset Details

Datase	Total Number of Videos	Average Duration	Common Split	Total Vocabulary	Conversion from Gaussia
MSVD	1970	10sec	12000/100/670	607399	13010
MSR-VTT	10000	20sec	6513/497/2900	1856523	29316

TABLE II
COMPARATIVE RESULTS OF DIFFERENT METHODS ON MSVD DATASET

Methods	BLUE-4	METEOR	CIDEr	ROUGE
POS-CG	52.5	34.1	88.7	71.3
ORG-TAL	54.3	36.4	95.2	73.9
SAAT	46.5	33.5	81.0	69.4
NACF	55.6	36.2	96.3	-
SGN	52.8	35.5	94.3	72.9
Ours	56.1	37.8	98.8	75.6

TABLE III
COMPARATIVE RESULTS OF DIFFERENT METHODS ON MSR-VTT DATASET

Methods	BLUE-4	METEOR	CIDEr	ROUGE
POS-CG	38.3	26.8	43.4	60.1
ORG-TAL	43.6	28.8	50.9	62.1
SAAT	460.5	28.2	49.1	60.9
NACF	42.0	28.7	51.4	-
SGN	40.8	28.3	49.5	60.8
Ours	44.9	29.8	53.7	63.8

GT:A cat on a coffee table bats at a puppy with its paw
Baseline:The puppy is playing with the cat
Ours:The cat on the table is hitting the dog on the ground

GT:The horse and rider trotted in the field
Baseline:An individual is riding a horse outdoors
Ours:A man is riding a horse in an open field surrounded by greenery

Fig. 4. Visualization results on MSVD Data

GT:Two ping pong players are in active competition against each other
Baseline:Two people are playing table tennis
Ours:Two men compete in a table tennis match in a large stadium

GT:an old man in formal wear is delivering speech to the audience gathered in the auditorium
Baseline:A man is standing on stage, discussing with a large audience
Ours:An elderly gentleman speaker is standing on a large stage delivering a speech

Fig. 5. Visualization results on MSR-VTT Data

In this paper's experiments, the standard Transformer structure is used as the baseline model. Figures 5 and 6 show the visual results of the proposed model (Multimodal Context Fusion Based Dense Video Captioning Method, MCF-DVC) and the baseline model on the MSVD and MSR-VTT datasets, respectively.

Based on the experimental results above, the model is able to accurately identify the subject, object, and verb in the sentences, particularly excelling in recognizing the number of participants in an activity. For example, it can correctly identify "A man" and "An elderly gentleman speaker". This further confirms that, under the influence of the multimodal context fusion encoder and attention mechanism, the model not only accurately identifies and associates the subject and object of the activity, but also reasonably infers the number of participants in the event. This enhances the semantic consistency and accuracy of the generated descriptions. These findings suggest that the proposed model possesses strong semantic understanding and expressive capabilities in the video description task, enabling it to generate descriptions that are more aligned with human intuition.

D.Ablation Experiments

To validate the rationality and effectiveness of the proposed model, this paper conducts ablation experiments and uses a controlled variable approach to analyze the contribution of each module to the overall performance of the model. The experimental results are shown in Tables IV and V. The ablation experiments consist of four parts, using the standard Transformer structure as the baseline model. A, B, and C represent the control experiments based on the baseline model, where only text features (such as road signs, subtitles, etc.) are fused, only the multimodal context fusion encoder is introduced, and only the SCN decoder is used, respectively. "Ours" refers to the complete model proposed in this paper.

As shown in Tables IV and V, each individually introduced module provides some improvement to the model, and the combination of all three modules in MCF-DVC yields the best performance. From the results, it can be observed that models A, B, and C, when compared to the baseline model, show improvements in multiple evaluation metrics to varying degrees. Notably, the CIDEr (C) score of model B shows the most significant improvement, increasing by 8.3% and 6.6% on the two datasets, respectively. This metric emphasizes the semantic consistency between the generated text and multiple reference descriptions, confirming that the multimodal context fusion encoder in model B effectively considers the matching of context and key vocabulary.

Furthermore, by comparing MCF-DVC with the baseline model, the experimental results demonstrate that each module proposed in this paper has a significant impact on the overall model's performance, validating the synergistic effect between different modules. These modules help the model better understand the relationships between different modalities and capture the subtle connections between events, thereby enhancing features and improving the accuracy and richness of the generated descriptions.

TABLE IV
COMPARISON OF ABLATION EXPERIMENT RESULTS. ON MSVD DATASET

Methods	BLUE-4	METEOR	CIDEr	ROUGE
Baseline	49.4	33.3	90.0	69.2
A	50.6	35.8	92.7	71.8
B	50.9	36.6	98.3	72.0
C	52.7	34.0	93.5	75.3
Ours	56.1	37.8	98.8	75.6

TABLE V
COMPARISON OF ABLATION EXPERIMENT RESULTS. ON MSR-VTT DATASET

Methods	BLUE-4	METEOR	CIDEr	ROUGE
Baseline	37.4	26.5	46.8	56.6
A	40.4	27.8	48.1	57.9
B	40.8	28.6	53.4	58.2
C	42.7	27.0	49.0	63.7
Ours	44.9	29.8	53.7	63.8

V. Conclusion

This paper proposes a Multimodal Context Fusion-based Dense Video Description method, which utilizes OCR technology to extract textual information and considers three different modalities: video, audio, and text features. This approach enhances the model's semantic understanding of video content and ensures the temporal consistency of multimodal features through sliding window and time alignment techniques. Based on this, a multimodal context fusion encoder is employed to capture the temporal and semantic relationships between events, enabling the deep fusion and alignment of multimodal features. Additionally, an SCN decoder is used to improve the semantic consistency and fluency of the generated descriptions through word-by-word generation and fine-grained semantic processing.

In the experimental section, comprehensive experiments are conducted on the MSVD and MSR-VTT datasets to validate the effectiveness of the proposed Multimodal Context Fusion Based Dense Video Captioning method. A series of ablation experiments are also performed to further analyze the contributions of each module. The experimental results show that the proposed model outperforms current state-of-the-art models on multiple common evaluation metrics, particularly in CIDEr, where the MCF-DVC model improves by 2.5% and 2.3% on the two datasets, respectively. Moreover, the descriptions generated by the model exhibit outstanding semantic consistency and logical coherence, capturing the temporal relationships and detailed information in the video more accurately. Compared to traditional methods, the generated descriptions are more natural, coherent, and accurate.

In future work, we plan to further optimize the model's decoding process and explore the integration of reinforcement learning to enhance the diversity and creativity of generated descriptions. By guiding the model to explore more diverse and rich generation paths during training, reinforcement learning can better capture long-range semantic dependencies. Additionally, we will focus on improving the model's inference speed, considering the introduction of knowledge distillation techniques to compress the complex multimodal model into a lightweight version, thereby enhancing its efficiency and real-time performance in practical applications.

References

[1] R. Krishna, K. Hata, F. Ren, et al., "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis*., Venice, 2017, pp. 706–715.

[2] T. Rahman, B. Xu, and L. Sigal, "Watch, listen and tell: Multi-modal weakly supervised dense event captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis*., 2019, pp. 8908–8917.

[3] V. Iashin and E. Rahtu, "Multi-modal dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 958–959.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[5] S. Hershey, S. Chaudhuri, D. P. W. Ellis, et al., "CNN architectures for large-scale audio classification," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 131–135.

[6] Zhe Gan, Chuang Gan, Xiao-Dong He, et al., "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5630–5639.

[7] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, 2002.

[8] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, New York: ACM, 2017, pp. 6000–6010.

[9] Chen Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A Joint Model for Video and Language Representation Learning," *arXiv preprint arXiv:1904.01766*, Apr. 2019. [Online]. Available: https://arxiv.org/abs/1904.01766. DOI: 10.48550/arXiv.1904.01766.

[10] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *arXiv preprint arXiv:1908.02265*, Aug. 2019. [Online]. Available: https://arxiv.org/abs/1908.02265. DOI: 10.48550/arXiv.1908.02265.

[11] J. Devlin, Ming-Wei Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.04805. DOI: 10.48550/arXiv.1810.04805.

[12] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" *arXiv preprint arXiv:2102.05095*, Feb. 2021. [Online]. Available: https://arxiv.org/abs/2102.05095. DOI: 10.48550/arXiv.2102.05095.

[13] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video Transformer Network," *arXiv preprint arXiv:2102.00719*, Feb. 2021. [Online]. Available: https://arxiv.org/abs/2102.00719. DOI: 10.48550/arXiv.2102.00719.

[14] H. Alamri, A. Bilic, M. Hu, A. Beedu, and I. Essa, "End-to-End Multimodal Representation Learning for Video Dialog," *arXiv preprint arXiv:2210.14512*, Oct. 2022. [Online]. Available: https://arxiv.org/abs/2210.14512. DOI: 10.48550/arXiv.2210.14512.

[15] C. S. C., "Multimodal attention-based transformer for video captioning," *Appl. Intell.*, vol. 53, pp. 23349–23368, 2023.

[16] Ye Liu, Si-Yuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiao-Hu Qie, "UMT: Unified Multi-modal Transformers for Joint Video Moment Retrieval and Highlight Detection," *arXiv e-prints*, Mar. 2022. [Online]. Available: https://arxiv.org/abs/2203.12745. DOI: 10.48550/arXiv.2203.12745.

[17] J. Johnson, A. Karpathy, and Fei-Fei Li, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4565-4574.

[18] Da-Li Yang and Chun Yuan, "Hierarchical Context Encoding for Events Captioning in Videos," *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 1288-1292.

[19] Jing-Wen Wang, Wen-h-Hao Jiang, Lin Ma, Wei Liu, and Yong Xu, " Bidirectional Attentive Fusion With Context Gating for Dense Video Captioning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7190-7198.

[20] Luo-Wei Zhou, Ying-Bo Zhou, J. J. Corso, R. Socher, and Cai-Ming Xiong, "End-to-End Dense Video Captioning with Masked Transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 8739–8748.

[21] Ye-Hao Li, Ting Yao, Ying-Wei Pan, Hong-Yang Chao, and Tao Mei, "Jointly localizing and describing events for dense video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7492–7500.

[22] M. J. Mun, Lin-Jie Yang, Zhou Ren, et al., "Streamlined dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 6588–6597.

[23] Lun Huang, Wen-Ming Wang, Jie Chen, and Xiao-Yong Wei, "Attention on Attention for Image Captioning," *arXiv e-prints*, Aug. 2019, [Online]. Available: https://arxiv.org/abs/1908.06954. DOI: 10.48550/arXiv.1908.06954.

[24] Teng Wang, Hui-Cheng Zheng, Ming-Jing Yu, et al., "Event-centric hierarchical representation for dense video captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1890–1900, May 2020.

[25] F. Rahutomo, and A. H. Ayatullah, "Indonesian dataset expansion of Microsoft Research Video Description Corpus and its similarity analysis," Kinetik: Game Technol., Inf. Syst., Comput. Netw., Comput. Electron., vol. 2018, pp. 26–319, 2018.

[26] Zi-Wei Yang, You-Jiang Xu, Hui-Yun Wang, et al., "Multirate multimodal video captioning," in *Proc. 25th ACM Int. Conf. Multimedia (MM '17)*, Mountain View, CA, USA, Oct. 23–27, 2017, New York: ACM, pp. 1877-1882.

[27] K. Papineni, S. Roukos, T. Ward, et al., "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, 2002, pp. 311–318.
[28] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *in Proc. Workshop Text Summarization Branches Out, Post-Conf. Workshop of ACL 2004*, 2004, pp. 74–81.
[29] M. Denkowski and A. Lavie, "Meteor Universal: Language-specific translation evaluation for any target language," *in Proc. 9th Workshop Stat. Mach. Translation*, 2014, pp. 376–380.
[30] R. Vedantam, L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.
[31] WANG B, MA L, ZHANG W, et al., "Controllable video captioning with POS sequence guidance based on gated fusion network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 142-103.
[32] Zi-Qi Zhang, Ya-Ya Shi, Chun-Feng Yuan, et al., "Object relational graph with teacher-recommended learning for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, 2020, pp. 13275–13285.
[33] Qi Zheng, Chao-Yue Wang, Da-Cheng Tao, "Syntax-aware action targeting for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, 2020, pp. 13093–13102.
[34] Bang Yang, Yue-Xian Zuo, Feng-Lin Liu, et al., "Non-autoregressive coarse-to-fine video captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3119–3127.
[35] H. RYU, S. KANG, H. KANG, et al., "Semantic grouping network for video captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2514–2522.