# Action-aware Linguistic Skeleton Optimization Network for Non-autoregressive Video Captioning

SHUQIN CHEN, School of Computer Science and Hubei Provincial Collaborative Innovation Center for Basic Education Information Technology Services, Hubei University of Education, Wuhan, China
XIAN ZHONG, Hubei Key Lab of Transportation Internet of Things, School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China and Rapid-Rich Object Search Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
YI ZHANG, School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China
LEI ZHU, ROAS Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China and Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
PING LI, Department of Computing and School of Design, The Hong Kong Polytechnic University, China
XIAOKANG YANG, MOE Key Laboratory of AI, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
BIN SHENG, Department of Computer Science and Engineering, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China

This work was supported in part by the National Natural Science Foundation of China under Grant 62271361; in part by the Natural Science Foundation of Hubei Province under Grant 2023AFB206; in part by the Hubei Institute of Education Science under Grant 2022ZA41; in part by the Scientific Research Foundation of Hubei University of Education for Talent Introduction under Grant ESRC20230009; in part by the Fundamental Research Funds for the Central Universities under Grant WHUTIOT2023-006; in part by the Hubei Provincial Collaborative Innovation Center for Basic Education Information Technology Services under Grant OFHUE202305.

Authors' Contact Information: Shuqin Chen, School of Computer Science and Hubei Provincial Collaborative Innovation Center for Basic Education Information Technology Services, Hubei University of Education, Wuhan, China; e-mail: csqcwx0801@163.com; Xian Zhong (corresponding author), Hubei Key Lab of Transportation Internet of Things, School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China and Rapid-Rich Object Search Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore; e-mail: zhongx@whut.edu.cn; Yi Zhang, School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China; e-mail: yi_zhang@whut.edu.cn; Lei Zhu, ROAS Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China and Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China; e-mail: leizhu@ust.hk; Ping Li, Department of Computing and School of Design, The Hong Kong Polytechnic University, Hong Kong, China; e-mail: p.li@polyu.edu.hk; Xiaokang Yang, MOE Key Laboratory of AI, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China; e-mail: xkyang@sjtu.edu.cn; Bin Sheng (corresponding author), Department of Computer Science and Engineering, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China; e-mail: shengbin@sjtu.edu.cn.

ACM 1551-6865/2024/10-ART326
https://doi.org/10.1145/3679203

Non-autoregressive video captioning methods generate visual words in parallel but often overlook semantic correlations among them, especially regarding verbs, leading to lower caption quality. To address this, we integrate action information of highlighted objects to enhance semantic connections among visual words. Our proposed Action-aware Language Skeleton Optimization Network (ALSO-Net) tackles the challenge of extracting action information across frames, improving understanding of complex context-dependent video actions and reducing sentence inconsistencies. ALSO-Net incorporates a linguistic skeleton tag generator to refine semantic correlations and a video action predictor to enhance verb prediction accuracy in video captions. We also address issues of unsatisfactory caption length and quality by jointly optimizing different levels of motion prediction loss. Experimental evaluation on prominent video captioning datasets demonstrates that ALSO-Net outperforms baseline methods by a significant margin and achieves competitive performance compared to state-of-the-art autoregressive methods with smaller model complexity and faster inference time.

CCS Concepts: • **Computing methodologies → Video summarization**; **Natural language generation**; **Neural networks**;

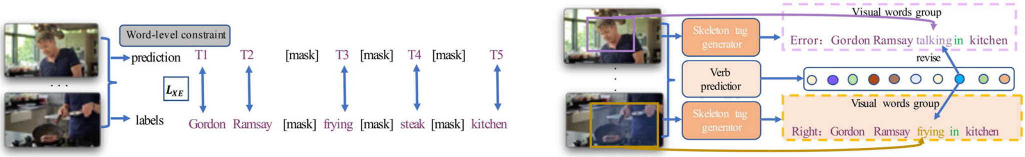
Additional Key Words and Phrases: Video Captioning, Non-Autoregressive Models, Visual-Language Alignment, Video Action Prediction, Semantic Dependencies

**ACM Reference format:**
Shuqin Chen, Xian Zhong, Yi Zhang, Lei Zhu, Ping Li, Xiaokang Yang, and Bin Sheng. 2024. Action-aware Linguistic Skeleton Optimization Network for Non-autoregressive Video Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 10, Article 326 (October 2024), 24 pages.
https://doi.org/10.1145/3679203

## 1 Introduction

Recent advancements in deep learning [21, 26, 42, 48, 69] have significantly propelled research in several cross-modal tasks [34, 50, 51, 70], particularly within multimedia and artificial intelligence. Among these, video captioning has rapidly emerged as a dynamic research field [43, 47, 65], focused on generating coherent natural-language descriptions of visual content. This field has broad applications, ranging from assisting visually impaired individuals to enhancing human–robot interactions.

Autoregressive methods, which sequentially generate captions word by word, are widely employed across various domains and typically utilize an encoder–decoder architecture to maintain visual and semantic coherence during caption generation [4, 33, 54]. However, the inherent sequential nature of these methods can restrict processing speed in certain applications. Consequently, non-autoregressive methods, which generate all words simultaneously, have emerged as a promising alternative due to their low latency. Despite their speed, non-autoregressive methods often suffer from weak dependencies among simultaneously generated words, resulting in captions with repeated or incorrect words. Innovations in **neural machine translation (NMT)** have addressed this issue through the use of latent variables that encode information about the target language sequence [40, 44]. In visual captioning, the integration of local autoregressive and global non-autoregressive methods has been identified as a practical compromise [14]. However, their application in video captioning remains limited, with only one iterative optimization-based method proposed to date [62]. This method, while innovative, requires an impractical number of iterations due to its iterative, sentence-level autoregressive nature. Yang et al. [62] highlight the significant role of visual words in guiding the caption generation process, especially in encouraging the production of scene-related words. Visual words, which correspond to relevant visual objects and actions, serve as a template for the generation of non-visual words, underscoring their importance in video



(a) Traditional Word-level Skeleton Constraints

(b) Intra-frame Dependency Mining and Independent Action Revision

Fig. 1. Illustration of the video captioning process guided by linguistic skeleton tags. (a) Traditional supervision of visual words, (b) Integration of a video action predictor branch in our model, which extracts high-level action information from the entire video to refine low-level action semantics identified within single frames by the skeleton tag generator. This method enhances semantic dependencies among visual words, enabling the effective integration of specific verbs (highlighted in color) and the generation of diverse, action-aware captions tailored to the video content.

captioning. However, the reliability of generating visual words non-autoregressively remains a concern. As shown in Figure 1(a), relying solely on word-level cross-entropy loss is insufficient to fully capture semantic relations within linguistic skeletons, as it fails to consider the linguistic skeleton as a cohesive unit.

In previous research [66], linguistic skeleton tags have been employed as sentence-level supervision for visual word sequences. This method utilizes **dual-scale visual-language bi-directional alignment (DVBA)** to establish the internal relevance of these tags, thereby enhancing the correlation among visual words within a single-frame image. Specifically, the method effectively captures the semantic dependencies among visual words by recognizing the relevance of objects depicted in a video frame. For instance, as shown in the upper part of Figure 1(b), attention mechanisms cluster "Gordon Ramsay" and "kitchen" into visual word groups. These clusters reflect their spatial relationships within the video frame, such as "Gordon Ramsay in the kitchen," indicating the preposition "in." Consequently, these visual word groups significantly strengthen semantic dependencies.

However, existing methods exhibit limitations. Associations captured within a single-frame image may not fully represent the cohesion of the entire video. Relying solely on single-frame image analyses to strengthen visual word dependencies can result in inaccuracies because describing actions within a video often requires summarizing successive observations, and the semantics expressed within a single frame frequently lack comprehensive temporal context. To more clearly distinguish and emphasize interaction levels both within and across frames, we adopt the definition from the field of image captioning [35], which characterizes "high-level semantic information" as a summary description of images. Building on this, we extend the concept to encompass video action semantic information, necessitating a summary across multiple frames. Consequently, for complex videos, low-level action semantic information from individual frames is inadequate compared to the high-level action semantic information conveyed by the entire video. Moreover, in complex videos where multiple events may be unrelated, employing a single-frame target semantics construction method can lead to errors when modeling the semantics of visual word sequences in the target language. To address these challenges, we propose the design of action prediction branches to extract high-level action information from the entire video. These prediction results are then integrated into the inputs of both the **length predictor (LP)** and the decoder, correcting the erroneous constraints imposed by low-level action information extracted by the skeleton tag generator. With this enhanced method, we optimize the entire process of semantic refinement for visual word sequences. We introduce an action-guided branch to accurately determine relevant actions, mitigating erroneous actions suggested by specific frames. This action information is subsequently integrated into the caption generator branch to refine the generation template, thereby improving semantic dependencies between visual words. Thus, enhancing the accuracy of the linguistic skeleton, which includes both noun and verb attributes, is a significant focus of our study. We observe that the correlation between visual words manifests primarily in two aspects: the association among nouns within a single frame and the semantic correlation between nouns and verbs that models the relationship among video objects and facilitates action prediction. By explicitly predicting actions, the captioning process is furnished with additional guidance beyond mere linguistic structuring.

Drawing on this method, we introduce the **Action-Aware Linguistic Skeleton Optimization Network (ALSO-Net)**, a novel method that incorporates skeleton tag prediction and a video **action predictor (AP)** within the Transformer framework. This method establishes semantic relationships among visual words through dual-scale video–text alignment. Additionally, we extend this mapping to the visual word sequences generated by non-autoregressive methods, enhancing semantic dependencies among visual words via the supervision of skeleton tags. This improvement significantly enhances the decoder's capacity to generate detailed captions guided by a more accurate visual context. ALSO-Net operates through a two-step process: it first identifies the target action present in the video and then selects the focal object to be included in the final caption.

Our work makes threefold significant contributions to the field of video captioning:

- We introduce a Video AP branch that dynamically enhances semantic connections among visual words. This branch leverages classification results to extract high-level action information from video frames, reinforcing semantic connections and addressing the complicated contextual dependencies in video actions.
- We develop ALSO-Net, a novel non-autoregressive video captioning framework. ALSO-Net effectively optimizes action prediction losses and reduces inconsistencies in generated sentences by acquiring multi-level visual representations at various granularities and associating them with their linguistic counterparts. This framework uniquely highlights the crucial connection between high-level action information and visual objects in non-autoregressive video captioning.
- We validate our model's high inference efficiency and its capability to produce reliable and coherent captions through extensive testing on **Microsoft research video to text (MSR-VTT)** and **Microsoft research video description (MSVD)** datasets. Our strategic integration of multilevel visual representations at different granularities demonstrates state-of-the-art performance, substantiating the effectiveness of ALSO-Net in practical applications.

## 2 Related Work

### 2.1 Video Captioning

In the field of video captioning, autoregressive methods [33] generate sentences token-by-token, each conditioned on the previously generated tokens, aiming to maximize the joint probability of the target words. However, this sequential word-by-word generation method has practical limitations. To address the shortcomings of non-autoregressive methods, recent advancements have focused on enhancing target sequence relevance through mechanisms such as latent variables, knowledge distillation, and iterative optimization. Nonetheless, these methods typically integrate target sequence relevance either in a one-off or a stepwise manner, which may not capture the full dynamics of sentence construction. In the field of NMT, Ran et al. [39] have developed a method that reconstructs the target sequence from source tokens, effectively incorporating the grammatical structures of the target language into the translation process.

Autoregressive methods are widely used in video captioning, with prior research largely concentrating on improving visual feature extraction and enhancing visual-semantic relevance. Li et al. [28] develop the **long short-term relation transformer (LSRT)** method, which constructs a video-specific long short-term graph and employs the G3RM for relational reasoning, effectively resolving issues such as redundant connections, over-smoothing, and ambiguity in relationships within video content. Additionally, Deng et al. [10] introduce the **syntax-guided hierarchical attention network (SHAN)** model, which integrates semantic and syntactic cues to better combine visual and contextual features in captioning, significantly improving the generation of non-visual words and model interpretability. In the domain of non-autoregressive frameworks, Fei et al. [13] introduce a method that uses location information to guide the generation of descriptions. Gao et al. [15] propose the Hierarchical Representation Network with Auxiliary Tasks, a framework aimed at high-quality video understanding enriched with semantics-aware cues. Yang et al. [62] utilize a masked language model to facilitate parallel caption generation, enhancing the accuracy and diversity of the captions. However, this method still relies on sequential decoding processes, which do not fully capture the semantic relationships among visual words, thus leaving room for further improvement in nonsequential caption generation techniques.

### 2.2 Non-Autoregressive Sequence Generation

The translation speed of NMT models is hindered by their autoregressive nature, which decodes target sentences word-by-word based on the translation history. In contrast, **non-autoregressive translation (NAT)** models decode all target words simultaneously, effectively overcoming the speed limitations associated with autoregressive translation. NAT achieves this by iteratively refining the initial translations using the source language input, which conceptually mimics sentence-level autoregressive processing. A notable method within NAT involves the use of knowledge distillation [67], where an autoregressive model, adept at detailed modeling, serves as a "teacher." The NAT, acting as a "student," learns the translation distributions block by block. This strategy allows NAT to achieve rapid translation speeds without sacrificing the quality of the output.

In video captioning, NAT [14, 62] eliminate sequential dependencies, enabling the simultaneous generation of all words and thus speeding up decoding. To address the performance gap between autoregressive and non-autoregressive captioning models, various strategies have been introduced, such as knowledge distillation, the incorporation of auxiliary regularization terms, and the optimization of decoder inputs. However, these methods often rely on traditional cross-entropy loss during training, which does not ensure sentence-level consistency.

### 2.3 Action-Aware Video Captioning

Pretrained 3D **convolutional neural network (CNN)** models such as Convolutional 3D [49], Inflated 3D ConvNet [3], and ResNet 3D [19] are commonly used as feature extractors in video captioning tasks to capture motion information. These models are typically pre-trained on motion recognition datasets like Kinetics-400 [23] or UCF-101 [46], excelling at detecting motion in video frames. Previous research [36, 38, 41, 62] has shown promising results by integrating 3D and 2D CNN features to achieve a more comprehensive video representation. Instead of directly merging 3D and 2D CNN features, other studies [1, 64] have explored novel methods to utilize 3D CNN features, also known as action features. Zheng et al. [64] apply the self-attention mechanism of the Transformer to capture global dependencies among multiple objects and extract action semantic information based on sequentially decoded subject–predicate–object syntax. However, this method involves serial decoding and incurs a quadratic computational cost relative to the large number of objects in a video, which contradicts the goal of fast, parallel decoding inherent in non-autoregressive generation algorithms. Bai et al. [1] explore object-level interaction and frame-level information through a conditional graph, but their method does not address the semantic correlation between nouns and verbs in video captions or explicitly predict actions. Furthermore, most autoregressive algorithms primarily rely on object-level features to extract action semantics, often overlooking the potential of appearance features from individual frames and motion features

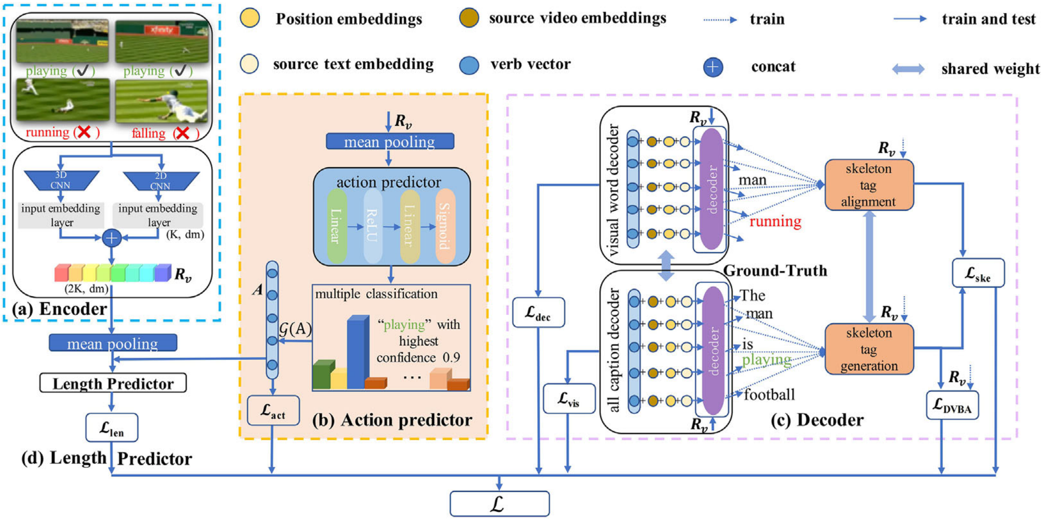


Fig. 2. Proposed ALSO-Net framework with dual branches: the video AP and the caption generator, both employing DVBA. (a) Inputs are sequences of video frames or clips of specific lengths, processed by pre-trained 2D/3D CNNs to extract visual features. (b) Video AP branch identifies key verbs of interest for the caption, which may be predicted by the DVBA-based caption generator branch, specified by the user, or predefined. (c) Caption generator branch then constructs the remaining words around these focused verbs to form a coherent caption. (d) LP estimates the overall length of the target caption. The entire model undergoes training with a joint optimization loss function to enhance performance.

from consecutive frames to generalize the primary content of the video, thus missing out on capturing global contextual information.

In contrast to previous studies, our work introduces a non-autoregressive video captioning algorithm that focuses on optimizing global high-level actions. We begin by addressing the limitations of traditional models in semantically modeling visual word sequences, particularly their failure to accurately recognize high-level actions in certain scenes. To overcome this issue, we employ a novel action recognition method that rectifies the limitations of single-frame action semantics. Our AP branch streamlines the process by compressing complete visual information and treating verb prediction as a multi-classification task. This method not only mitigates issues related to redundant connections but also significantly reduces the model's parameter count through the use of two linear layers in the prediction head. Through the integration of action features that capture object interactions, our method achieves a higher accuracy in generating captions that faithfully represent the involved actions and objects. To the best of our knowledge, this marks the initial incorporation of prominently emphasized action features in non-autoregressive video captioning, thereby representing a substantial advancement in the field.

## 3 Proposed Method

### 3.1 Overall of Framework

The proposed ALSO-Net, depicted in Figure 2, consists of two main branches: a video AP branch and a caption generator branch. The video AP branch is trained to predict actions within the video by minimizing an element-wise logistic loss function, and its outputs are integrated into the caption generator. The caption generator branch is composed of four key components: a Transformer-based video caption generator, a skeleton tag generation module, a skeleton alignment module, and a specially designed DVBA-based loss. In subsequent sections, we will provide detailed descriptions

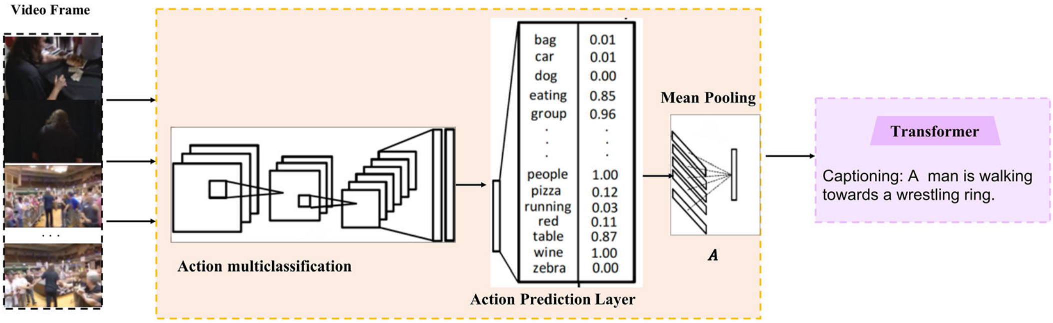


Fig. 3. Action prediction model fine-tuned on the target multi-label dataset. For each test video frame, selected proposal regions are processed through a shared CNN. The outputs from various proposals are aggregated via mean pooling to produce the final multi-label prediction, yielding a high-level action representation denoted as $\boldsymbol{A}$.

of these components and explain how they are synergistically combined within our framework to generate more accurate visual captions. The entire model is trained through the joint optimization of the loss function.

### 3.2 Visual Encoder

Both advanced and traditional video captioning methods commonly employ the encoder–decoder framework. Here, image and motion features are processed separately by two distinct encoders, resulting in a sequence of consecutive video clips with a length $K$. These multi-modal features, including both image and motion features, are concatenated to create $\boldsymbol{R}$, which is subsequently input into the decoder.

To reduce video redundancy, we initially sample a fixed number of frames (length $K$). These frames are then fed into pre-trained 2D/3D CNNs to extract two types of visual features: $\boldsymbol{V}_a = \{v_k\}_{k=1}^{K} \in \mathbb{R}^{K \times d_v}$ and $\boldsymbol{V}_m = \{v_k\}_{k=1}^{K} \in \mathbb{R}^{K \times d_v}$. These representations are further encoded using an **input embedding layer (IEL)**, resulting in $f_{\mathrm{IEL}}(\boldsymbol{V}_v) = \boldsymbol{R} \in \mathbb{R}^{K \times d_m}$, which can be formalized as follows:

$$f_{\mathrm{IEL}}\left(\boldsymbol{V}_v\right) = \mathrm{BN}\left(\boldsymbol{G}\bar{\boldsymbol{V}} + \left(1 - \boldsymbol{G}\right)\hat{\boldsymbol{V}}\right), \tag{1}$$

where $\bar{\boldsymbol{V}} = \boldsymbol{V}_v \boldsymbol{W}_{e1}$, $\hat{\boldsymbol{V}} = \tanh(\bar{\boldsymbol{V}} \boldsymbol{W}_{e2})$, $\boldsymbol{G} = \sigma(\bar{\boldsymbol{V}} \boldsymbol{W}_{e3})$, $\boldsymbol{V}_v = \boldsymbol{V}_a$ or $\boldsymbol{V}_m$, BN denotes batch normalization, $\boldsymbol{W}_{e1} \in \mathbb{R}^{d_n \times d_m}$, and $\{\boldsymbol{W}_{e2}, \boldsymbol{W}_{e3}\} \in \mathbb{R}^{d_m \times d_m}$. $\sigma$ signifies the sigmoid function. $d_v$ and $d_m$ denote the feature dimensions before and after $f_{\mathrm{IEL}}(\boldsymbol{V}_v)$, respectively. Next, we concatenate these two modalities to obtain $\boldsymbol{R}_v = \{v_i\}_{i=1}^{2K} \in \mathbb{R}^{2K \times d_m}$.

### 3.3 AP

To overcome the constraint of relying solely on single-frame semantics to establish semantic connections among visual words in the target language, and inspired by the success of **non-autoregressive coarse-to-fine (NACF)** [62] where complete visual information is directly compressed to supplement the decoder's visual domain information, we introduce a novel action recognition predictor. This predictor compresses the entire video, predicting the semantics of its primary actions, thereby augmenting the textual domain information of both the LP and the decoder.

The video AP anticipates the action portrayed in the given video, treating verb prediction as a multi-classification task, illustrated in Figure 3. To accomplish this, we employ a simple linear predictor, as shown in Figure 2(b). Initially, we compile an action corpus based on the complete dataset, encompassing all potential action types. Each video is then associated with at least one action type from this corpus. Specifically, we represent the ground truth action as $\boldsymbol{A}^* = \{\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \cdots, \boldsymbol{a}_M^*\} \in \mathbb{R}^M$, where $M$ denotes the number of verbs in the corpus. If a video is associated with action $i$, $\boldsymbol{a}_i^* = 1$; otherwise, $\boldsymbol{a}_i^* = 0$. The formulaic structure of the AP is as follows:

$$\boldsymbol{A} = f_{\text{AP}}\left(\boldsymbol{R}_v\right) = \text{Sigmoid}\left(\text{ReLU}\left(\text{MP}\left(\boldsymbol{R}_v\right)\boldsymbol{W}_{v1}\right)\boldsymbol{W}_{v2}\right), \tag{2}$$

$$\mathcal{G}\left(\boldsymbol{A}\right) = \text{Concat}\left(\boldsymbol{A}\boldsymbol{W}_{v3}, \cdots, \boldsymbol{A}\boldsymbol{W}_{v3}\right) \in \mathbb{R}^{N\times d_m}, \tag{3}$$

where MP denotes mean pooling, $\boldsymbol{W}_{v1} \in \mathbb{R}^{d_m\times d_m}$, $\boldsymbol{W}_{v2} \in \mathbb{R}^{d_m\times M}$, and $\boldsymbol{W}_{v3} \in \mathbb{R}^{M\times d_m}$ are the parameters subject to learning. However, unlike length prediction, the long-tail effect of action classification is pronounced due to the larger number of verb corpora compared to possible lengths. Therefore, treating all categories equally is impractical, even with constraints like **Kullback–Leibler divergence (KLD)** or multiple binary cross-entropy. Otherwise, the actions of all videos may be classified into a limited number of categories. Following Wu et al. [58], we refrain from constraining the negative labels of a single sample to diminish their confidence. Instead, we solely enhance the confidence of positive labels to alleviate the difficulty of such problems. Specifically, we minimize the element-wise logistic loss function $\mathcal{L}_{\text{act}}$ to train the AP:

$$\mathcal{L}_{\text{act}} = \sum_{j=1}^{M} \log\left(1 + \exp\left(-\boldsymbol{a}_j^*\boldsymbol{a}_j\right)\right). \tag{4}$$

Finally, since the dimensions of the predicted results may not match the input dimensions of the subsequent text domain module, a linear mapping step is necessary. This process entails adjusting the dimensions and replicating them $N_m$ times, where $N_m$ represents the maximum length of the input decoder text during the training phase:

$$\bar{\boldsymbol{A}} = \mathcal{G}(\boldsymbol{A}) = \boldsymbol{A}\boldsymbol{W}_{a3}, \tag{5}$$

$$\hat{\boldsymbol{A}} = \text{Concat}(\bar{\boldsymbol{A}}_1, \cdots, \bar{\boldsymbol{A}}_N) \in \mathbb{R}^{N_m\times d_m}. \tag{6}$$

### 3.4 Length Prediction

In contrast to autoregressive methods, which stop decoding upon encountering <EOS> token, non-autoregressive methods typically incorporate an LP to forecast the length distribution $\boldsymbol{L} \in \mathbb{R}^{\mathbb{N}}$. However, beyond the encoder output $\boldsymbol{R}_v$, we enhance the text domain information of the LP with the high-level action semantics $\boldsymbol{A}$ predicted by the AP. This entails using two linear layers for one-hot type length prediction:

$$\boldsymbol{L} = f_{\text{LP}}(\boldsymbol{R}) = \text{Softmax}\left(\text{ReLU}\left(\text{MP}\left(\boldsymbol{R}_v \oplus \boldsymbol{A}\right)\boldsymbol{W}_{l1}\right)\boldsymbol{W}_{l2}\right), \tag{7}$$

where MP and $\oplus$ denote mean pooling and matrix-vector addition, respectively. $\boldsymbol{W}_{l1} \in \mathbb{R}^{d_m\times d_m}$ and $\boldsymbol{W}_{l2} \in \mathbb{R}^{d_m\times N}$ are trainable, with $N$ representing the maximum training length. We minimize KLD between the predicted length distribution $\boldsymbol{L}$ and the ground-truth distribution $\boldsymbol{L}^*$, where each $\boldsymbol{l}_j^*$ in $\boldsymbol{L}^*$ represents the probability of sentences being of length $j$:

$$\mathcal{L}_{\text{len}} = \mathcal{D}_{\text{KL}}\left(\boldsymbol{L}^*\|\boldsymbol{L}\right) = -\sum_{j=1}^{N} \boldsymbol{l}_j^* \log\frac{\boldsymbol{l}_j}{\boldsymbol{l}_j^*}. \tag{8}$$

### 3.5 Caption Decoder and Visual Word Decoder

The cloze training method [11] aligns well with our nonautoregressive method for network training. In this method, given a ground-truth sentence $\boldsymbol{Y}^*$, we randomly replace some tokens with <mask> tokens according to a specified ratio $\alpha$. This results in a partially observed sequence $\boldsymbol{Y}_{\text{obs}}$ and a masked (unobserved) sequence $Y_{\text{mask}} = \boldsymbol{Y}^* \backslash Y_{\text{obs}}$. Subsequently, we present the visual representation $\boldsymbol{R}_v$ along with $Y_{\text{obs}}$ to the decoder, aiming to predict the probability distribution of $Y_{\text{mask}}$:

$$p_\theta\left(Y_{\text{mask}} | Y_{\text{obs}}, \boldsymbol{R}_v \oplus \boldsymbol{A}\right) = \prod_{\boldsymbol{y}_i \in Y_{\text{mask}}} f_{\text{dec}}\left(\boldsymbol{y}_i | Y_{\text{obs}}, \boldsymbol{R}_v \oplus \boldsymbol{A}\right), \tag{9}$$

where $f_{\text{dec}}$ represents the transformation within the decoder. We aim to minimize the negative log-likelihood

$$\mathcal{L}_{\text{dec}} = -\log\left(p_\theta\left(Y_{\text{mask}} | Y_{\text{obs}}, \boldsymbol{R}_v \oplus \boldsymbol{A}\right)\right), \tag{10}$$

where $\alpha$ follows a uniform distribution ranging from $\beta_1$ to $\beta_2$, ensuring that the network is trained on examples of varying difficulties.

Similar to constructing training samples in a masked language model, we begin with a ground-truth sentence $\boldsymbol{Y}^*$ of length $N_g$ and replace certain words with <mask> to create the corresponding target sequence $\boldsymbol{Y}^{\text{vis}} = \{\boldsymbol{y}_n^{\text{vis}}\}_{n=1}^{N_g}$

$$\boldsymbol{y}_n^{\text{vis}} = \begin{cases} \boldsymbol{y}_n^* & \text{POS}\left(\boldsymbol{y}_n^*\right) \in \{\text{noun}, \text{verb}\}, \\ \text{<mask>} & \text{otherwise}, \end{cases} \tag{11}$$

where POS$(\cdot)$ denotes the part-of-speech of a word. When generating visual words, we pair the visual representation $\boldsymbol{R}_v$ is paired with $Y_{\text{obs}}^{\text{vis}} = \boldsymbol{\varnothing}_{[\text{vis}]}$ (a sequence containing solely the special token <vis>), and provide this combination to the decoder for visual template generation. The corresponding loss function is formulated as follows:

$$\mathcal{L}_{\text{vis}} = -\sum_{\boldsymbol{y}_n^{\text{vis}} \in \boldsymbol{Y}^{\text{vis}}} \log\left(p_\theta\left(\boldsymbol{y}_n^{\text{vis}} | \boldsymbol{\varnothing}_{\text{<vis>}}, \boldsymbol{R}_v \oplus \boldsymbol{A}\right)\right). \tag{12}$$

### 3.6 Skeleton Tags Generation Module

Visual words, distinct from non-visual ones, faithfully represent elements within video frames. We leverage features extracted from video frames closely linked to particular scene-related words for fusion. Precisely, given a comprehensive ground-truth linguistic skeleton $\boldsymbol{Y}^{\text{vis}} = \{\boldsymbol{y}_n^{\text{vis}}\}_{n=1}^{N_g}$, we utilize attention mechanisms to associate these visual words with fixed-length visual groups $\boldsymbol{G} = \{\boldsymbol{g}_i\}_{i=1}^{2K} \in \mathbb{R}^{2K \times d_m}$

$$\boldsymbol{G} = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\boldsymbol{M}\right)\boldsymbol{V}, \tag{13}$$

$$\boldsymbol{Y}_p^{\text{vis}} = \boldsymbol{Y}^{\text{vis}} + \text{PE}\left(1, N_g\right), \tag{14}$$

where $\boldsymbol{Q} = \boldsymbol{R}_v\boldsymbol{W}_q$, $\boldsymbol{K} = \boldsymbol{Y}_p^{\text{vis}}\boldsymbol{W}_k$, and $\boldsymbol{V} = \boldsymbol{Y}_p^{\text{vis}}\boldsymbol{W}_v$. In Equation (14), PE$(\cdot)$ introduces position embeddings to ensure the distinctiveness of visual words in different positions during mapping. Here, $d_k$ represents the dimension of $\boldsymbol{K}$, while $\boldsymbol{M}$ serves to mitigate the influence of aligning <mask>. Represented as a $2K$ by $N_g$ matrix, each row vector $\boldsymbol{V}_s = \{v_n^{\text{vis}}\}_{n=1}^{N_g}$ is defined as follows:

$$v_n^{\text{vis}} = \begin{cases} 1 & \text{if } \boldsymbol{y}_n^{\text{vis}} \in \boldsymbol{Y}^{\text{vis}} \text{ and } \boldsymbol{y}_n^{\text{vis}} = \boldsymbol{y}_n^* \\ 0 & \text{if } \boldsymbol{y}_n^{\text{vis}} \in \boldsymbol{Y}^{\text{vis}} \text{ and } \boldsymbol{y}_n^{\text{vis}} = \text{<mask>}. \end{cases} \tag{15}$$

Following Equation (13), the variable-length annotated visual word sequence is transformed into fixed-length visual word sequence $\boldsymbol{S}$ (*i.e.*, Concat$(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_{2K}) \in \mathbb{R}^{1 \times 2K \times d_m}$). Similarly, we map the non-fixed predicted linguistic skeleton into the same space, denoted as $\boldsymbol{S}^{\text{pre}} = \text{Concat}(\boldsymbol{g}_1^{\text{pre}}, \cdots, \boldsymbol{g}_{2K}^{\text{pre}}) \in \mathbb{R}^{1 \times 2K \times dm}$ using Equations (13), (14), and (15).

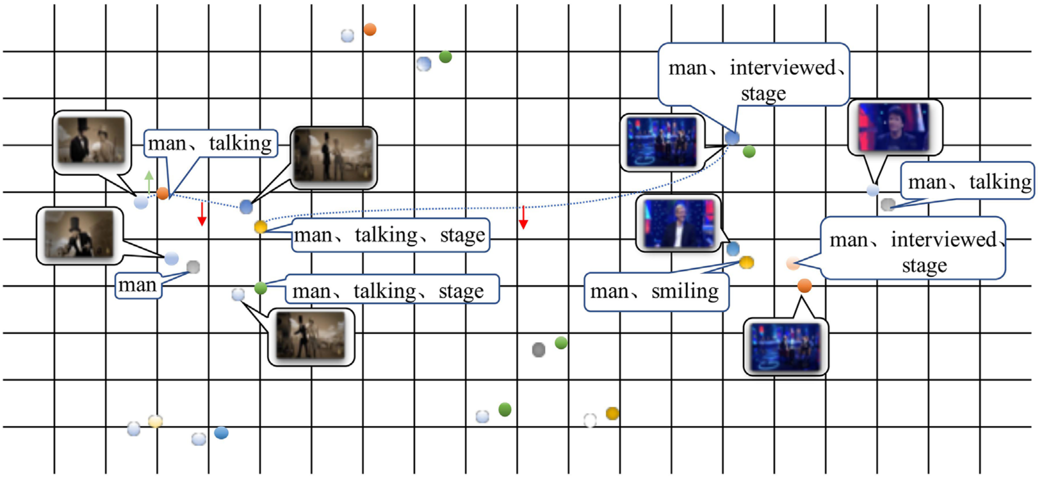


Fig. 4. DVBA loss implemented to reduce disparity between video frames and corresponding visual word groups in video–text alignment. The upward green arrow signifies a decrease in the gap between the two hidden states, and the downward red arrow indicates an increase in this gap.

## 3.7 DVBA–Based Loss

Skeleton-level tags focus on the distribution within the tags. Although $S$ provides uniqueness crucial for ground-truth tags, it lacks the ability to establish relationships between visual words in different positions. To enrich $S$ with intra-dependencies, constraints are necessary during the tag generation process. We attribute physical meaning to the tags generated in the preceding section, where each visual group ($\boldsymbol{g}_i \in \boldsymbol{G}$) corresponds to one video frame, denoted as $\boldsymbol{v}_i \in \boldsymbol{R}_v$. This method ensures the intra-dependencies of skeleton tags by aligning visual and linguistic information. Therefore, we devise a dual-scale alignment method more suitable for skeleton tag generation, as depicted in Figure 4.

### 3.7.1 *Intra-Tag Alignment.*

By minimizing the cosine similarity between $\boldsymbol{g}_i$ and $\boldsymbol{v}_i$, we establish positive alignment between frames and visual groups, formulated as

$$\cos\left(\boldsymbol{g}_i, \boldsymbol{v}_i\right) = \frac{\boldsymbol{g}_i {\boldsymbol{v}_i}^T}{\left\|\boldsymbol{g}_i\right\| \left\|\boldsymbol{v}_i\right\|}. \tag{16}$$

While we address redundant continuous frames in the video, a challenge persists due to the abundance of similar visual groups post-alignment. To distinguish between different visual groups $\boldsymbol{g}_i \in \boldsymbol{G}$, we employ $\boldsymbol{g}_j \in \boldsymbol{G}$ as negative samples. As a result, the intra-tag alignment losses are defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{intra-t2v}} &= -\frac{1}{2K}\sum_{i}^{2K} \log \frac{\exp\cos\left(\boldsymbol{g}_i, \boldsymbol{v}_i\right)}{\sum_j^{2K} \exp\cos(\boldsymbol{g}_i, \boldsymbol{v}_j)}, \\ \mathcal{L}_{\text{intra-v2t}} &= -\frac{1}{2K}\sum_{i}^{2K} \log \frac{\exp\cos\left(\boldsymbol{g}_i, \boldsymbol{v}_i\right)}{\sum_j^{2K} \exp\cos(\boldsymbol{g}_j, \boldsymbol{v}_i)}, \end{aligned} \tag{17}$$

where $\mathcal{L}_{\text{intra-t2v}}$ and $\mathcal{L}_{\text{intra-v2t}}$ denote the loss functions for text-to-video and video-to-text alignment, respectively.

*3.7.2 Inter-Tag Alignment.* To distinguish between similar captions and their corresponding videos, we perform inter-tag alignment. Initially, we compute the cosine similarity between positive pairs $(\boldsymbol{G}_k \in \{G\}_1^B, (\boldsymbol{R}_v)_k)$, where $\boldsymbol{G}_k$ corresponds to $(\boldsymbol{R}_v)_k$, expressed as

$$\cos\left(\boldsymbol{G}_k, (\boldsymbol{R}_v)_k\right) = \sum_{i=1}^{2K} \cos\left(\boldsymbol{g}_i, \boldsymbol{v}_i\right). \tag{18}$$

Next, we utilize $(\boldsymbol{G}_k, (\boldsymbol{R}_v)_m)$ (where $\boldsymbol{G}_k$ does not correspond to $(\boldsymbol{R}_v)_m$ in a batch $B$) as negative pairs. The optimization objective is to maximize $\cos(\boldsymbol{G}_k, (\boldsymbol{R}_v)_k)$ and minimize $\cos(\boldsymbol{G}_k, (\boldsymbol{R}_v)_m)$, which are formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{inter-t2v}} &= -\frac{1}{B}\sum_{k}^{B} \log \frac{\exp\cos\left(\boldsymbol{G}_k, (\boldsymbol{R}_v)_k\right)}{\sum_m^B \exp\cos\left(\boldsymbol{G}_k, (\boldsymbol{R}_v)_m\right)}, \\ \mathcal{L}_{\text{inter-v2t}} &= -\frac{1}{B}\sum_{k}^{B} \log \frac{\exp\cos\left(\boldsymbol{G}_k, (\boldsymbol{R}_v)_k\right)}{\sum_m^B \exp\cos\left(\boldsymbol{G}_m, (\boldsymbol{R}_v)_k\right)}, \end{aligned} \tag{19}$$

where $\mathcal{L}_{\text{inter-t2v}}$ and $\mathcal{L}_{\text{inter-v2t}}$ represent the loss functions for video-to-text and text-to-video alignment, respectively.

*3.7.3 DVBA Overall Loss Function.* We adopt the parameter settings of the previous research work [66], combining the losses from the two scales described above results in the dual-scale alignment loss:

$$\mathcal{L}_{\text{DVBA}} = \mathcal{L}_{\text{intra-t2v}} + \mathcal{L}_{\text{intra-v2t}} + \mathcal{L}_{\text{inter-t2v}} + \mathcal{L}_{\text{inter-v2t}}. \tag{20}$$

## 3.8 Training and Inference

*3.8.1 Training.* Skeleton tags encode robust dependencies among visual words, facilitating the regulation of the predicted skeleton sequence. KLD effectively quantifies the disparity between the predicted $\boldsymbol{S}^{\text{pre}}$ and the tags $\boldsymbol{S}$, expressed as

$$\mathcal{L}_{\text{ske}} = \mathcal{D}_{\text{KL}}\left(\boldsymbol{S}^{\text{pre}} \| \boldsymbol{S}\right) = -\sum_{j=1}^{2K\times d_m} \boldsymbol{g}_j^{*} \log \frac{\boldsymbol{g}_j^{\text{pre}}}{\boldsymbol{g}_j^{*}}. \tag{21}$$

To achieve a better equilibrium between visual word generation and DVBA loss functions, we introduce two parameters, $\lambda_1$ and $\lambda_2$. Therefore, the network's overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{len}} + \mathcal{L}_{\text{dec}} + \lambda_1\mathcal{L}_{\text{vis}} + \lambda_2\mathcal{L}_{\text{DVBA}} + \mathcal{L}_{\text{ske}} + \lambda_3\mathcal{L}_{\text{act}}. \tag{22}$$

Following NACF [62], we set the parameters $\mathcal{L}_{\text{len}}$, $\mathcal{L}_{\text{dec}}$, and $\mathcal{L}_{\text{ske}}$ to 1, and $\lambda_1$ to 0.8. Furthermore, for both MSR-VTT and MSVD, we set $\lambda_2$ and $\lambda_3$ to 1.

*3.8.2 Inference.* During the inference phase, we utilize only the caption generator and AP to derive the visual words template, which acts as input to the decoder. This is followed by the generation of the complete caption.

# 4 Experimental Results

## 4.1 Datasets

MSR-VTT dataset [59] includes 10,000 videos spanning 20 distinct categories, each paired with 20 captions created by 1,327 workers. For evaluation purposes, we use publicly available splits: 6,513 videos for training, 497 for validation, and 2,990 for testing.

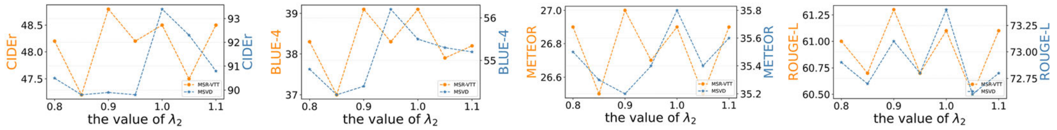


Fig. 5. Illustration of performance across diverse metrics while varying the hyper-parameter $\lambda_2$.

MSVD dataset [18] consists of 1,970 short YouTube clips, each with approximately 40 English captions, totaling 70,028 annotations from Amazon Mechanical Turk workers. Videos range from 10 to 25 seconds. We split the dataset into three subsets: 1,200 videos for training, 100 for validation, and 670 for testing.

VATEX dataset [55] presents 41,269 video clips, each with 10 descriptions, utilized as follows: 25,991 clips for training, 3,000 for validation, and 6,000 for testing.

ActivityNet Captions dataset [25] includes 10,009 training and 4,917 validation videos, averaging 3.65 event segments per video. For practical evaluation, we partition the validation set into two subsets: ae-val with 2,460 videos for validation and ae-test with 2,457 videos for testing.

YouCookII dataset [68] comprises 1,333 training videos and 457 validation videos, with each video averaging 7.7 event segments. We report results on the validation set.

## 4.2 Metrics

To quantitatively evaluate ALSO-Net, we employ four established metrics: BLEU (B-1, 2, 3, and 4) [37], METEOR (M) [2], ROUGE-L (R) [30], and CIDEr (C) [53]. These metrics assess the quality of generated captions by comparing them to ground-truth sentences, with higher scores indicating better sentence generation. CIDEr is particularly valued in captioning tasks for its alignment with human judgment, while BLEU-4 focuses on $n$-gram similarity, indicative of caption fluency. We use the standard evaluation software from MSCOCO [31] server, with a particular emphasis on BLEU-4 and CIDEr due to their relevance in assessing fluency and specificity, respectively.

In addition, we adopt two supplementary metrics commonly used in the image captioning domain [8]: average length and vocabulary usage. The average length metric assesses the typical caption length, and the vocabulary usage metric evaluates the diversity of vocabulary employed in the captions.

## 4.3 Implementation Details

For video feature extraction, we employ ResNet-101-based 2,048-dimensional appearance features from ImageNet and ResNeXt-101-based 2,048-dimensional motion features from Kinetics. The parameter $K$ is set empirically to 8 for each modality. Regarding sequence length, we set $N$ to 30 for MSR-VTT and 20 for MSVD. In our model architecture, we use a single decoder layer with model dimensions $d_m$ of 512, hidden dimensions $d_v$ of 2,048, and 8 attention heads per layer. To regularize the model, we apply dropout with a rate of 0.5 and $\ell_2$ weight decay of 5e-4. Optimization is performed using the adaptive moment estimation (Adam) optimizer [24] for 50 epochs, starting with an initial learning rate of 5e-3. All experiments are conducted on two NVIDIA Tesla PH402 SKU 200 GPUs.

The impact of different $\lambda_2$ values in Equation (22) on the model's performance, considering metrics such as BLEU-4, CIDEr, METEOR, and ROUGE-L, is illustrated in Figure 5. The results reveal a significant influence of $\lambda_2$ variation on the model's performance, particularly concerning CIDEr. Therefore, conducting multiple experiments is vital to achieving a balance between the contributions of different loss components and obtaining optimal results. Following a comprehensive analysis of all metrics, we empirically set $\lambda_2$ to 1.0 for both BLEU-4 and CIDEr, as mentioned earlier. Additionally, we establish the value of $\lambda_2$ as 1.0 for the combined loss functions within the network.

### 4.4 Noisy Parallel Decoding (NPD)

We employ the well-established technique of NPD [56], following NACF [62]. This method consists of three steps: first, the model selects the top-$B$ length candidates from the predicted length $L$ and simultaneously generates $B$ candidate sentences. Next, these $B$ candidates undergo re-scoring using an autoregressive counterpart—an independently trained autoregressive baseline model with the same structure as the original model. Finally, the sentence with the highest confidence score among the $B$ candidates is chosen as the final hypothesis.

### 4.5 Iterative Optimization Strategy

Several non-autoregressive NMT systems employ iterative optimization algorithms at the sentence level to improve translation quality. The primary method involves the following steps: (1) Retaining high-confidence words while replacing relatively low-confidence words with a <mask> token to create a new sequence. (2) Using the newly generated sequence as the input for the decoder's header layer in a second decoding pass to integrate valid information from the secondary decoding input and improve decoding accuracy. (3) Iterating the above two steps multiple times until a predetermined number of iterations is reached. In this study, in addition to the iterative strategy utilized in NACF [62], we explore three distinct iterative strategies. The main difference among these strategies lies in the coverage rate and method of word replacement at each iteration, as described below:

- *Mask-Predict (MP)* [17] follows a coverage rate that linearly decreases with the number of iterations to replace low-confidence words.
- *Easy-First (EF)* [62] prioritizes replacing $q$ words with the highest confidence in each iteration.
- *Left-to-Right (L2R)* [62] extends the EF method by sequentially reserving new $q$ words in each iteration from the previous coverage position, starting from the left, and adding them to the existing sequence for decoding.

### 4.6 Comparisons with State-of-the-Art Methods

Table 1 presents a detailed comparison between our proposed ALSO-Net and existing state-of-the-art autoregressive methods on MSR-VTT and MSVD. As a non-autoregressive video captioning method, ALSO-Net is inherently more lightweight and efficient. Importantly, ALSO-Net not only outperforms current leading models but also delivers competitive results compared to most autoregressive video captioning methods. Compared to RNN-based methods like SibNet [33], ALSO-Net achieves superior performance across four key metrics without employing part-of-speech tagging, surpassing SibNet by margins of 2.2% and 3.5% under BLEU-4 and CIDEr, respectively. When compared to FrameSel [29], our model achieves substantial improvements, with gains of 10.5%, 4.7%, and 26.7% in these metrics, respectively. While ALSO-Net slightly trails behind STGCN [36] in ROUGE-L on MSVD, it ranks second, underscoring its effectiveness. Although ALSO-Net does not surpass STGCN and ORG-TRL [63] in terms of METEOR and CIDEr metrics on MSVD, its strength lies in its rapid inference capabilities. Additionally, ALSO-Net methods exhibit near-optimal performance levels while maintaining its fast processing advantage.

On MSR-VTT, our ALSO-Net exhibits strong competitive performance relative to other autoregressive methods. Specifically, when compared to SHAN [10], ALSO-Net with L2R-NP registers improvements of 5.5%, 2.1%, 2.6%, and 4.7% across the respective metrics. Likewise, against LSRT [28], it achieves enhancements of 2.1%, 1.6%, and 3.6% on the latter three metrics. Furthermore, ALSO-Net sets the benchmark among non-autoregressive methods. For instance, compared to NACF [62], it shows substantial gains of 12.9%, 9.1%, 1.5%, and 8.5% across all four metrics. This

Table 1. Performance Comparison of BLEU-4, METEOR, ROUGE-L, and CIDEr Scores with State-of-the-Art Methods on MSR-VTT and MSVD

| | Model | Venue | MSR-VTT | | | | MSVD | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | B-4 | M | R | C | B-4 | M | R | C |
| AT | Two-stream [16] | TPAMI'20 | 39.7 | 27.0 | - | 42.1 | 54.3 | 33.5 | - | 72.8 |
| | STAT [61] | TMM'20 | 39.3 | 27.1 | - | 43.8 | 52.0 | 33.3 | - | 73.8 |
| | VideoTRM [4] | ACM MM'20 | 38.8 | 27.0 | - | 44.7 | - | - | - | - |
| | STGCN [36] | CVPR'20 | 40.5 | 28.3 | 60.9 | 47.1 | 52.2 | 36.9 | 73.9 | 93.0 |
| | MAD-SAP [20] | TIP'20 | 41.3 | 28.3 | 61.4 | 48.5 | 53.3 | 35.4 | 72.0 | 90.8 |
| | SAAT [64] | CVPR'20 | 40.5 | 28.2 | 60.9 | 49.1 | 46.5 | 33.5 | 69.4 | 81.0 |
| | PMI-CAP [6] | ECCV'20 | 42.1 | 28.7 | - | 49.4 | 54.6 | 36.4 | - | 95.1 |
| | ORG-TRL [63] | CVPR'20 | 43.6 | 28.8 | 62.1 | 50.9 | 54.3 | 36.4 | 73.9 | 95.2 |
| | SBAT [22] | IJCAI'20 | 42.9 | 28.9 | 61.5 | 51.6 | 53.1 | 35.3 | 72.3 | 89.5 |
| | TTA [52] | PR'21 | 41.4 | 27.7 | 61.1 | 46.7 | 52.0 | 34.0 | 70.5 | 81.2 |
| | SibNet [33] | TPAMI'21 | 41.2 | 27.8 | 60.2 | 48.6 | 55.7 | 35.5 | 72.6 | 88.8 |
| | AR-B [62] | AAAI'21 | 42.0 | 28.7 | - | 49.1 | 48.7 | 35.3 | - | 91.8 |
| | SGN [41] | AAAI'21 | 40.8 | 28.3 | 60.8 | 49.5 | 52.8 | 35.5 | 72.9 | 94.3 |
| | MGRMP [7] | ICCV'21 | 41.7 | 28.9 | 62.1 | 51.4 | 55.8 | 36.9 | 74.5 | 98.5 |
| | FrameSel [29] | TCSVT'22 | 38.4 | 27.2 | 59.7 | 44.1 | 50.4 | 34.2 | 70.4 | 73.7 |
| | SHAN [10] | TCSVT'22 | 39.7 | 28.3 | 60.4 | 49.0 | 54.3 | 35.3 | 72.2 | 91.3 |
| | LSRT [28] | TIP'22 | 42.6 | 28.3 | 61.0 | 49.5 | 55.6 | 37.1 | 73.5 | 98.5 |
| | TVRD [57] | TCSVT'22 | 43.0 | 28.7 | 62.2 | 51.8 | 50.5 | 34.5 | 71.7 | 84.3 |
| | R-ConvED [5] | TOMM'23 | 40.4 | 28.1 | - | 47.9 | 53.5 | 34.6 | - | 82.4 |
| | EFFECT [12] | TOMM'23 | 41.4 | 28.4 | 60.5 | 48.8 | 56.9 | 36.6 | 74.2 | 98.5 |
| | RSFD [65] | AAAI'23 | 43.4 | 29.3 | 62.3 | 53.1 | 51.2 | 35.7 | 72.9 | 96.7 |
| NAT | NACF [62] (baseline)[a] | AAAI'21 | 37.1 | 26.5 | 61.1 | 47.3 | 54.1 | 35.2 | 73.5 | 91.0 |
| | O2NA [32] | ACL'21 | 41.6 | 28.5 | **62.4** | 51.1 | 55.4 | **37.4** | **74.5** | **96.4** |
| | ALSO-Net *w* NPD (ours) | | 39.1 | 27.0 | 61.3 | 48.9 | 55.5 | 35.8 | 73.7 | 93.4 |
| | ALSO-Net *w* L2R-NPD (ours) | | **41.9** | **28.9** | 62.0 | **51.3** | **55.7** | 35.9 | 73.0 | 89.0 |

[a] indicates the reproduced method. Best results are highlighted in bold.

highlights our method's proficiency in understanding semantic relationships among visual elements in complex scenarios. In conclusion, as evidenced by Table 1, ALSO-Net not only corrects inaccuracies in caption sequences but also significantly elevates the quality of the generated captions. On MSR-VTT, it achieves remarkable improvements over FrameSel and delivers performance on par with the best video captioning methods on MSVD. Notably, our DVBA-based caption generator is instrumental in achieving outstanding results across various metrics assessing visual word sequences.

On MSR-VTT, index scores are notably lower compared to those on MSVD, which can be attributed to several factors. Primarily, MSVD typically features videos depicting single events, whereas MSR-VTT often includes videos that correspond to multiple distinct events. This discrepancy significantly increases the complexity of the captioning task. Moreover, the uniform sampling of visual features as model inputs may lead to our AP within ALSO-Net inadvertently using irrelevant video features for action prediction, which can disrupt the captioning process. Nonetheless, when compared with O2NA [32], a model that leverages a multi-input guidance corpus to enhance the diversity of video descriptions, our ALSO-Net achieves improvements of 0.7%, 1.4%, and 0.3% in

Table 2. Performance Comparison of BLEU-4, METEOR, ROUGE-L, and CIDEr Scores on VATEX

| | Method | Venue | B-4 | M | R | C |
|---|---|---|---|---|---|---|
| AT | VATEX [55] | ICCV'19 | 28.4 | 21.7 | 47.0 | 45.1 |
| | ORG-TRL [63] | CVPR'20 | 32.1 | 22.2 | 48.9 | 49.7 |
| NAT | NACF [62] (baseline)[a] | AAAI'21 | 26.7 | 20.2 | 47.0 | 39.1 |
| | ALSO-Net (ours) | | **27.8** | **20.6** | **46.8** | **40.4** |

[a] indicates the reproduced method. Best results are highlighted in bold.

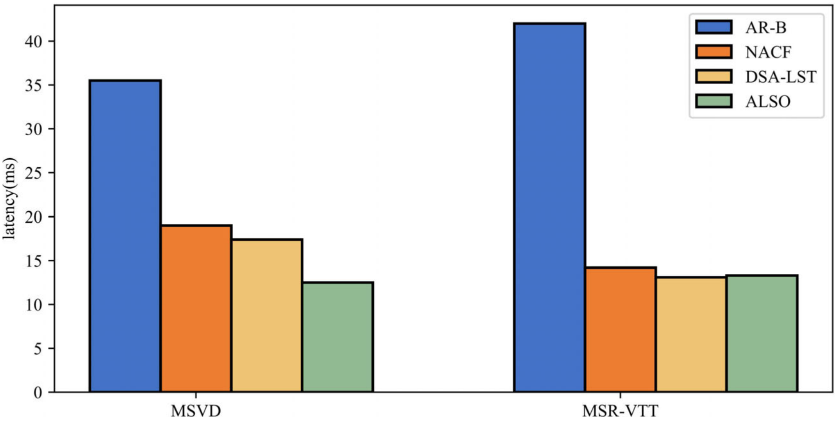


Fig. 6. Comparison of latency between autoregressive and non-autoregressive decoding strategies on MSR-VTT and MSVD, highlighting the significant improvement in inference speed achieved by the proposed ALSO-Net.

BLEU-4, METEOR, and CIDEr, respectively. These gains underscore ALSO-Net's enhanced capability to understand complex, context-dependent video actions and to minimize inconsistencies in generated sentences by effectively extracting action information across frames in more complex video content. Furthermore, due to resource constraints, all methods in this study utilize only eight visual frames for decoding. In scenarios where videos capture multiple events, this limited number of frames restricts the amount of effective information available, posing challenges in accurately summarizing the global high-level actions depicted in the videos. Therefore, there is potential to further improve our model's performance by implementing more effective sampling methods and increasing the frequency of sampling.

Additionally, we evaluate the performance of our model on the public testing set of VATEX, as shown in Table 2. Compared to MSVD and MSR-VTT, VATEX features longer titles and presents a greater challenge for non-autoregressive algorithms. Despite these challenges, our model achieves improvements of 4.1%, 2.0%, and 3.3% in BLEU-4, METEOR, and CIDEr, respectively, compared to the baseline. These results highlight the effectiveness of incorporating a video AP branch into our model, which significantly enhances the comprehension of complex, contextually relevant video actions.

In Figure 6, our proposed ALSO-Net shows comparable performance to traditional autoregressive methods while achieving a significant increase in inference speed, without the need for additional features or manual labeling. Autoregressive caption generation methods, like **autoregressive baseline (AR-B)** [62], sequentially generate each word based on the previously generated output,

Table 3. Performance Comparison of Different Variants of the Proposed Skeleton Tags across Various Scales

| Skeleton | Intra | Inter | AP | MSR-VTT | | | MSVD | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | B-4 | M | C | B-4 | M | C |
| ○ | ○ | ○ | ○ | 37.1 | 26.5 | 47.3 | 54.1 | 35.2 | 91.0 |
| ● | ● | ○ | ○ | 37.9 | 26.6 | 47.8 | 54.8 | 35.3 | 91.6 |
| ● | ○ | ● | ○ | 37.4 | 26.8 | 48.7 | 54.2 | 35.2 | 91.3 |
| ● | ● | ● | ○ | 38.8 | 26.6 | 48.8 | 55.3 | 35.3 | 90.7 |
| ● | ● | ● | ● | **39.1** | **27.0** | **48.9** | **55.5** | **35.8** | **93.4** |

Best results are highlighted in bold.

which substantially increases inference latency. Conversely, **non-autoregressive decoding (NAT)** has gained popularity in natural language processing for its ability to generate words in parallel, thereby significantly speeding up inference. Prominent examples of NAT methods include NACF [62] and our ALSO-Net. Our comparison between AR-B and ALSO-Net demonstrates that ALSO-Net not only matches AR-B in performance but also drastically improves decoding efficiency, reducing latency by 28.7 ms. Moreover, ALSO-Net outperforms the non-autoregressive NACF in both model performance and inference speed.

## 4.7 Ablation Study

*4.7.1 Skeleton Tags.* To assess the impact of the proposed skeleton tags and various alignment methods, we conducted an ablation study examining two video-language alignment strategies at different scales. Table 3 details the performance under different experimental conditions, where "Skeleton," "Intra," "Inter," and "AP" represent skeleton tags, intra-tags alignment, inter-tags alignment, and action predictor, respectively.

Analysis of the table's first three rows indicates that both intra-tags and inter-tags alignments significantly enhance the dependencies among visual words. Additionally, using semantic information encoded in the skeleton tags, rather than solely depending on direct language grammar rules, offers clear benefits for our non-autoregressive method. The comparison between the fourth row and the second and third rows reveals that minimizing redundancy in skeleton tags and effectively distinguishing between similar tags yield improved results across various scales while preserving efficiency. The final row of the table demonstrates that our proposed ALSO-Net, which integrates a skeleton tag generator and a video AP, successfully refines the semantic correlation between visual words and rectifies verb errors. This integration facilitates the generation of more precise and descriptive captions.

*4.7.2 AP.* To assess the impact of the proposed video AP branch, we conducted an ablation study within autoregressive frameworks. Table 4 details performance across different conditions, with "*w* AP" denoting the inclusion of the AP. Compared to the baseline, our method demonstrates improvements of 2.2%, 1.2%, 0.3%, and 0.3% on four metrics respectively within the ActivityNet dataset. In the YoucookII dataset, it shows gains of 0.6%, 0.3%, and 1.1% on the latter three metrics. These results clearly affirm the positive influence of the video AP branch across various decoding frameworks, underscoring its wide applicability and effectiveness in enhancing video captioning performance.

To further assess the effectiveness of our proposed ALSO-Net method, we conducted several ablation experiments to clarify the contribution of each component within the model. The outcomes of these evaluations are summarized in Table 5, where "appearance" and "motion" denote the use

Table 4. Performance Comparison of BLEU-4, METEOR, ROUGE-L, and CIDEr Scores with the Proposed AP across Autoregressive Frameworks

| Model | Venue | ActivityNet | | | | YouCookII | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | B-4 | M | R | C | B-4 | M | R | C |
| Trans.-XL [9] | ACL'19 | 10.3 | 14.9 | 30.3 | 21.7 | - | - | - | - |
| MART [27] | ACL'20 | 9.8 | 15.6 | 30.9 | 22.2 | 8.0 | 15.9 | - | 35.7 |
| Memory Trans. [45] | CVPR'21 | 11.7 | 15.6 | - | 26.6 | - | - | - | - |
| VLTinT [60] (baseline)[a] | AAAI'23 | 13.4 | 17.2 | 36.0 | 30.2 | **9.1** | 17.3 | 34.3 | 43.7 |
| VLTinT *w* AP (ours) | | **13.7** | **17.4** | **36.1** | **30.3** | 8.9 | **17.4** | **34.4** | **44.2** |

[a] Indicates the reproduced method. Best results are highlighted in bold.

Table 5. Performance Comparison of Various Iterations of the Proposed AP across Different Types of Features

| Model | MSR-VTT | | | MSVD | | |
|---|---|---|---|---|---|---|
| | B-4 | M | C | B-4 | M | C |
| ALSO-Net *w/o* AP | 38.8 | 26.6 | 48.8 | 55.3 | 35.3 | 90.7 |
| ALSO-Net *w* appearance | 37.4 | 26.3 | 48.1 | 53.7 | 35.0 | 88.6 |
| ALSO-Net *w* motion | 38.2 | 26.7 | 48.4 | 53.9 | 35.0 | 89.5 |
| ALSO-Net *w* both | **39.1** | **27.0** | **48.9** | **55.5** | **35.8** | **93.4** |

Best results are highlighted in bold.

of appearance and motion features, respectively. The results detail four configurations of ALSO-Net method: "*w/o* AP", which operates without the AP; "*w* appearance" and "*w* motion", which utilize solely the appearance or motion features respectively; and "*w* both", which incorporates both features along with the AP. These ablation studies demonstrate the critical role of the AP in ALSO-Net. It is clear that using only one type of feature–either appearance or motion–along with the AP can limit the model's expressive capacity, resulting in reduced performance. In contrast, the integration of both appearance and motion features with the AP significantly enhances the model's performance, underscoring the importance of a multifaceted method in action prediction.

## 4.8 Comparison with Diversity and Different Decoding Algorithms

*4.8.1 Diversity.* Both of our proposed models consistently outperform NACF on at least one diversity metric, providing a quantitative measure of caption diversity. Notably, our ALSO-Net achieves a significant enhancement in average caption length, with a relative improvement ranging from 6.38% to 6.49% compared to NACF [62] and DSA-LST [66]. This superior performance in average length indicates that the captions generated by ALSO-Net encompass a broader range of information and subtleties present in the video content. This suggests an increased complexity and diversity in the captions, reflecting more nuanced and comprehensive video interpretation.

*4.8.2 Decoding Algorithms.* Table 6 compares the inference performance and latency of ALSO-Net using three different sentence-level iterative optimization algorithms alongside the commonly used NPD strategy [56]. Specifically, MP, EF, and L2R represent the Mask-Predict, Easy-Fit,, and Left-to-Right optimization algorithms, respectively. Additionally, NPD serves as the base optimization strategy for ALSO-Net, with combinations such as MP-NPD, EF-NPD, and L2R-NPD indicating the integration of these iterative algorithms with NPD strategy.

Table 6. Performance Comparison of Different Variants of the Proposed ALSO-Net Using Various Decoding Algorithms

| Algorithm | MSR-VTT | | | | | | | | MSVD | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B-1 | B-2 | B-3 | B-4 | M | R | C | Latency | B-1 | B-2 | B-3 | B-4 | M | R | C | Latency |
| MP | 75.5 | 61.1 | 47.2 | 34.8 | 25.7 | 59.4 | 43.9 | 30.8 ms | 80.2 | 66.3 | 55.7 | 45.1 | 33.6 | 70.8 | 81.5 | 29.5 ms |
| EF | 75.6 | 61.2 | 47.4 | 35.2 | 25.8 | 59.1 | 44.9 | 58.7 ms | 80.1 | 66.1 | 55.4 | 44.7 | 33.5 | 70.6 | 80.8 | 42.0 ms |
| L2R | 76.1 | 62.0 | 48.4 | 36.1 | 26.2 | 59.2 | 43.3 | 50.6 ms | 80.6 | 66.6 | 55.7 | 44.8 | 33.6 | 70.8 | 81.2 | 41.0 ms |
| MP-NPD | 81.2 | 67.2 | 53.5 | 41.1 | 28.4 | 61.6 | 51.0 | 33.7 ms | 84.5 | 74.5 | 65.4 | 55.6 | 35.1 | 72.3 | 89.8 | 30.3 ms |
| EF-NPD | **81.6** | 67.9 | 54.2 | 41.7 | 28.7 | 61.8 | **51.5** | 57.3 ms | 83.4 | 73.1 | 64.0 | 54.2 | 34.5 | 71.6 | 88.0 | 44.2 ms |
| L2R-NPD | **81.6** | **68.0** | **54.4** | **41.9** | **28.9** | **62.0** | 51.3 | 51.4 ms | 84.6 | 74.4 | 65.4 | **55.7** | **35.9** | 73.0 | 89.0 | 44.6 ms |
| NPD | 80.8 | 66.1 | 51.7 | 39.1 | 27.0 | 61.3 | 48.9 | **13.3** ms | **86.0** | **75.5** | **66.0** | 55.5 | 35.8 | **73.7** | **93.4** | **12.5** ms |

Best results are highlighted in bold.

Analysis of the data from the first three rows compared to the fourth through sixth rows reveals that NPD algorithm significantly enhances the metrics while maintaining relatively stable inference latency. This finding suggests that all types of iterative optimization algorithms are compatible with NPD method. The synergy arises because NPD algorithm focuses on ensuring consistency between length prediction and decoding results, while the iterative optimization algorithms prioritize constructing the semantics of the target language. These optimization methods serve distinct yet complementary roles, enhancing the overall effectiveness of the decoding process.

Comparing the results from rows four to six with those of the last row in the table, it becomes apparent that sentence-level iterative optimization algorithms do not always enhance model performance. In fact, for MSVD, there is an observable trend towards over-optimization. This is likely due to the relative simplicity of MSVD, where annotations primarily consist of visual words that reflect straightforward video events. ALSO-Net method, which focuses on modeling the semantics between visual words, can experience adverse effects from the introduction of additional iterative optimization algorithms in such contexts. On the other hand, MSR-VTT benefits more from continuous iterative optimization of the initial decoding sequence based on visual words. This iterative process aids in generating non-visual words, using visual words as a reference. However, it is important to note that this method significantly increases inference time. The delay is primarily caused by the cyclic decoding process of the decoder, compounded by the average length of texts in the training corpus. Therefore, employing iterative optimization to tackle non-autoregressive challenges may not always be the most efficient method, particularly in scenarios where time efficiency is crucial.

### 4.9 Qualitative Results

Figure 7 displays the visualization of cross-modal cross-attention weights on two types of features within our proposed ALSO-Net. These examples highlight the model's accurate verb prediction for target words, ensuring syntactic correctness. Moreover, the model primarily concentrates on the most pertinent features for most words, influenced by both semantic and syntactic cues. This focused attention on word-related features substantially improves the accuracy of predictions for target words. Additionally, the distribution of cross-modal cross-attention weights across features corresponds with human common sense. This alignment provides explicit cues about the reasoning process for each target word, thereby increasing the interpretability of the video captioning process.

Figure 8 presents examples of captions generated by ALSO-Net, contrasting them with those produced by NACF [62]. While NACF captions utilize a refined visual word template and include detailed verbs, ALSO-Net offers significant advantages: (1) *Precision in Verb Description:* In the first

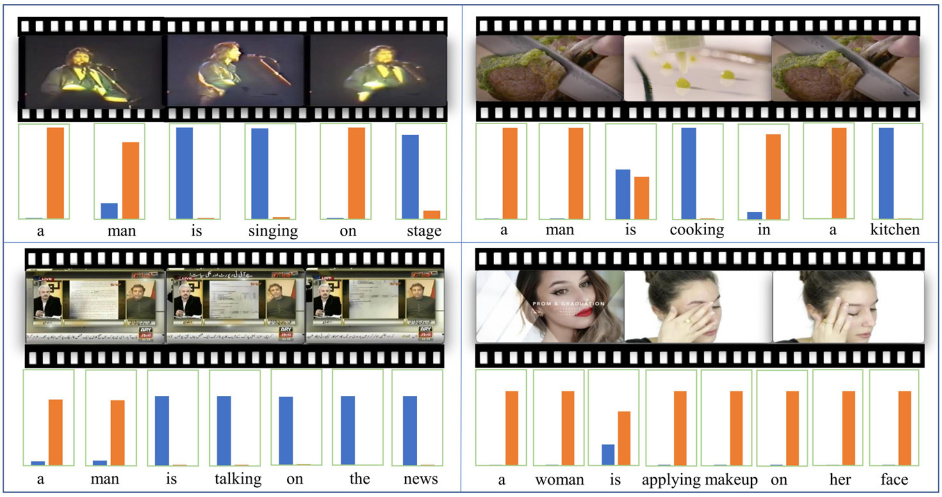
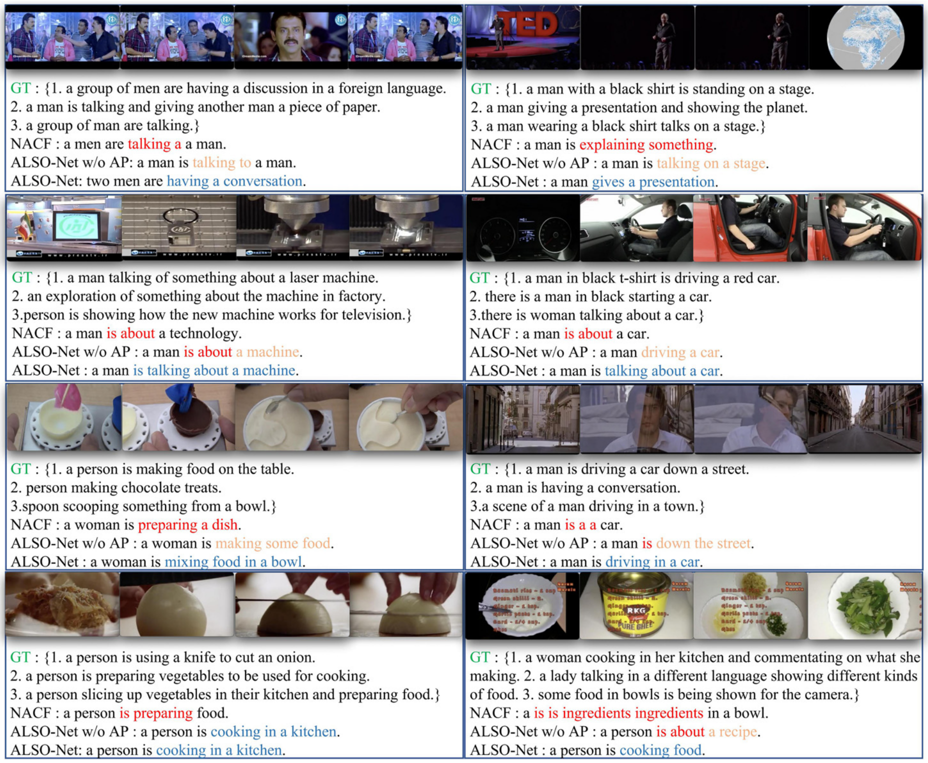


Fig. 7. Visualizations of cross-attention weights in ALSO-Net on MSR-VTT, including histograms of weights assigned to two types of features: 3D motion features (blue bars) and 2D video features (orange bars).



GT : {1. a group of men are having a discussion in a foreign language.
2. a man is talking and giving another man a piece of paper.
3. a group of man are talking.}
NACF : a men are talking a a man.
ALSO-Net w/o AP: a man is talking to a man.
ALSO-Net: two men are having a conversation.

GT : {1. a man with a black shirt is standing on a stage.
2. a man giving a presentation and showing the planet.
3. a man wearing a black shirt talks on a stage.}
NACF : a man is explaining something.
ALSO-Net w/o AP : a man is talking on a stage.
ALSO-Net : a man gives a presentation.

GT : {1. a man talking of something about a laser machine.
2. an exploration of something about the machine in factory.
3.person is showing how the new machine works for television.}
NACF : a man is about a technology.
ALSO-Net w/o AP : a man is about a machine.
ALSO-Net : a man is talking about a machine.

GT : {1. a man in black t-shirt is driving a red car.
2. there is a man in black starting a car.
3.there is woman talking about a car.}
NACF : a man is about a car.
ALSO-Net w/o AP : a man driving a car.
ALSO-Net : a man is talking about a car.

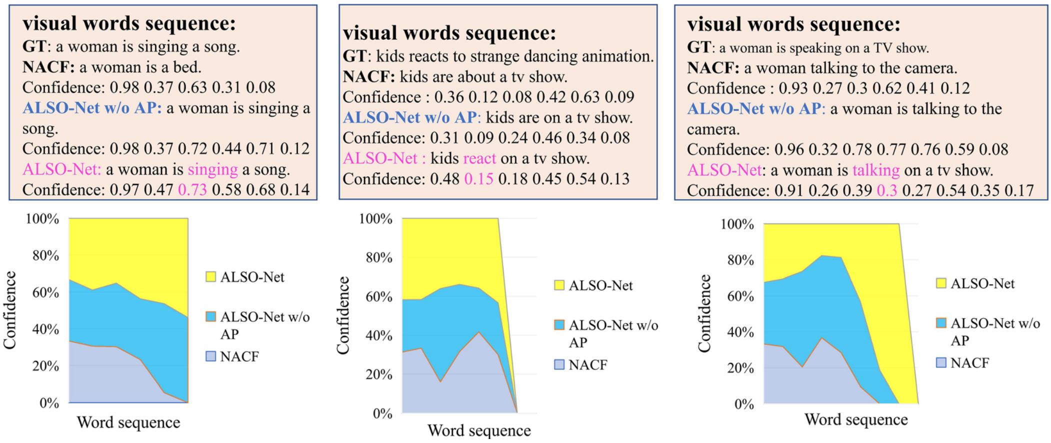
GT : {1. a person is making food on the table.
2. person making chocolate treats.
3.spoon scooping something from a bowl.}
NACF : a woman is preparing a dish.
ALSO-Net w/o AP : a woman is making some food.
ALSO-Net : a woman is mixing food in a bowl.

GT : {1. a man is driving a car down a street.
2. a man is having a conversation.
3.a scene of a man driving in a town.}
NACF : a man is a a car.
ALSO-Net w/o AP : a man is down the street.
ALSO-Net : a man is driving in a car.

GT : {1. a person is using a knife to cut an onion.
2. a person is preparing vegetables to be used for cooking.
3. a person slicing up vegetables in their kitchen and preparing food.}
NACF : a person is preparing food.
ALSO-Net w/o AP : a person is cooking in a kitchen.
ALSO-Net: a person is cooking in a kitchen.

GT : {1. a woman cooking in her kitchen and commentating on what she making. 2. a lady talking in a different language showing different kinds of food. 3. some food in bowls is being shown for the camera.}
NACF : a is is ingredients ingredients in a bowl.
ALSO-Net w/o AP : a person is about a recipe.
ALSO-Net : a person is cooking food.

Fig. 8. Examples of videos paired with corresponding captions from MSR-VTT, comparing captions generated by NACF [62], our proposed ALSO-Net without the AP (ALSO-Net *w/o* AP), and ground-truth captions by human annotators. Words highlighted in blue demonstrate that ALSO-Net effectively emphasizes fine-grained verbs and selects superior visual word groups, producing high-quality video descriptions.



**visual words sequence:**
**GT**: a woman is singing a song.
**NACF:** a woman is a bed.
Confidence: 0.98 0.37 0.63 0.31 0.08
**ALSO-Net w/o AP:** a woman is singing a song.
Confidence: 0.98 0.37 0.72 0.44 0.71 0.12
ALSO-Net: a woman is singing a song.
Confidence: 0.97 0.47 0.73 0.58 0.68 0.14

**visual words sequence:**
**GT**: kids reacts to strange dancing animation.
**NACF:** kids are about a tv show.
Confidence : 0.36 0.12 0.08 0.42 0.63 0.09
**ALSO-Net w/o AP**: kids are on a tv show.
Confidence: 0.31 0.09 0.24 0.46 0.34 0.08
ALSO-Net : kids react on a tv show.
Confidence: 0.48 0.15 0.18 0.45 0.54 0.13

**visual words sequence:**
**GT**: a woman is speaking on a TV show.
**NACF:** a woman talking to the camera.
Confidence : 0.93 0.27 0.3 0.62 0.41 0.12
**ALSO-Net w/o AP**: a woman is talking to the camera.
Confidence: 0.96 0.32 0.78 0.77 0.76 0.59 0.08
ALSO-Net: a woman is talking on a tv show.
Confidence: 0.91 0.26 0.39 0.3 0.27 0.54 0.35 0.17

Fig. 9. Example of linguistic skeleton optimization on MSR-VTT. We display the results of skeleton generation obtained in the first stage. Additionally, the full captions generated in the second stage, along with the confidence scores for each word, are presented as area line graphs. The size of the generated area corresponds to the quality of the sentences produced by our proposed method, with larger areas indicating superior sentence generation performance.

and second examples, NACF yields incorrect descriptions involving small objects, such as "talking a a man." In contrast, ALSO-Net accurately generates descriptions like "talking to a man," showcasing its superior ability to capture enhanced action features essential for precise verb articulation. (2) *Guidance through Skeleton Tags*: ALSO-Net uses skeleton tags to refine the coarse visual template, providing crucial guidance for generating captions. For instance, in the third and fourth examples, ALSO-Net accurately identifies the sequence "talking about a machine," which serves as a vital guideline for the caption model. This use of skeleton tags not only ensures more precise captions but also aids in identifying additional elements, such as "talking about a car." Additionally, in the last two examples, ALSO-Net generates captions that incorporate more visual words and effectively utilizes the AP, resulting in increased caption complexity. This is evident in phrases such as "in a house" and "talking," demonstrating the model's ability to enhance the detail and depth of the narrative.

Figure 9 illustrates the optimization process of ALSO-Net from three perspectives, showing distinct improvements over the NACF method. Specifically, ALSO-Net *w/o* AP enhances the semantic connections between certain visual words and improves comprehension of global higher-order actions. For example, in the second case, ALSO-Net predicts the action "react," suggesting that the children are not only present in a TV program but are also reacting within it. This contrasts with ALSO-Net *w/o* AP model's simpler assertion that they "are in a TV show," a statement based on a single frame, while the action "react" is derived by synthesizing the entire video content. In another instance, ALSO-Net *w/o* AP model mistakenly associates the action with the incorrect subject "camera." In contrast, ALSO-Net correctly interprets the action as "talking," which leads to a more accurate depiction of the spatial setting as "on a TV show." Further analysis of the confidence levels in verb generation by ALSO-Net with and without the AP reveals notable enhancements. In the second example, confidence in the verb increases from 0.09 ("are") to 0.15 ("react"). Similarly, in the third example, confidence jumps from 0.28 ("is") to 0.92 ("is talking"). These results demonstrate that the AP not only encapsulates comprehensive global action information but also significantly boosts confidence in verb accuracy and the generation of spatial semantics. Given that the baseline model primarily depends on a visual word template method, these improvements are particularly significant.

## 5 Conclusion

In this work, we introduce a novel framework, the ALSO-Net, designed to address the limitations of non-autoregressive video captioning models, particularly their inadequate grasp of video action. Our method synergistically combines low-level action semantics from individual frames with high-level action information from the entire video to enrich semantic relations among visual words. To further enhance these relations, we incorporate visual word sequences within skeleton tags to better align video frames. Recognizing that relying solely on individual frames can lead to an insufficient representation of verbs, we have integrated a video AP branch. This addition encourages the model to focus on and accurately describe verbs based on action classification results, thus refining the caption generation process and producing more precise and descriptive captions. Our experimental evaluations on two benchmark datasets demonstrate that ALSO-Net significantly improves the accuracy of visual word predictions and increases the richness of visual words in the generated captions, outperforming current state-of-the-art video captioning methods. Looking ahead, we propose to explore the integration of graph convolutional networks with ALSO-Net to further augment caption generation performance. Leveraging graph structures and contextual dependencies is anticipated to enhance caption quality and description expression.

## References

[1] Yang Bai, Junyan Wang, Yang Long, Bingzhang Hu, Yang Song, Maurice Pagnucco, and Yu Guan. 2021. Discriminative Latent Semantic Graph for Video Captioning. In *Proc. ACM Multimedia*, 3556–3564.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proc. Assoc. Comput. Linguist. Workshops*, 65–72.

[3] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 4724–4733.

[4] Jingwen Chen and Hongyang Chao. 2020. VideoTRM: Pre-training for Video Captioning Challenge 2020. In *Proc. ACM Multimedia*, 4605–4609.

[5] Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Hongyang Chao, and Tao Mei. 2023. Retrieval Augmented Convolutional Encoder-Decoder Networks for Video Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 1s (2023), 48:1–48:24.

[6] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. 2020. Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos. In *Proc. Eur. Conf. Comput. Vis,* 333–351.

[7] Shaoxiang Chen and Yu-Gang Jiang. 2021. Motion Guided Region Message Passing for Video Captioning. In *Proc. IEEE/CVF Int. Conf. Comput. Vis,* 1523–1532.

[8] Bo Dai, Sanja Fidler, and Dahua Lin. 2018. A Neural Compositional Paradigm for Image Captioning. In *Adv. Neural Inf. Process. Syst,* 656–666.

[9] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proc. Assoc. Comput. Linguist,* 2978–2988.

[10] Jincan Deng, Liang Li, Beichen Zhang, Shuhui Wang, Zhengjun Zha, and Qingming Huang. 2022. Syntax-Guided Hierarchical Attention Network for Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* 32, 2 (2022), 880–892.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proc. North Am. Chapter Assoc. Comput. Linguist,* 4171–4186.

[12] Shanshan Dong, Tian-Zi Niu, Xin Luo, Wu Liu, and Xinshun Xu. 2023. Semantic Embedding Guided Attention with Explicit Visual Feature Fusion for Video Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 2 (2023), 68:1–68:18.

[13] Zhengcong Fei. 2019. Fast Image Caption Generation with Position Alignment. arXiv:1912.06365. Retrieved from https://arxiv.org/abs/1912.06365

[14] Zhengcong Fei. 2021. Partially Non-Autoregressive Image Captioning. In *Proc. AAAI Conf. Artif. Intell*, 1309–1316.

[15] Lianli Gao, Yu Lei, Pengpeng Zeng, Jingkuan Song, Meng Wang, and Heng Tao Shen. 2022. Hierarchical Representation Network with Auxiliary Tasks for Video Captioning and Video Question Answering. *IEEE Trans. Image Process.* 31 (2022), 202–215.

[16] Lianli Gao, Xiangpeng Li, Jingkuan Song, and Heng Tao Shen. 2020. Hierarchical LSTMs with Adaptive Attention for Visual Captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 5 (2020), 1112–1131.

[17] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In *Proc. Conf. Empirical Methods Nat. Lang. Process,* 6111–6120.

[18] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2013. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition. In *Proc. IEEE/CVF Int. Conf. Comput. Vis,* 2712–2719.

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 3154–3160.

[20] Yiqing Huang, Jiansheng Chen, Wanli Ouyang, Weitao Wan, and Youze Xue. 2020. Image Captioning with End-to-End Attribute Detection and Subsequent Attributes Prediction. *IEEE Trans. Image Process.* 29 (2020), 4013–4026.

[21] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Zheng Wang, Xiao Wang, Junjun Jiang, and Chia-Wen Lin. 2021. Rain-Free and Residue Hand-in-Hand: A Progressive Coupled Network for Real-Time Image Deraining. *IEEE Trans. Image Process.* 30 (2021), 7404–7418.

[22] Tao Jin, Siyu Huang, Ming Chen, Yingming Li, and Zhongfei Zhang. 2020. SBAT: Video Captioning with Sparse Boundary-Aware Transformer, In *Proc. Int. Joint Conf. Artif. Intell*, 630–636.

[23] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950. Retrieved from https://arxiv.org/abs/1705.06950

[24] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. Int. Conf. Learn. Represent*. 1–15.

[25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *Proc. IEEE/CVF Int. Conf. Comput. Vis*, 706–715.

[26] Yun Lan, Ruimin Hu, Xin Xu, Dengshi Li, Chao Wang, and Xiaochen Wang. 2023. From Collective Attribute Association of Groups to Precise Attribute Association of Individuals. *IEEE Trans. Multimedia* 25 (2023), 1547–1554.

[27] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *Proc. Assoc. Comput. Linguist,* 2603–2614.

[28] Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. 2022a. Long Short-Term Relation Transformer with Global Gating for Video Captioning. *IEEE Trans. Image Process.* 31 (2022), 2726–2738.

[29] Linghui Li, Yongdong Zhang, Sheng Tang, Lingxi Xie, Xiaoyong Li, and Qi Tian. 2022b. Adaptive Spatial Location with Balanced Loss for Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* 32, 1 (2022), 17–30.

[30] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. Eur. Conf. Comput. Vis,* 740–755.

[32] Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. 2021. O2NA: An Object-Oriented Non-Autoregressive Approach for Controllable Video Captioning, In *Proc. Assoc. Comput. Linguist. Findings,* 281–292.

[33] Sheng Liu, Zhou Ren, and Junsong Yuan. 2021. SibNet: Sibling Convolutional Encoder for Video Captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 9 (2021), 3259–3272.

[34] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2023. Entity-Enhanced Adaptive Reconstruction Network for Weakly Supervised Referring Expression Grounding. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3 (2023), 3003–3018.

[35] Xiaoxiao Liu and Qingyang Xu. 2021. Adaptive Attention-Based High-Level Semantic Introduction for Image Caption. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 4 (2021), 128:1–128:22.

[36] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. 2020. Spatio-Temporal Graph for Video Captioning With Knowledge Distillation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 10867–10876.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. Assoc. Comput. Linguist,* 311–318.

[38] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. 2019. Memory-Attended Recurrent Network for Video Captioning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 8347–8356.

[39] Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2020. Learning to Recover from Multi-Modality Errors for Non-Autoregressive Neural Machine Translation. In *Proc. Assoc. Comput. Linguist,* 3059–3069.

[40] Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2021. Guiding Non-Autoregressive Neural Machine Translation Decoding with Reordering Information. In *Proc. AAAI Conf. Artif. Intell,* 13727–13735.

[41] Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D. Yoo. 2021. Semantic Grouping Network for Video Captioning. In *Proc. AAAI Conf. Artif. Intell,* 2514–2522.

[42] Bin Sheng, Ping Li, Riaz Ali, and C. L. Philip Chen. 2022. Improving Video Temporal Consistency via Broad Learning System. *IEEE Trans. Cybern.* 52, 7 (2022), 6662–6675.

[43] Yaya Shi, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2023. Learning Video-Text Aligned Representations for Video Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 2 (2023), 63:1–63:21.

[44] Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. 2020. Latent-Variable Non-Autoregressive Neural Machine Translation with Deterministic Inference Using a Delta Posterior. In *Proc. AAAI Conf. Artif. Intell,* 8846–8853.

[45] Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards Diverse Paragraph Captioning for Untrimmed Videos. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 11245–11254.

[46] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402. Retrieved from https://arxiv.org/abs/1212.0402

[47] Zhixin Sun, Shuqin Chen, and Luo Zhong. 2022. Visual-aware Attention Dual-stream Decoder for Video Captioning. In *Proc. IEEE Int. Conf. Multimedia Expo, 1–6.*

[48] Chunwei Tian, Menghua Zheng, Wangmeng Zuo, Shichao Zhang, Yanning Zhang, and Chia-Wen Lin. 2024. A Cross Transformer for Image Denoising. *Inf. Fusion* 102 (2024), 102043.

[49] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc. IEEE/CVF Int. Conf. Comput. Vis,* 4489–4497.

[50] Yunbin Tu, Liang Li, Li Su, Junping Du, Ke Lu, and Qingming Huang. 2023. Viewpoint-Adaptive Representation Disentanglement Network for Change Captioning. *IEEE Trans. Image Process.* 32 (2023), 2620–2635.

[51] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. 2023. Self-supervised Cross-view Representation Reconstruction for Change Captioning. In *Proc. IEEE/CVF Int. Conf. Comput. Vis,* 2793–2803.

[52] Yunbin Tu, Chang Zhou, Junjun Guo, Shengxiang Gao, and Zhengtao Yu. 2021. Enhancing the Alignment between Target Words and Corresponding Frames for Video Captioning. *Pattern Recognit.* 111 (2021), 107702.

[53] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 4566–4575.

[54] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018. Reconstruction Network for Video Captioning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 7622–7631.

[55] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *Proc. IEEE/CVF Int. Conf. Comput. Vis,* 4580–4590.

[56] Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019a. Non-Autoregressive Machine Translation with Auxiliary Regularization. In *Proc. AAAI Conf. Artif. Intell,* 5377–5384.

[57] Bofeng Wu, Guocheng Niu, Jun Yu, Xinyan Xiao, Jian Zhang, and Hua Wu. 2022. Towards Knowledge-Aware Video Captioning via Transitive Visual Relationship Detection. *IEEE Trans. Circuits Syst. Video Technol.* 32, 10 (2022), 6753–6765.

[58] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 203–212.

[59] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 5288–5296.

[60] Kashu Yamazaki, Khoa Vo, Quang Sang Truong, Bhiksha Raj, and Ngan Le. 2023. VLTinT: Visual-Linguistic Transformer-in-Transformer for Coherent Video Paragraph Captioning. In *Proc. AAAI Conf. Artif. Intell,* 3081–3090.

[61] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. 2020. STAT: Spatial-Temporal Attention Mechanism for Video Captioning. *IEEE Trans. Multimedia* 22, 1 (2020), 229–241.

[62] Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. 2021. Non-Autoregressive Coarse-to-Fine Video Captioning. In *Proc. AAAI Conf. Artif. Intell,* 3119–3127.

[63] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. Object Relational Graph with Teacher-Recommended Learning for Video Captioning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit,* 13275–13285.

[64] Qi Zheng, Chaoyue Wang, and Dacheng Tao. 2020. Syntax-Aware Action Targeting for Video Captioning. In *Proc. Int. Joint Conf. Artif. Intell,* 13093–13102.

[65] Xian Zhong, Zipeng Li, Shuqin Chen, Kui Jiang, Chen Chen, and Mang Ye. 2023. Refined Semantic Enhancement towards Frequency Diffusion for Video Captioning. In *Proc. AAAI Conf. Artif. Intell,* 3724–3732.

[66] Xian Zhong, Yi Zhang, Shuqin Chen, Zhixin Sun, Huantao Zheng, and Kui Jiang. 2022. Dual-scale Alignment-Based Transformer on Linguistic Skeleton Tags for Non-Autoregressive Video Captioning. In *Proc. IEEE Int. Conf. Multimedia Expo.*

[67] Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding Knowledge Distillation in Non-autoregressive Machine Translation. In *Proc. Int. Conf. Learn.* Represent.

[68] Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *Proc. AAAI Conf. Artif. Intell,* 7590–7598.

[69] Yu Zhou, Zhihua Chen, Ping Li, Haitao Song, C. L. Philip Chen, and Bin Sheng. 2023. FSAD-Net: Feedback Spatial Attention Dehazing Network. *IEEE Trans. Neural Networks Learn. Syst.* 34, 10 (2023), 7719–7733.

[70] Lei Zhu, Xize Wu, Jingjing Li, Zheng Zhang, Weili Guan, and Heng Tao Shen. 2023. Work Together: Correlation-Identity Reconstruction Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Trans. Knowl. Data Eng.* 35, 9 (2023), 8838–8851.

Received 20 August 2023; revised 6 May 2024; accepted 13 July 2024