Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.046418

ARTICLE

A Video Captioning Method by Semantic Topic-Guided Generation

Ou Ye, Xinli Wei, Zhenhua Yu*, Yan Fu and Ying Yang

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, 710054, China

*Corresponding Author: Zhenhua Yu. Email: zhenhua_yu@163.com

Received: 30 September 2023 Accepted: 28 November 2023 Published: 30 January 2024

ABSTRACT

In the video captioning methods based on an encoder-decoder, limited visual features are extracted by an encoder, and a natural sentence of the video content is generated using a decoder. However, this kind of method is dependent on a single video input source and few visual labels, and there is a problem with semantic alignment between video contents and generated natural sentences, which are not suitable for accurately comprehending and describing the video contents. To address this issue, this paper proposes a video captioning method by semantic topic-guided generation. First, a 3D convolutional neural network is utilized to extract the spatiotemporal features of videos during the encoding. Then, the semantic topics of video data are extracted using the visual labels retrieved from similar video data. In the decoding, a decoder is constructed by combining a novel Enhance-TopK sampling algorithm with a Generative Pre-trained Transformer-2 deep neural network, which decreases the influence of "deviation" in the semantic mapping process between videos and texts by jointly decoding a baseline and semantic topics of video contents. During this process, the designed Enhance-TopK sampling algorithm can alleviate a long-tail problem by dynamically adjusting the probability distribution of the predicted words. Finally, the experiments are conducted on two publicly used Microsoft Research Video Description and Microsoft Research-Video to Text datasets. The experimental results demonstrate that the proposed method outperforms several state-of-art approaches. Specifically, the performance indicators Bilingual Evaluation Understudy, Metric for Evaluation of Translation with Explicit Ordering, Recall Oriented Understudy for Gisting Evaluation-longest common subsequence, and Consensus-based Image Description Evaluation of the proposed method are improved by 1.2%, 0.1%, 0.3%, and 2.4% on the Microsoft Research Video Description dataset, and 0.1%, 1.0%, 0.1%, and 2.8% on the Microsoft Research-Video to Text dataset, respectively, compared with the existing video captioning methods. As a result, the proposed method can generate video captioning that is more closely aligned with human natural language expression habits.

KEYWORDS

Video captioning; encoder-decoder; semantic topic; jointly decoding; Enhance-TopK sampling

1 Introduction

At present, videos have become an essential carrier of information dissemination and an important source of daily human life, learning, and knowledge acquisition, such as object detection studies, video surveillance technology popularization, and film entertainment [1–4]. The studies of video captioning have become a hotspot in computer vision and cross-modal content cognition [5,6]. Their purpose is to enhance the intelligent understanding and analysis of video data by using natural language to describe and interpret the content of videos, aiming to achieve structured summarization and re-expression of visual content. In recent years, video captioning studies have gained attention for their applications in navigation assistance, human-computer interaction, automatic interpretation, and video monitoring [7].

Since 2010, there have been two primary video captioning methods: template-based video captioning [8] and retrieval-based video captioning [9]. The template-based video captioning methods require many manually designed annotations, resulting in a single syntactic structure and limited sentence diversity; in contrast, the retrieval-based video captioning methods are easily limited by the retrieval samples, which makes it difficult to generate accurate natural sentences. Currently, with the in-depth studies of deep learning in image processing [10] and machine translation [11], deep network models with encoder-decoder have been applied to generate video captioning [12,13]. This kind of model regards to video captioning as a process of "translation". In the encoding phase, the visual contents of video data are encoded into feature vectors using an encoder; in the decoding phase, a decoder is utilized to map feature vectors to generate semantic aligned natural sentences. In general, the video captioning methods based on encoder-decoder can achieve caption generations that are more in line with human language habits through cross-media semantic alignment. However, since the rich and diverse presentation forms of video content, relying solely on the visual features extracted by an encoder can easily ignore some details and be affected by noise and changes in appearance features, which are not conducive to accurately comprehending and describing the video contents. To address this issue, an attention mechanism is introduced into a video captioning method based on the encoder-decoder in [14]. This method improves the quality of generated sentences by assigning weights to optimize and select the temporal features. At present, although attention mechanisms can improve the generated accuracy of video captioning, a semantic alignment issue due to a single input source limits accurate video captioning generation.

Considering that the semantic topic expression is not only limited to text information but also can convey visual information, which can effectively converge cross-modal semantic elements in the visual and linguistic domains and alleviate the limitations of limited visual features. Furthermore, incorporating the semantic topics of video data can also effectively guide the caption generation of the video contents. Therefore, a video captioning method based on semantic topic-guided generation is proposed, which guides the generation of video captioning using external semantic topic information. This method utilizes the Convolutional 3D (C3D) model to extract advanced spatiotemporal features of video data, then generates video captioning by constructing a decoder Enhance-TopK Generative Pre-rained Transformer-2 (EGPT-2) based on the extracted semantic topics of the video data. Taking Fig. 1 as an example, when input video data of "a woman rides a horse through the forest", even though the scene "forest" and the object action "through" are not easy to obtain in the encoding and decoding phases, they are still possible to utilize a Latent Dirichlet Allocation model to construct a semantic topic of the video data to guide and predict the video content of "girl riding through the forest" by retrieving "woman", "horse", "riding" and other keywords contained in a video with similar semantics. In this process, to alleviate the long tail problem existing in the decoding phase and make the generation of video captioning more accurate, a novel sampling algorithm named Enhance-TopK is designed, which calculates the probability distribution of the topic correlation coefficient to affect the probability distribution of the predictive words, further ensuring the accurate generation of the video captioning. As shown in the decoding phase of Fig. 1, the generated video captioning is "a girl is riding a horse through the forest" instead of passing through "fields" or "wilderness".

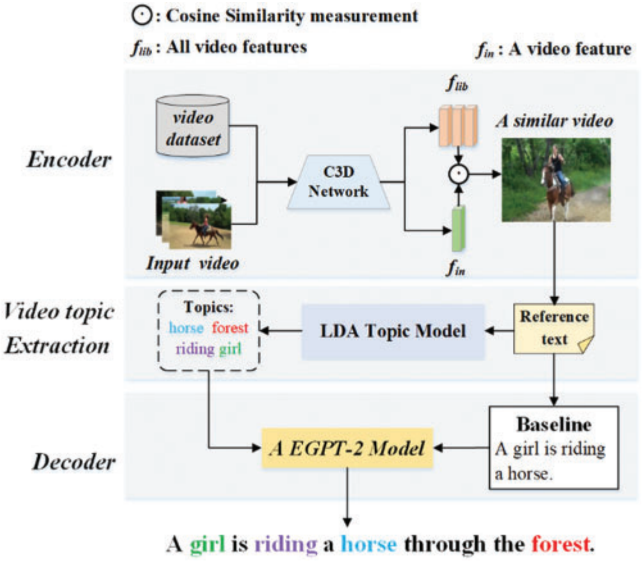

⊙: Cosine Similarity measurement
f_{lib}: All video features
f_{in}: A video feature
Encoder
video dataset
Input video
C3D Network
f_{lib}
f_{in}
A similar video
Video topic Extraction
Topics: horse forest riding girl
LDA Topic Model
Reference text
Decoder
A EGPT-2 Model
Baseline A girl is riding a horse.
A girl is riding a horse through the forest.

Figure 1: The general idea of a video captioning method based on semantic topic-guided generation

This paper's contributions are summarized as follows: (1) an innovative idea of utilizing video semantic topics as a "bridge" to guide the video captioning generation is proposed, which can improve the semantic alignment between generated natural sentences and video content, thereby reducing the "deviation" effect in the semantic mapping process between video and text; (2) in the decoding phase, a baseline caption of a video and its semantic topic are jointly decoded through an EGPT-2 deep network model. In this process, a sampling algorithm named Enhance-TopK is designed to enhance the topic impact on the next prediction word, thereby adjusting the probability distribution of the prediction words. The purpose is to ensure that the generated video captioning is consistent with the video topic and keeps the natural sentences smooth; (3) the proposed method is verified on the Microsoft Research Video Description (MSVD) and Microsoft Research - Video to Text (MSR-VTT) datasets with a significant difference in size to evaluate the different effects of semantic topics on the generation of video captioning, which provides help for in-depth studies on how the topic information of video content can guide the video captioning generation.

This paper is divided into several sections. First, Section 2 presents a brief review of related video captioning. Subsequently, Section 3 introduces a video captioning method based on semantic topic-guided generation. The performance of this method is then verified through experiments in Section 4. Finally, we conclude and provide the future study in Section 5.

2 Related Works

In the early stages of video caption studies, video captioning methods mainly consisted of template-based and retrieval-based video captioning methods. Template-based video captioning methods involve two phases: syntax generation and semantic constraints. Syntax generation refers to detecting attributes, concepts, and relations of objects in videos using designed manual features; semantic constraint refers to pre-setting fixed sentence templates or syntax rules, and then the detected objects are combined into the final natural language descriptions [15]. For example, a study in [16] first detects objects in images, and predicts the attributes of objects and prepositional relations between objects; then a Conditional Random Fields (CRF) model is constructed to predict labels in the form of <object–attribute–preposition>triplet and generate natural sentences based on linguistic templates. Yang et al. [17] first detected the features and scenes of objects in the images, then predict the verbs, scenes, and prepositions that makeup sentences. Afterward, a hidden Markov model is used to infer the optimal quaternion of <noun–verb–scene–preposition> to form the sentence. Finally, a description sentence of the video content is generated. However, this method relies more on attribute detections and template settings, resulting in poor performance of the final generated video captioning. The retrieval-based video captioning methods are to transform the video captioning generation task into a video retrieval task. Specifically, video captioning is generated by retrieving sentences with similar semantics to the target video from manually constructed sentence sets. For instance, the global similarities of the query images are calculated on a constructed large image dataset, and the retrieved sentences are used as the natural sentences of video contents [18]. The work in [19] utilized the potential semantic information of the images to generate a new video captioning through a triple group <visual clues–corpus statistics–available descriptions>. In general, although retrieval-based video captioning methods can retrieve natural sentences like the semantics of the video contents, due to the limited scale of retrieval samples, they are challenging to generate video captioning with a high fit to the video contents. Under some extreme conditions, the generated video captioning may have nothing to do with the video contents themselves.

With the in-depth studies of video captioning approaches, the encoder-decoder models combining convolutional neural networks with recurrent neural networks [20] have been widely used to generate video captioning. For example, Venugopalan et al. [21] utilized an encoder of the AlexNet model to extract video features. Then, these features are fed into a short-term long memory (LSTM) network model to generate the natural sentences. However, this method ignores the influence of the temporal features on the semantic representation of video data. To address this issue, the study in [22] constructed a long-short-term graph (LSTG) to capture the short-term spatial semantic relationship and long-term conversion dependency relationship between several visual objects, then fully exploring the spatiotemporal relationships between visual objects, and implements the inference of object relationship on LSTG through the global gated module, further improving the quality of generated video captioning. Zheng et al. [23] suggested that existing methods prioritize the accuracy of object category prediction in generated video captioning. However, it often overlooks object interactions. Hence, a Syntax-Aware Action Targeting model is designed to detect semantic objects and dynamic information to learn the actions in videos to improve the accuracy of action prediction in the generated video captioning. The work in [24] proposed a semantic-based video keyframe method that extracts the initial video keyframe using a convolutional neural network model and the feature windows. Then the keyframes are automatically marked by the image caption network. Finally, a pre-interactive LSTM network model is proposed to generate video captioning, which can fully extract the semantic feature of a video using the video keyframes to improve the generation of video captioning. However, the quality of the generated video captioning depends on the accuracy of keyframe annotations. Liu et al. [25] proposed an Unpaired Video Captioning with Visual Injection system to address the issue that sufficient paired data is unavailable for many targeted languages. However, the semantic information of video captioning is still limited. The study in [26] presented a novel global-local encoder to obtain video captioning across frames by generating a rich semantic vocabulary. However, this method only enriches semantic vocabulary by encoding different visual features, and the source of semantic information is relatively single. Recent studies of video captioning utilized an encoder-decoder to extract the visual features from the video data and then decode these features with a decoder to produce human-like sentences that match the video content semantically. However, video content is diverse and complex, which can result in an encoder losing some critical visual information during the encoding process. This loss of part visual information can make it difficult to accurately convey the video content during the decoding phase.

There have been related studies to address this issue. For example, the work in [27] proposed a dual-stream learning method with multiple instances and labels, which minimizes the semantic gap between the original video and the generated captions through a dual-learning mechanism of captioning generation and video reconstruction. Furthermore, a Textual-Temporal Attention Model was proposed in [28]. This model introduces pre-detected visual labels from the videos and selects the most relevant visual labels based on the contextual background. Subsequently, the temporal attention mechanism is used to enhance semantic consistency between visual content and generated sentences. Nevertheless, whether it is simple to label extraction or enhance the semantic alignment between video and text by stacking a multi-layer attention mechanism, it may introduce irrelevant semantic noise, which may further amplify the impact of "semantic deviation" on the generation of video captioning.

3 The Proposed Method

To alleviate the problem of "semantic bias" caused by limited visual information, a video captioning method based on semantic topic-guided generation is proposed, and the overall idea is shown in Fig. 2. This approach comprises an encoder, a semantic topic extraction module, and a decoder. In the encoding phase, the encoder in this methodology utilizes a C3D deep network model to capture the spatiotemporal features. At the stage of semantic topic extraction of videos, the retrieval of video reference sentences and extraction of semantic topics are mainly implemented. Due to the visual consistency between the video reference sentences and the video contents, the topic words with the reference text R related to the contents of an input video are extracted in this phase. Then, a probability set of topic words is obtained, denoted as $P_{topic} = \{P_{w1}, P_{w2}, \ldots, P_{wn}\}$, where p_{wi} denotes a topic probability of the i^{th} topic word. Finally, we select the top n (n = 6 is set in this paper) topic words $T_v = \{t_1, t_2, \ldots, t_n\}$ in the probability ranking P_{topic} as a semantic topic of an input video. In the decoding phase, aiming to strengthen the semantic alignment between the video contents and the generated natural sentences under the guidance of the semantic topics, an EGPT-2 decoder is constructed to jointly decode a baseline caption C_b and the semantic topics of a given input video. In this process, aiming to improve the accuracy of the predicted words, an Enhance-TopK sampling algorithm is designed to calculate the probability distribution of the topic-related coefficients λ affecting the predicted words, which is used to enhance the influence of the video semantic topics on the prediction words in the next moment.

3.1 *An Encoder Construction of C3D Deep Neural Network*

Since video data have both temporality and spatiality, the C3D deep network model can extract both appearance and motion features from videos simultaneously, a pre-trained C3D deep network model with eight convolution layers, five pooling layers, two fully connected layers, and a softmax output layer was used in this paper. Specifically, we take any i^{th} video clip v_i in a video dataset $V = \{v_1, v_2, \ldots, v_N\}$ as the input video to inject into a C3D deep network model and extract the spatiotemporal features of a given video. Suppose a video clip v_i has l video frames. Each video frame's size is $w \times h$ and has the number of c channels. We designate the values of w and h are all 112 and $c = 3$. In addition, we refer to the parameter settings in [15] to set the convolution kernel size $3 \times 3 \times 3$ and the step size 1 × 1 × 1. To preserve the spatiotemporal features of the videos, we configure the size of the pooling kernel in the initial pooling layer as 1 × 2 × 2 with the step size as 1 × 2 × 2, and the remaining 3D pooling kernel size with step size 2 × 2 × 2 for implementing the convolution operations, and the maximum pooling processes. In the convolution operation, the C3D deep network model performs the convolution operation by stacking the 3D convolution kernel and the cube formed by multiple consecutive frames. Therefore, the feature map generated by the current convolution layer can be used to capture spatiotemporal features of multiple consecutive frames in the upper layer of the network. Formally, an eigenvalue in the location (x, y, z) on the j^{th} feature map of the i^{th} layer in the C3D deep network model can be acquired using Eq. (1):

$$f_{ij}^{xyz} = \mathrm{Relu}\left[\sum_{s,t,r}^{S_i-1,T_i-1,R_i-1} w_{ijm}^{str}\ f_{(i-1)}^{(x+s)(y+t)(z+r)} + b_{ij}\right] \tag{1}$$

where Relu (•) denotes an activation function of the Rectified Linear Unit, m denotes the number of index relations linking a group of feature maps in the upper layer to the feature map in the current layer; S_i, T_i and R_i are the height, width, and size in the temporal dimension of the 3D convolution kernels, respectively; w_{ijm}^{str} denotes an eigenvalue of the point (s, t, r) linked to the m^{th} feature map in the upper layer, and b_{ij} denotes the bias of the current feature map. After eight convolutional layers and five maximum pooling operations, two fully connected layers are utilized to extract a spatiotemporal feature vector $f_{v_{in}}$ of size [1, 4096] to represent the global features of the given input videos. Moreover, the C3D deep network model employs a dropout function with a dropout probability of 0.5 and an activation function of the Rectified Linear Unit to prevent overfitting.

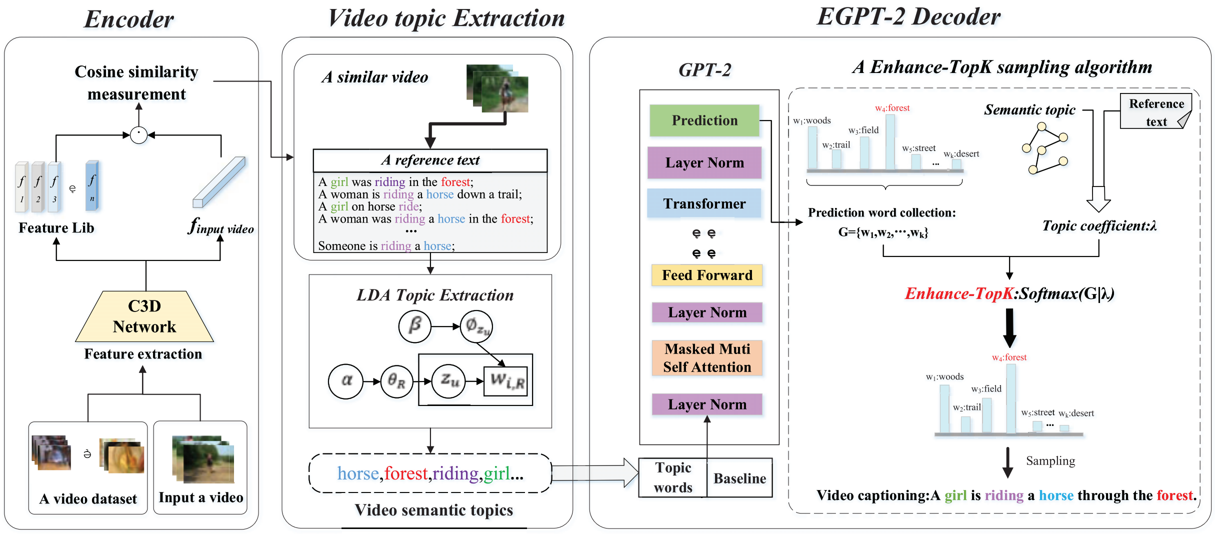


Figure 2: The video captioning method based on semantic topic-guided generation

3.2 Semantic Topic Extraction of Videos

In this paper, the reference sentences of videos refer to the natural sentences related to this video content, which are derived from the description sentences of other videos like this video content. Since the reference sentences of the input videos contain valuable semantic information that aids in describing video topics, the semantic topic distribution of the video data can be extracted according to the reference sentences of the videos.

First of all, the matching pairs of "video clip–reference sentences" in a video dataset are defined as $v_i = \left\{v_i\colon y_i^1, y_i^2, \ldots, y_i^j\right\}$,$i \in [1, N]$, $j \in [1, J]$, where N denotes the total number of videos in a video dataset, and y_i^j denotes the j^{th} reference sentence of the i^{th} video data in a video dataset V. Then, the feature vectors of the input video are used to retrieve the video data with similar semantics, and the several reference sentences of this video are spliced to form a reference text R, which is expressed as $R = \left\{y_i^1, y_i^2, \ldots, y_i^j\right\}$. Subsequently, a reference text R is randomly selected as a baseline caption of the video v_i. During the test phase of the topic extraction of video content, since the input video v_{in} does not have to correspond to the reference sentences, a video v_{res} that is most similar to v_{in} a video dataset V is retrieved by using the similarity measurement between videos and learning a semantic topic of the given input video v_{in} with a reference text R of the video v_{res}.

Before retrieving similar video data, each video in the video datasets can be represented as a high-dimensional feature vector by using a C3D deep network model. Then, similar video data are retrieved by calculating the similarities between the given input video v_{in} and all other videos in a video dataset. According to [29], similar video retrieval methods based on Euclidean distance may be significantly affected by the size of the visual representation dimension, while cosine similarity can still maintain the accuracy of similarity calculation when the dimension is high. Moreover, Wang et al. [30] calculated the cosine distance between the global feature vectors of video data and the corresponding features of text captioning by weighting to measure the similarity of different videos. Therefore, the similarities between an input video and all other videos in the video set are calculated through the cosine distance in this paper, as shown in Eq. (2):

$$\mathrm{sim}\left(f_{v_{in}}, f_{v_i}\right) = \frac{f_{v_{in}} \cdot f_{v_i}}{\left|f_{v_{in}}\right| \times \left|f_{v_i}\right|} \tag{2}$$

where $f_{v_{in}}$ denotes a feature vector of an input video; f_{v_i} denotes a feature vector of any one video data in the dataset, and sim $(\cdot)$ denotes a function of cosine similarity measurement between the input video v_{in} and any i^{th} video data v_i. Here, we select a video v_{res} with the highest cosine similarity as the similar video of v_{in}.

Through the above phases, similar video data v_{res} with the given input video v_{in} can be retrieved, and a reference text R can be obtained. On this basis, this reference text can be utilized to extract a semantic topic of the given input video. Since the topic model based on the Latent Dirichlet Allocation [31] can obtain the topics of documents in the form of a probability distribution, and it is used in the studies of text clustering [32] and automatic summarization [33], the semantic topics of videos are extracted using the topic model based on Latent Dirichlet Allocation. The topic extraction process of video semantics can be regarded as a generation process of "videos→texts→topics→words". Here, it should be noted that since the reference text R comes from the baselines of similar video data, the above words extracted from R using a topic model can serve as visual labels to accurately describe the semantic topics of video data.

Specifically, a reference text R of a video is extracted first, and then a topic probability based on this reference text is calculated according to Eq. (3):

$$Z_{R,u} \sim \mathrm{Multinomial}\left(\theta_R\right)\ u = 1, \ldots, U \tag{3}$$

where $Z_{R,u}$ denotes the u^{th} topic probability of a reference text R, and obeys the multinomial distribution with parameters. Here, θ_R denotes a topic probability distribution of reference text R, and obeys the Dirichlet distribution $\theta_R \sim \mathrm{Dirichlet}\left(\alpha\right)$, where α denotes a priori parameter of the Dirichlet distribution, $\alpha = 1/U$ is set in this paper. Moreover, U denotes the number of topics. Since all videos in a video dataset V can be divided into several categories, for example, a dataset named MSR-VTT is divided into 20 categories, $U = 20$ which is set for this video dataset.

On this basis, the probability distribution of topic words under the u^{th} topic can be obtained through Eq. (4):

$$w^i_{R,u} \sim \text{Multinomial}\left(\Phi_{k=Z_{R,u}}\right) \tag{4}$$

where $w^i_{R,u}$ denotes the i^{th} topic word of reference text R under the u^{th} topic, and obeys multiple distributions with the parameters. Here, $\Phi_{k=Z_{R,u}}$ denotes the only generated topic words under the u^{th} topic, and obeys the Dirichlet distribution $\Phi_{k=Z_{R,u}} \sim \text{Dirichlet}\,(\beta)$, β denotes the priori parameter of the Dirichlet distribution, which is set to 0.01.

When the topic words under all topics in the reference text R have been extracted, the semantic topics of the given input video v_{in} can be obtained through the calculation of Eq. (5):

$$P\left(w^i_{R,u}\left|z_{R,u}\right.\right) = \left(z_{R,u}\left|\left(\text{sim}\left(f_{v_{in}},f_{v_i}\right) \to R\right),\theta_R\right.\right) \cdot P\left(\Phi_{k=z_{R,u}}\right) \tag{5}$$

where $p\left(w^i_{R,u}\left|z_{R,u}\right.\right)$ denotes the probability of the i^{th} topic word w_i in the u^{th} topic under reference text R. On this basis, we can obtain a co-occurrence frequency matrix $\mathbf{E}$ of topic words by using Eq. (6).

$$\mathbf{E} = \begin{bmatrix} p\left(w^1_{R,1}\left|z_{R,1}\right.\right) & p\left(w^2_{R,1}\left|z_{R,1}\right.\right) & \cdots & p\left(w^n_{R,1}\left|z_{R,1}\right.\right) \\ p\left(w^1_{R,2}\left|z_{R,2}\right.\right) & p\left(w^2_{R,2}\left|z_{R,2}\right.\right) & \cdots & p\left(w^n_{R,2}\left|z_{R,2}\right.\right) \\ \vdots & \vdots & \ddots & \vdots \\ p\left(w^1_{R,U}\left|z_{R,U}\right.\right) & p\left(w^2_{R,U}\left|z_{R,U}\right.\right) & \cdots & p\left(w^n_{R,U}\left|z_{R,U}\right.\right) \end{bmatrix} \tag{6}$$

For any one row or column in the matrix $\mathbf{E}$, the element values are sorted in reverse order. Hence, we can obtain the top n topic words as a semantic topic $T_v = \{w_1, w_2, \ldots, w_n\}$ of the input video v_{in}.

3.3 The Construction of an EGPT-2 Decoder

Inspired by the open domain question and answer tasks [34], video captioning is not generated directly based on visual features, but a GPT-2 language model [35] is introduced to generate video captioning. Since the GPT-2 language model is a language model that has been pre-trained by a large number of unsupervised data, it can learn rich semantic information from external text data, and reduce the dependence on labeled data. In the decoding phase, aiming to obtain external semantic information from the video semantic topics to enrich the semantics of video captioning, the visual semantic topics based on the GPT-2 language model and a baseline caption of video content are jointly decoded. In addition, to alleviate the long-tail problem that appears during the decoding phase, a novel Enhance-TopK sampling algorithm is designed to combine with the GPT-2 language model to construct an EGPT-2 decoder. This decoder can adjust the probability distribution of the predictive words by combining the topic correlation coefficient λ, which can reduce the impact of predicted repeating irrelevant words on the accuracy of the generated video captioning.

As shown in Fig. 3, the EGPT-2 decoder contains a stacked 12-layer Transformer module. Specifically, the input data f_{in} of the EGPT-2 decoder is defined as the form of a <baseline captions–topics>. First, a Byte Pair Encoding (BPE) model is utilized to encode each word f_{in} as a token corresponding to a vocabulary, which effectively merges the subword information and is conducive to processing words outside the vocabulary. In addition, the learnable position coding way and the pre-training weight of the EGPT-2 network model are adopted to initialize the word embedding of each token f_{in} to obtain the input matrix $\mathbf{F_{in}}$. Finally, the matrix $\mathbf{F_{in}}$ is input into a Transformer module stacked inside the EGPT-2 decoder, which captures the context dependency between a baseline caption and topic semantics of videos through a multi-head self-attention mechanism. After the Transformer module at this layer finishes processing, the results are input into the Transformer module at the lower layer to continue embedding until the Transformer module at the last layer outputs a context feature vector f_t with the topic information. Here f_t has a size of [1, 768], which can be obtained by using Eqs. (7) and (8):

$$\mathbf{head}_{\mathbf{i}} = \mathrm{SelfAttention}\left(\mathbf{F}_{\mathbf{in}} \cdot \boldsymbol{\mu}_{\mathbf{Q}}, \mathbf{F}_{\mathbf{in}} \cdot \boldsymbol{\mu}_{\mathbf{K}}, \mathbf{F}_{\mathbf{in}} \cdot \boldsymbol{\mu}_{\mathbf{V}}\right) \tag{7}$$

$$f_t = \mathrm{Transformer}_j\left(\mathrm{concat}\left(\mathbf{head}_{\mathbf{1}}, \ldots, \mathbf{head}_{\mathbf{n}}\right) \cdot \boldsymbol{\mu}_{\mathbf{s}}\right) \tag{8}$$

where $\mathbf{head}_i$ denotes the semantic features learned by the self-attention mechanism in the i^{th} feature subspace, and $i \in [1, m]$, selfAttention$(\cdot)$ denotes a self-attention function. $\mathbf{F}_{\mathbf{in}}$ denotes a word embedding matrix of all tokens in the input data, $\boldsymbol{\mu}_{\mathbf{Q}}$, $\boldsymbol{\mu}_{\mathbf{K}}$ and $\boldsymbol{\mu}_{\mathbf{V}}$ are learnable parameters, the "$\cdot$" operator denotes the dot product operation of matrixes. Transformer$(\cdot)$ denotes an output of the j^{th} layer in a Transformer module ($j = 12$, $m = 8$ are set in this paper), concat$(\cdot)$ denotes the combination of semantic features learned by the self-attention mechanism in different subspaces, and $\boldsymbol{\mu}_{\mathbf{s}}$ is a transformation matrix, which is used to maintain the size invariance of semantic features.

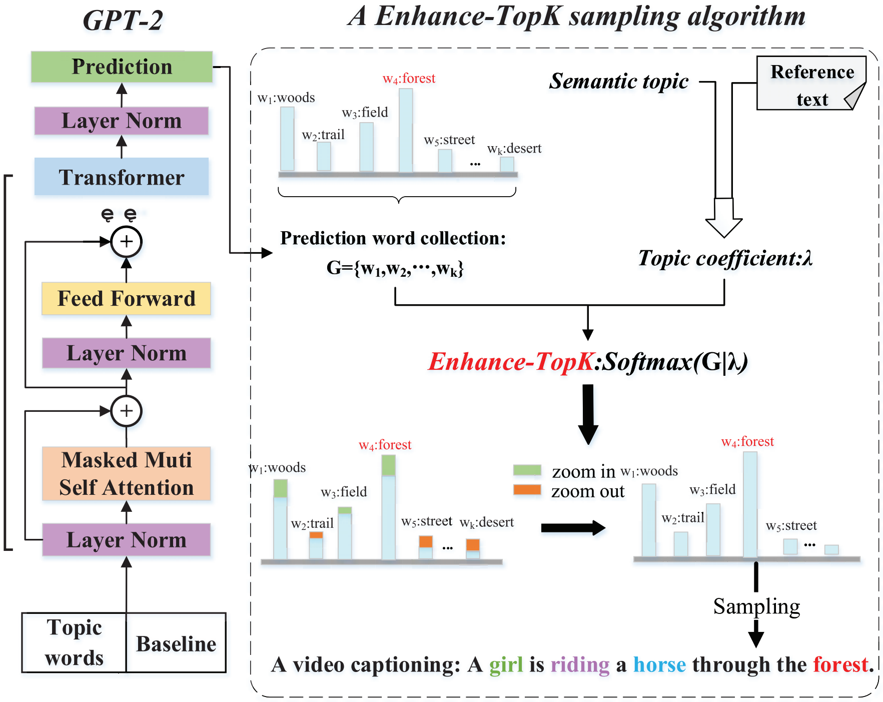


Figure 3: The designed process of an EGPT-2 decoder

Subsequently, the feature vector f_t is multiplied with the word embedding matrix $\mathbf{D}_{\mathbf{emb}}$ of vocabulary D, and the result is normalized to obtain an initial probabilities set of predicted words $G_p = \left\{p_{w_1}, \ldots p_{w_i}, \ldots, p_{w_n}\right\}$, where p_{w_i} denotes the co-occurrence probability of the i^{th} prediction word in the vocabulary D; p_{w_n}denotes the co-occurrence probability of the n^{th} prediction word in the vocabulary

D. The calculation is shown in Eq. (9):

$$p_{w_i} = softmax\,(f_t \cdot \mathbf{D}_{\mathbf{emb}}) \tag{9}$$

where $softmax\,(\cdot)$ denotes an activation function and the size of $\mathbf{D}_{\mathbf{emb}}$ is [768, 50257]. When using a greedy sampling algorithm for word predictions, a predicted word x_t at moment t can be obtained by using Eq. (10):

$$x_t = \max\left(p_{w_1}, \ldots, p_{w_n}\right) \tag{10}$$

According to Eq. (10), the greedy sampling algorithm selects the word with the highest probability for prediction at each time step. However, this method often leads to repetitive or meaningless video descriptions. When using pure random sampling, the length of the predicted word set G_p is 50257, which results in a long-tail problem that can cause generated sentences to be illogical and difficult to read. To alleviate this issue, references [36] and [37] utilized a decoding method based on beam search, which can improve the accuracy of the generated natural sentences. However, such methods are prone to fall into local optimization that leads to generating a more rigid, incoherent, or repetitive loop of the natural sentence. To suppress the long-tail effect and make the generated video captioning satisfy the diversity of language expressions while maintaining the smoothness of the natural sentences, a novel Enhance-TopK sampling algorithm is constructed in this paper, as shown in Fig. 4. This sampling algorithm uses the feature vectors of video data to calculate a topic correlation coefficient, which is utilized to adjust the probability distribution of predicted words to generate words related to the semantic topic of video data.

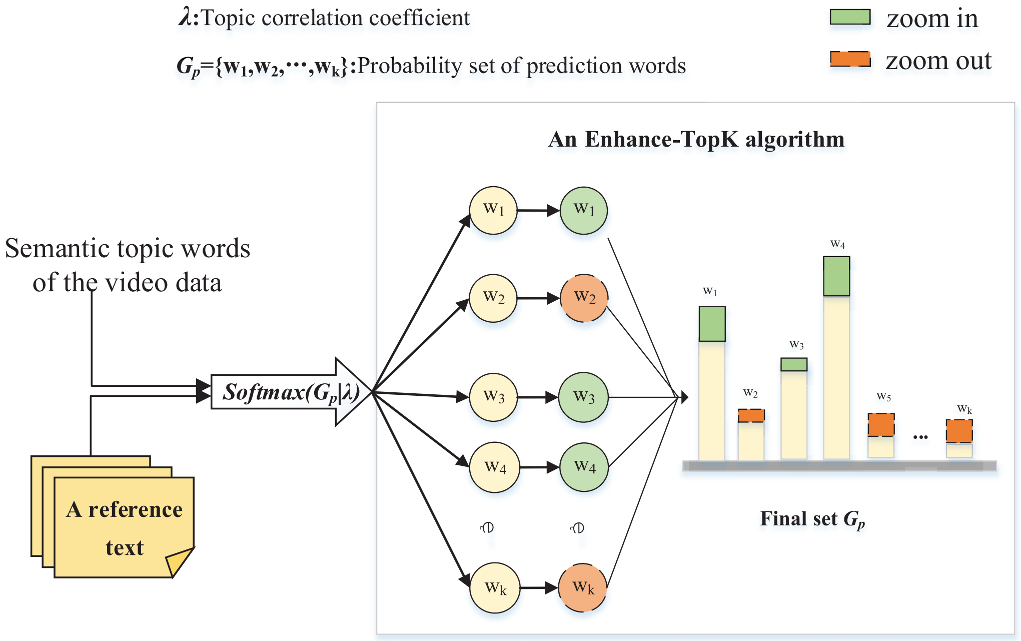

λ:Topic correlation coefficient
G_p={w1,w2,⋯,wk}:Probability set of prediction words
🟩 zoom in
🟧 zoom out
An Enhance-TopK algorithm
Semantic topic words of the video data
Softmax(G_p|λ)
A reference text
w1
w2
w3
w4
wk
w1
w2
w3
w4
w5
wk
...
Final set G_p

Figure 4: The idea of an Enhance-TopK sampling algorithm

Compared to existing Top-K sampling algorithms, the Enhance-TopK sampling algorithm considers the influence of the topics on the predicted word probability distribution. The “zoom in” in Fig. 4 shows the words that are greatly affected by the topic in G_p by using a topic correlation coefficient λ. For instance, the occurrence probabilities of words "w_1", "w_3" and "w_4" in G_p are further influenced by the topics under the reinforcement of λ. Moreover, "zoom-out" shows that the occurrence probability of words in a probability set of predicted words G_p will be further reduced, which is insensitive to the semantic topics at this moment, such as "w_2" and "w_k".

Specifically, an Enhance-TopK sampling algorithm first splits the topic words of an input video v_{in} to form the semantic topic T_v, as shown in Eq (11):

$$T_v = (t_1 \oplus t_2 \oplus, \ldots, \oplus t_n) \tag{11}$$

where t_i denotes the i^{th} topic word of a video; "⊕" denotes an operation of concatenating different topic words. Then, a pre-trained Sentence-Transformer deep network model [38] was used to implement word embedding calculation for a semantic topic T_v and a reference sentence y_i, as shown in Eq. (12):

$$\mathbf{f}_{y_i} = \text{sentenceTransformer}\left(y_i^w \left|\mathbf{head}_\mathbf{j}, \theta\right.\right) \tag{12}$$

where $\mathbf{f}_{y_i}$ denotes a feature vector of the reference sentence y_i; the sentenceTransformer $(\cdot)$ denotes the function of the Sentence-Transformer deep network model; y_i^w denotes the word embedding of the w^{th} word in the reference sentence y_i; $\mathbf{head}_\mathbf{j}$ denotes a semantic feature output in the j^{th} subspace, and θ denotes the super-parameters of a Sentence-Transformer deep network model. Similarly, a feature vector $\boldsymbol{f}_{T_V}$ of the semantic topic T_v can be obtained.

On this basis, a topic correlation coefficient λ of the input video v_{in} can be calculated by using the semantic topic T_v and the reference sentence y_i, as shown in Eq. (13):

$$\lambda = \frac{\sum_{i=1}^{m} \frac{f_{T_V} \cdot f_{y_i}}{\|f_{T_V}\| \cdot \|f_{y_i}\|}}{m} \tag{13}$$

where m is the number of sentences in a reference text R.

Then, the first Top-K words (we set $K = 30$ in this paper) in the prediction vocabulary D are selected in reverse order of probability as an initial sampling set $G_w = \left\{p_{w_1}, p_{w_2}, \ldots, p_{w_K}\right\}$ at the moment t. To enhance the influence of the semantic topics on word prediction, we combine the topic correlation coefficients λ to renormalize the sampling set G_w, as shown in Eq. (14):

$$\text{softmax}\left(G_w|\lambda\right) = \frac{exp\left(\frac{p_{w_i}}{\lambda}\right)}{\sum_{K=1}^{K} exp\left(\frac{p_{w_K}}{\lambda}\right)} \tag{14}$$

where K is the number of predicted words in the sampling set G_w. In addition, the probabilities of predicted words that are outside the sampling set G_w are set to 0. On this basis, a prediction word x_t^* at moment t can be obtained by using Eq. (15):

$$x_t^* = \text{softmax}\left(G_w\,|\lambda\right) \tag{15}$$

Finally, a prediction word at each moment is connected until the next prediction word is marked with <endoftext> to complete the generation of video captioning.

3.4 Training

In the encoding phase, a pre-trained C3D deep network model is used to extract the spatiotemporal features of the videos. The input size of the C3D deep network model is 3 × 16 × 112 × 112. In the semantic topic extraction of videos phase, the reference sentences of each video in the training set are spliced to form a video reference text R. After performing word segmentation and removing stop words, the semantic topics are extracted from videos using Latent Dirichlet Allocation. In the decoding phase, an EGPT-2 deep network model is constructed to jointly decode the baseline captions and semantic topics of videos, and predict the words to generate video captioning with the help of semantic topics. In the training phase, the input f_{in} of an EGPT-2 deep network model is a <baseline captions–topics–reference texts> triple. Here, this triple is abbreviated as$f_{in} = \{C_b, T_v, R\}$, and the generation probability P_θ of any video captioning y can be calculated by using Eq. (16). It should be noted that the form of input data is only $<$ baseline captions–topics $>$ in the test phase.

$$P_\theta(y) = \prod_{t=1}^{T} P_\theta\left(x_t^* \left| S_{<t}^*,\ T_v, C_b, R\right.\right) \tag{16}$$

where a sequence $S_{<t}^*$ consists of all words predicted before moment t.

Finally, the EGPT-2 network model is trained using the minimum negative log-likelihood loss function by using Eq. (17).

$$\mathcal{L}_\theta = -\sum_{i=1}^{|S|} \sum_{t=1}^{T^i} \log P_\theta\left(x_t^* \left| S_{<t}^*,\ C_b, T_v, R\right.\right) \tag{17}$$

where $|S|$ denotes the size of the training set, and T^i denotes the length of the i^{th} sentence.

In addition, this model is trained with a batch size of 8 for the MSVD dataset and 16 for the MSR-VTT dataset. The number of training iterations is set to 20, and the learning rate is set to 1×10^{-4}. Finally, aiming to avoid overfitting, the regularization rate of a Dropout algorithm is set to 0.1, and an Adam optimizer in the Stochastic Gradient Descent [39] is used for training optimization. When the training is completed, the minimum loss value on the training sets is used as the metric to select the best parameters of the EGPT-2 deep network model.

4 Experimental Evaluation

4.1 Dataset and Evaluation Indicator

This paper evaluates the proposed method's performance through extensive experiments using MSVD [40] and MSR-VTT [41] datasets. MSVD is a collection of 1970 open-domain video clips from YouTube covering various topics such as sports and cooking. Each video has around 40 natural sentences as labels. The dataset is partitioned into three sets: a training set of 1200 video clips, a validation set of 100 video clips, and a test set of 670 video clips. On the other hand, MSR-VTT is a larger dataset consisting of 10,000 videos and 200,000 captions. Each video has an average of 20 annotation sentences. The dataset is divided into a training set of 6513 video clips, a validation set of 609 video clips, and a test set of 2878 video clips.

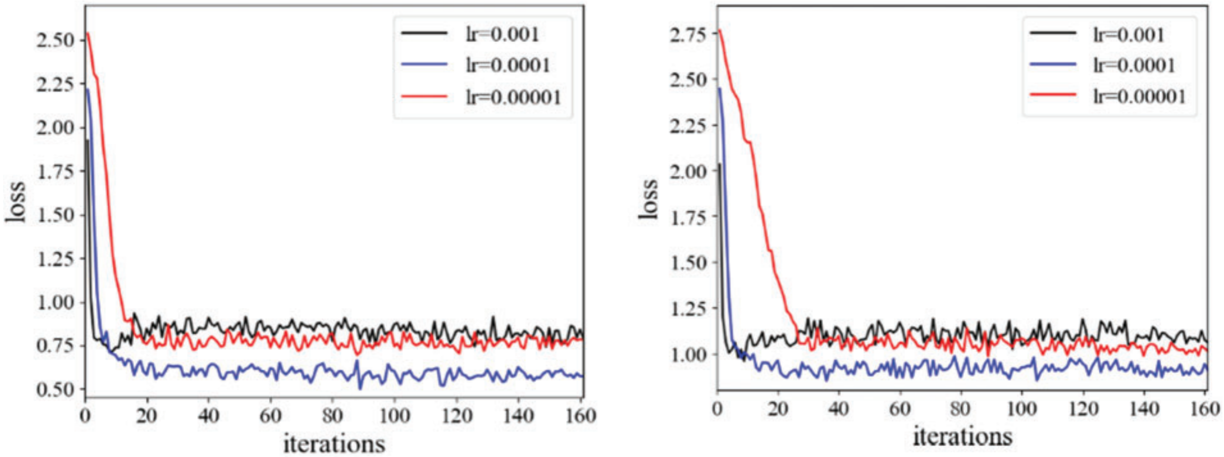
Currently, the evaluation indicators Bilingual Evaluation Understudy (BLEU) [42], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [43], Recall Oriented Understudy for Gisting Evaluation-longest common subsequence (ROUGE-L) [44], and Consensus-based Image Description Evaluation (CIDEr) [45] are widely utilized to evaluate the quality of the generated video captioning. Since the number of n-grams overlapped between the generated and reference sentences can better reflect the quality of video captioning, the performance of the BLEU-4 indicator in the experiment has also received attention.

The experiments are conducted using Python 3.8 programming language and Pytorch 1.7.0 framework for model training on a Linux operating system. The GPU used in the experiments is RTX A5000, with 42GB memory and 100GB hard disk size. Moreover, CUDA11.0 with cuDNN8.0 is utilized to accelerate the proposed method's computation.

4.2 Experimental Results Analysis

1) Training analysis of the proposed method

The proposed method utilizes 48000 reference sentences and 130260 reference sentences for model training in MSVD and MSR-VTT datasets, respectively. In the experiments, the optimal performance is achieved by setting the learning rate to 1×10^{-3}, 1×10^{-4}, and 1×10^{-5}, respectively. The experimental results are shown in Fig. 5. After 20 iterations, the loss function converges smoothly, and when the learning rate is set to 1×10^{-4}, the loss on the two datasets is the lowest, respectively. In addition, it can also be seen from Table 1 that when the learning rate is 1×10^{-4}, the indicators of the proposed method are optimal. It is considered that when the learning rate is very small, this method may fall into local optimization. On the other hand, if the learning rate is set too high, the output error of the proposed method will have a larger influence on the network parameters during the backpropagation process. This, in turn, causes the parameters to update too rapidly, making it difficult to converge. When the learning rate is 1×10^{-4}, the defined negative log-likelihood loss function (Eq. (17)) can converge and achieve the optimal training effect by updating the network parameters in each iteration.



(a) The curves of loss variation for different learning rates on the MSVD dataset

(b) The curves of loss variation for different learning rates on the MSR-VTT dataset

Figure 5: The curve of loss variation during the EGPT-2 deep network model training on different datasets

Table 1: Comparison of results with different learning rates on the MSR-VTT dataset

lr	BLEU4	METEOR	ROUGE-L	CIDEr
$lr = 10^{-3}$	38.5	27.3	59.8	48.8
$\boldsymbol{lr} = 10^{-4}$	**40.8**	**28.5**	**61.1**	**50.5**
$lr = 10^{-5}$	39.0	27.6	60.2	49.4

To evaluate the effectiveness of the proposed method, the performance indicators at different learning rates are further compared on the MSR-VTT dataset. The experiment results are presented in Table 1.

The results in Table 1 demonstrate the effectiveness of the proposed method in generating video captions that are coherent with the video content without overfitting.

2) Ablation experiments

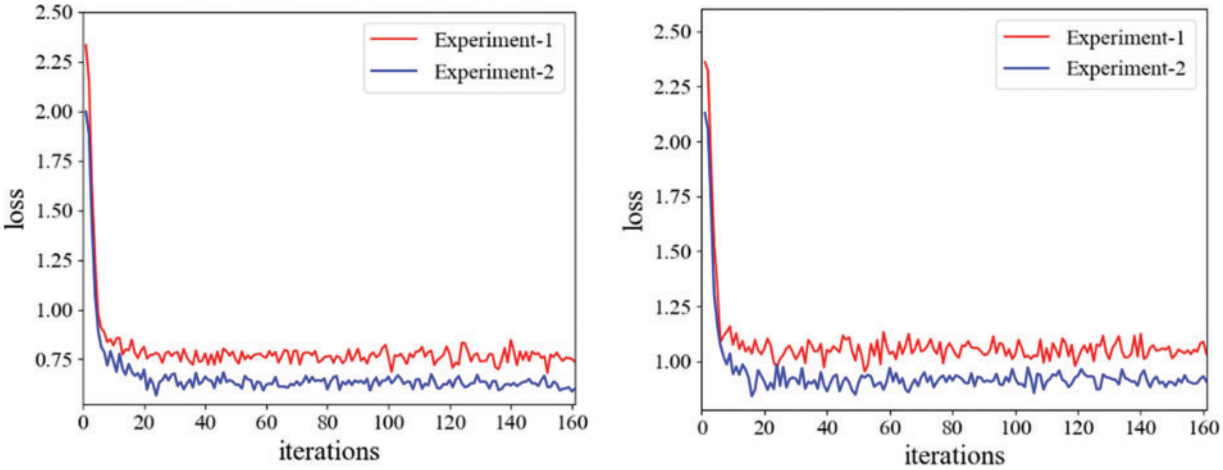
To clarify the respective contributions of the video semantic topics and the Enhance-TopK sampling algorithm to prediction words, the following ablation experiments are conducted in two datasets.

① Experiment-1: the input data of the proposed method only contain video clips and their baseline captions C_b, and the target words are predicted by a limited number of semantic features of baseline captions.

② Experiment-2: the input data of the proposed method contain video clips, their baseline captions C_b, and semantic topics T_v, and the baseline captions and semantic topics are jointly decoded to verify the effectiveness of introducing video semantic topics.

③ Experiment-3: based on the above experiments, the Enhance-TopK sampling algorithm is used in the decoding phase to verify its effectiveness.

It should be noted that to ensure the training effect of the EGPT-2 deep network model, Experiment-1 and Experiment-2 both utilize the normal greedy sampling algorithm in the training phase, and Experiment-3 is conducted based on the above experiments, which is only used the verify the effectiveness of the Enhance-TopK sampling algorithm, and does not involve the training phase. Therefore, the comparison results of training the EGPT-2 deep network model through Experiment-1 and Experiment-2 are shown in Fig. 6.



(a) The comparison of training results on the MSVD dataset

(b) The comparison of training results on the MSR-VTT dataset

Figure 6: The comparison results of training the EGPT-2 deep network model

From Fig. 6, it can be seen that Experiment-1 and Experiment-2 both start to converge smoothly after 20 iterations, and the loss of Experiment-2 is significantly lower than Experiment-1. It is considered that since Experiment-2 embeds the semantic topic features of videos in the encoding and decoding process, it can predict the target words with the help of semantic topics to improve the accuracy of video captioning. For example, when the semantic topic of an input video contains the semantics “slicing” and “cucumber”, the generated video captioning will be more specific, and there will be no blurred descriptions such as "A person is cooking". To further verify the impact of the semantic topics on the model performance, the comparative experiments separately are conducted on the two datasets, and the experimental results are shown in Tables 2 and 3.

Table 2: Comparison of results of ablation experiments on the MSVD dataset

Model	BLEU4	METEOR	ROUGE-L	CIDEr
Experiment-1	50.3	34.4	69.5	85.1
Experiment-2	**52.6**	**35.4**	**71.8**	**88.3 (3.2%↑)**

Table 3: Comparison of results of ablation experiments on the MSR-VTT dataset

Model	BLEU4	METEOR	ROUGE-L	CIDEr
Experiment-1	39.3	27.7	59.4	48.6
Experiment-2	**40.8**	**28.5**	**61.1**	**50.5 (1.9%↑)**

Tables 2 and 3 explore the impact of topics on model performance through ablation experiments. It can be seen that all indicators of Experiment-2 are superior to Experiment-1 on both datasets, especially on the indicator CIDEr, which has been significantly improved by 3.2% on MSVD and 1.9% on MSR-VTT, respectively. This is because Experiment-2 captures the topic information of video content through video semantic topics. During the decoding stage, additional information can be obtained from the video semantic topic features to predict the generation of sentences, making the generated video captioning both consistent with the video content and accurate. Specifically, the CIDEr metric is calculated using the TF-IDF algorithm, which assigns low weights to infrequent *n*-grams and high weights to the core *n*-grams in a sentence. A high CIDEr score indicates that the decoder can further accurately capture the semantic information of video data using video semantic topics. Thereby ensuring that the generated sentences of video content are more consistent with human consensus.

In addition, the Enhanced-TopK sampling used in the decoding phase is also effective. As shown in Tables 4 and 5, Experiment-3 has achieved good results on both datasets. It is considered that since the Enhance-TopK sampling algorithm utilizes the topic correlation coefficient λ, the probability of words affected by the topic in the prediction word set G is enhanced. By adjusting the probability distribution of the predicted words, the long-tail effect in the decoding stage is suppressed, and the accuracy of generated video captioning is further improved.

Table 4: Experiment-3 experimental results on the MSVD dataset

Method	BLEU4	METEOR	ROUGE-L	CIDEr
Experiment-2	52.6	35.4	71.8	88.3
Experiment-3	**53.5**	**35.8**	**72.7**	**90.1**

Table 5: Experiment-3 experimental results on the MSR-VTT dataset

Method	BLEU4	METEOR	ROUGE-L	CIDEr
Experiment-2	40.8	28.5	61.1	50.5
Experiment-3	**41.4**	**29.4**	**62.0**	**51.9**

3) Parameter setting

① *The sampling interval of the Enhance-TopK algorithm*

To ensure the generated video captioning conforms to the video content, which maintains the smoothness and rationality of the natural language, a novel Enhance-TopK sampling algorithm is constructed in the decoding stage. In this algorithm, to determine the value of the optimal initial sampling interval K, K is set to 10, 20, 30, 40, and 50 for experimental verification on the MSVD video dataset. The experimental results are shown in Table 6.

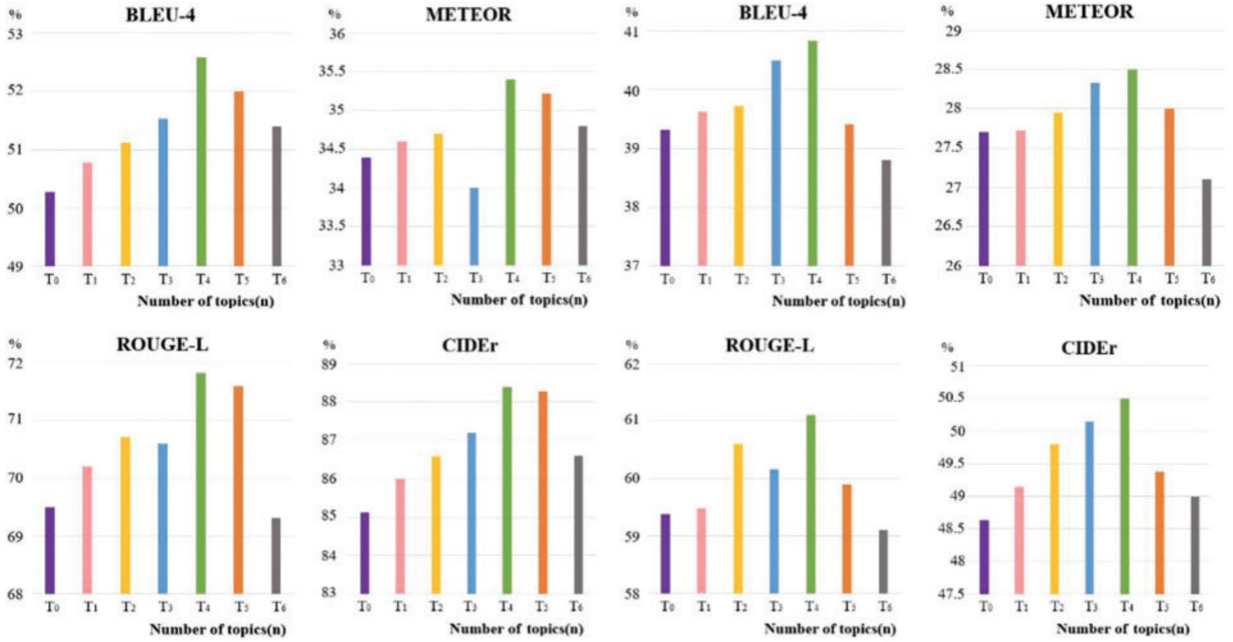
Table 6: Comparison of experimental results on different sampling intervals of Top-K

Top-K	BLEU4	METEOR	ROUGE-L	CIDEr
$K = 10$	48.6	33.6	66.7	83.4
$K = 20$	49.4	33.8	67.3	84.9
$\boldsymbol{K = 30}$	**50.3**	**34.4**	**69.5**	**85.1**
$K = 40$	49.9	34.2	66.2	84.7
$K = 50$	49.2	33.0	67.7	83.2

The results in Table 6 show that when the K value of the sampling interval is equal to 30, the score of each indicator is highest. It is considered that when the K value is less than 30, the sampling interval is small, leading to the next predicted word being likely to be sampled by using the maximum probability to generate repeated or flat sentences, such as the sentence "A woman is riding horse riding horse" resulting in the poor semantics of the generated video captioning. When the K value is greater than 30, the sampling interval is too large, which may cause the mode to sample the long tail words and make the generated sentences impassable. Therefore, the K value of the initial sampling interval is set to 30.

② *Parameter setting of the semantic topic quantity*

The method proposed in this paper can learn additional semantic information by embedding the semantic topic of the video content to predict the next word. In this process, the parameter setting of the semantic topic quantity is very important. To obtain an appropriate parameter of visual semantic topics, 0 to 6 topic words are set for each video to verify the performance of the method. The experimental results are shown in Fig. 7.



(a) Comparison of performance on the MSVD dataset

(b) Comparison of performance on the MSR-VTT dataset

Figure 7: Comparison of performance metrics for different number of topics on MSVD and MSRVTT datasets

To evaluate the impact of the Enhance-TopK sampling algorithm on the performance of the proposed method, comparative experiments with different numbers of topic words are conducted in Experiment-2. It can be seen from Fig. 7 that when the number of topic words is 4, all indicators of the proposed method on the two datasets can achieve the best ones. It is considered that when the number of topic words is small, the EGPT-2 decoder will lose some key objects and actions in the video content. On the contrary, when too many topic words are selected, some irrelevant noise will be introduced into the EGPT-2 decoder, leading to the generated video captioning deviating from the video content. For example, when generating the sentence “A monkey is grasping the dog’s tail.”, if the number of topic words is set to 0, the generated video captioning may lose some key information about the monkey, such as “attack” and “grab” actions. When the number of topic words is set to 6, the irrelevant semantic noise such as “animals” and “roadside” may affect the generation of video captioning that is inconsistent with the video content, such as the sentence “Monkeys are dragging an animal off the road.”. Therefore, according to the results in Fig. 7, the topic words are finally set to 4 for each video data.

4) Comparative analysis with different video captioning methods

To evaluate the performance of the proposed method, it is compared with other video captioning studies. Tables 7 and 8 show the comparison results on the MSVD and MSR-VTT datasets.

It can be seen from Tables 7 and 8 that the proposed method shows better performance on two datasets compared with S2VT [20], SAAT [23], PI-LSTM [24], TTA [28], CSA-SR [46], GRU-EVE

[47], ADL [48], RecNet [49], STM [50], TDDF [51], PickNet [52], and CoSB [53] methods, especially the advantages of CIDEr indicator is obvious. It is considered that since the CIDEr indicator is used to measure the amount of video-related information contained in the contents generated by the models, and the method proposed utilizes the EGPT-2 decoder to jointly decode the semantic topics and baseline of video captioning simultaneously. This alignment of prediction words with video content enhances semantic consistency, and the generated video captioning contains more semantic information about video content. In addition, the Enhance-TopK sampling algorithm can be used in the vocabulary prediction stage to improve the accuracy of word prediction, which makes the generated video captioning smooth and closer to the semantics of video content.

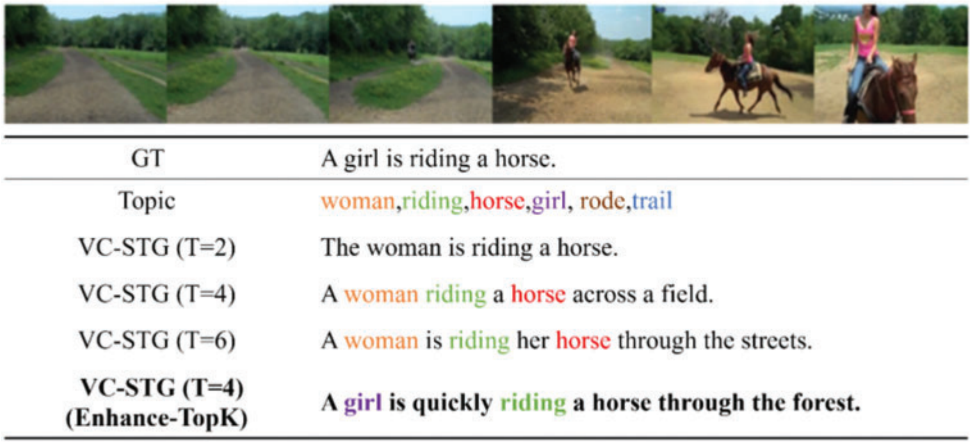
Table 7: Comparison of experimental results of multiple methods on the MSVD dataset

Model	BLEU4	METEOR	ROUGE-L	CIDEr
S2VT [20]	—	29.8	—	—
SAAT [23]	46.5	33.5	69.4	81.0
PI-LSTM [24]	48.7	34.0	69.8	85.4
TTA [28]	51.8	35.5	72.4	87.7
CSA-SR [46]	52.2	35.6	72.7	83.4
GRU-EVE [47]	47.9	35.0	71.5	78.1
ADL [48]	53.1	35.7	70.4	81.6
RecNet [49]	52.3	34.1	69.8	80.3
aLSTM [50]	50.8	33.3	—	74.8
TDDF [51]	45.8	33.3	—	73.8
PickNet [52]	52.3	33.3	69.6	76.5
The proposed method	**53.5**	**35.8**	72.7	**90.1**

Table 8: Comparison of experimental results of multiple methods on the MSR-VTT dataset

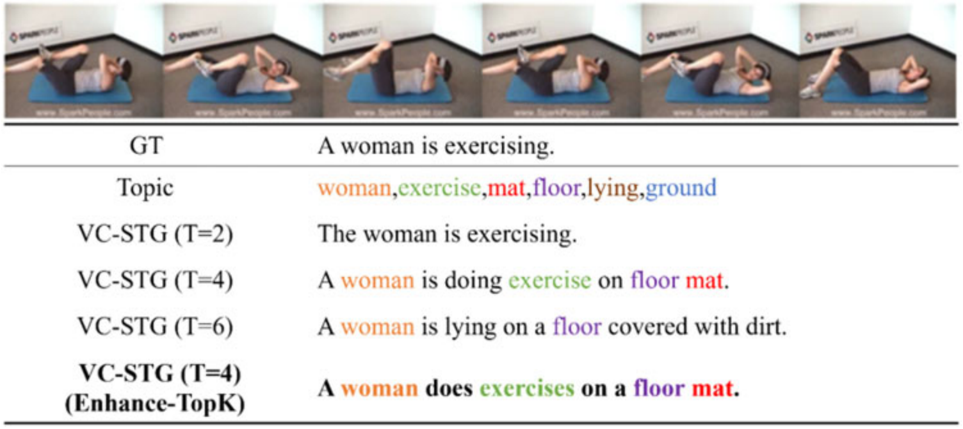
Model	BLEU4	METEOR	ROUGE-L	CIDEr
S2VT [20]	—	25.7	—	—
SAAT [23]	40.5	28.2	60.9	49.1
TTA [28]	41.4	27.7	61.1	46.7
CSA-SR [46]	41.3	28.2	61.9	48.6
GRU-EVE [47]	38.3	28.4	60.7	48.1
RecNet [49]	39.1	26.6	59.3	42.7
aLSTM [50]	38.0	26.1	—	43.2
TDDF [51]	37.3	27.8	—	43.8
PickNet [52]	41.3	27.7	59.8	44.1
CoSB [53]	41.4	27.8	61.0	46.5
The proposed method	**41.4**	**29.4**	**62.0**	**51.9**

Finally, we utilize the proposed method (denoted as VC-STG) to generate the video captioning on the MSVD and MSR-VTT datasets, as shown in Figs. 8 and 9. In Fig. 8, "GT" denotes a reference sentence of a video, and "Topic" denotes the topic words of a video. To reflect the impact of the semantic topics on the performance of the proposed method, 2, 4, and 6 topic words are selected to compare the results of the video captioning. It can be seen from Fig. 8 that when the given input video contains the baseline and topics of the video content, it will output a relatively complete video captioning. However, when the number of topic words is large (the number of topic words $T = 6$), semantic noise may be introduced into the EGPT-2 decoder, resulting in a deviation in the generated video captioning, such as "through the streets.". When the optimal number of topic words is $T = 4$, the Enhance-TopK sampling algorithm can be used to help generate more accurate video captioning, such as the generated word "quickly" and the phrase "through the forest".



GT	A girl is riding a horse.
Topic	woman,riding,horse,girl, rode,trail
VC-STG (T=2)	The woman is riding a horse.
VC-STG (T=4)	A woman riding a horse across a field.
VC-STG (T=6)	A woman is riding her horse through the streets.
VC-STG (T=4) (Enhance-TopK)	**A girl is quickly riding a horse through the forest.**

Figure 8: A video captioning under the MSVD test set



GT	A woman is exercising.
Topic	woman,exercise,mat,floor,lying,ground
VC-STG (T=2)	The woman is exercising.
VC-STG (T=4)	A woman is doing exercise on floor mat.
VC-STG (T=6)	A woman is lying on a floor covered with dirt.
VC-STG (T=4) (Enhance-TopK)	**A woman does exercises on a floor mat.**

Figure 9: A video captioning under the MSR-VTT test set

In addition, the impact of semantic topics on video captioning is verified. When the number of topic words is 4 ($T = 4$), the accuracy of prediction words can be significantly improved. However, having too few or too many topic words cannot improve video captioning. For example, the error message “lying on a floor covered with dirt.” may be predicted according to the words “lying”, “floor” and “ground”.

It can be seen from Figs. 8 and 9 that when the number of topic words is 4, the method proposed can generate a video captioning that conforms to the video content. On this basis, when the Enhance-TopK sampling algorithm is used in the decoding stage, the generated video captioning is more accurate and the effect is better.

5 Conclusion

To address the issue that the existing video captioning methods developed based on encoder-decoder rely on a single video input source, this paper proposes a video captioning method based on semantic topic-guided generation to improve the accuracy of video captioning, which can enhance the alignment between visual information and natural language by introducing the semantic topics of video data and guide the generation of video captioning. The proposed method is verified with two common MSVD and MSR-VTT datasets. The experimental results demonstrate that the proposed method outperforms several state-of-art approaches. Specifically, the performance indicators BLEU, METEOR, ROUGE-L, and CIDEr of the proposed method are improved by 1.2%, 0.1%, 0.3%, and 2.4% on the MSVD dataset, and 0.1%, 1.0%, 0.1%, and 2.8% on the MSR-VTT dataset, respectively, compared with the existing video captioning methods. The introduction of semantic topics can be effective in generating topic-related video captioning and improving the generation effect of video content by a decoder. However, since the limited baselines of video data in public video datasets, the extraction of video semantic topics is restricted. Future studies will introduce target detection algorithms to capture fine-grained semantic information and combine the attention mechanism to eliminate irrelevant and interfering semantic information.

Acknowledgement: The authors would like to express their gratitude to the College of Computer Science and Technology, Xi’an University of Science and Technology (XUST), Xi’an, China, and China Postdoctoral Science Foundation (CPSF) for providing valuable support to carry out this work at XUST.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 61873277, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2020JQ-758, and in part by the Chinese Postdoctoral Science Foundation under Grant 2020M673446.

Author Contributions: Study conception and design: Ou Ye; data collection: Ou Ye, Xinli Wei; analysis and interpretation of results: Zhenhua Yu, Xinli Wei; draft manuscript preparation: Yan Fu, Ying Yang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Datasets used for generating the results reported in Sections 3 and 4 are available at https://opendatalab.com/OpenDataLab/MSVD and https://www.kaggle.com/datasets/vishnutheepb/msrvtt?resource=download.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. Khurana and U. Deshpande, "Video question-answering techniques, benchmark datasets evaluation metrics leveraging video captioning: A comprehensive survey," *IEEE Access*, vol. 9, pp. 43799–43823, 2021.

[2] Z. Yu, A. Sohail, M. Jamil, O. A. Beg and J. M. R. S. Tavares, "Hybrid algorithm for the classification of fractal designs and images," *Fractals*, 2022. https://doi.org/10.1142/S0218348X23400030

[3] Z. Yu, H. Gao, X. Cong, N. Wu and H. H. Song, "A survey on cyber-physical systems security," *IEEE Internet of Things Journal*, 2023. https://doi.org/10.1109/JIOT.2023.3289625

[4] M. Adimoolam, S. Mohan, A. John and G. Srivastava, "A novel technique to detect and track multiple objects in dynamic video surveillance systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 112–120, 2022.

[5] L. Gao, Y. Lei, P. Zeng, J. Song, M. Wang *et al.,* "Hierarchical representation network with auxiliary tasks for video captioning and video question answering," *IEEE Transactions on Image Processing*, vol. 31, pp. 202–215, 2022.

[6] N. Aafaq, A. Mian, W. Liu, N. Akhtar and M. Shah, "Cross-domain modality fusion for dense video captioning," *IEEE Transaction on Aritificial Intelligence*, vol. 3, no. 5, pp. 763–777, 2022.

[7] J. Perez-Martin, B. Bustos, S. J. F. Guimarães, I. Sipiran, J. Pérez *et al.,* "A comprehensive review of the video-to-text problem," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 1–75, 2021.

[8] R. Xu, C. Xiong, W. Chen and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *Proc. of the Twenty-Ninth AAAI Conf. on Artificial Intelligence*, Austin, Texas, USA, pp. 2346–2352, 2015.

[9] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg and Y. Choi, "Collective generation of natural image descriptions," in *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea, pp. 359–368, 2012.

[10] C. Cheng, C. Li, Y. Han and Y. Zhu, "A semi-supervised deep learning image caption model based on Pseudo Label and N-gram," *International Journal of Approximate Reasoning*, vol. 131, pp. 93–107, 2021.

[11] R. Leblond, J. Alayrac, L. Sifre, M. Pislar, L. Jean-Baptiste *et al.,* "Machine translation decoding beyond beam search," in *Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, pp. 8410–8434, 2021.

[12] K. Lin, Z. Gan and L. Wang, "Augmented partial mutual learning with frame masking for video captioning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, California, USA, pp. 2047–2055, 2021.

[13] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng *et al.*, "End-to-end dense video captioning with parallel decoding," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 6827–6837, 2021.

[14] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal *et al.,* "Describing videos by exploiting temporal structure," in *IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 4507–4515, 2015.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *2015 IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 4489–4497, 2015.

[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li *et al.,* "BabyTalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[17] Y. Yang, C. Teo, H. Daumé and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, pp. 444–454, 2011.

[18] V. Ordonez, G. Kulkarni and T. L. Ber, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. of the 24th Int. Conf. on Neural Information Processing Systems (NIPS'11)*, Red Hook, NY, USA: Curran Associates Inc., pp. 1143–1151, 2011.

[19] A. Gupta, Y. Verma and C. Jawahar, "Choosing linguistics over vision to describe images," in *Proc. of the AAAI Conf. on Artificial Intelligence*, California, USA, pp. 606–612, 2012.

[20] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell *et al.,* "Sequence to sequence–video to text," in *IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 4534–4542, 2015.

[21] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney *et al.,* "Translating videos to natural language using deep recurrent neural networks," in *Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, USA, pp. 1494–1504, 2015.

[22] L. Li, X. Gao, J. Deng, Y. Tu, Z. J. Zha *et al.,* "Long short-term relation transformer with global gating for video captioning," *IEEE Transactions on Image Processing*, vol. 31, pp. 2726–2738, 2022.

[23] Q. Zheng, C. Wang and D. Tao, "Syntax-aware action targeting for video captioning," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 13093–13102, 2020.

[24] Y. Li, X. Cui and X. Jin, "Research on video captioning method based on semantic key frame," in *2nd Asia-Pacific Conf. on Communications Technology and Computer Science (ACCTCS)*, Shenyang, China, pp. 39–44, 2022.

[25] F. Liu, X. Wu, C. You, S. Ge, Y. Zou *et al.,* "Aligning source visual and target language domains for unpaired video captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9255–9268, 2022.

[26] L. Yan, S. Ma, Q. Wang, Y. Chen, X. Zhang *et al.,* "Video captioning using global-local representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6642–6656, 2022.

[27] W. Ji and R. Wang, "A multi-instance multi-label dual learning approach for video captioning," *ACM Transactions on Multimidia Computing Communications and Applications*, vol. 17, no. 2s, pp. 1–18, 2021.

[28] Y. Tu, C. Zhou, J. Guo, S. Gao and Z. Yu, "Enhancing the alignment between target words and corresponding frames for video captioning," *Pattern Recognition*, vol. 111, pp. 1–11, 2021.

[29] W. Hu, L. Wu, M. Jian, Y. Chen and H. Yu, "Cosine metric supervised deep hashing with balanced similarity," *Neuro Computing*, vol. 448, pp. 94–105, 2021.

[30] X. Wang, L. Zhu and Y. Yang, "T2VLAD: Global-local sequence alignment for text-video retrieval," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 5075–5084, 2021.

[31] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[32] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas and B. Jiang, "Evaluation of clustering and topic modeling methods over health-related tweets and emails," *Artificial Intelligence in Medicine*, vol. 117, pp. 102096, 2021.

[33] R. Rani and D. K. Lobiyal, "An extractive text summarization approach using tagged-LDA based topic modeling," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3275–3305, 2021.

[34] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio *et al.,* "TAP: Text-aware pre-training for text-VQA and text-caption," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 8747–8757, 2021.

[35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei *et al.,* "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, pp. 1–24, 2019.

[36] C. Deng, S. Chen, D. Chen, Y. He and Q. Wu, "Sketch, ground, and refine: Top-down dense video captioning," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 234–243, 2021.

[37] Z. Zhang, C. Yuan, Y. Shan, B. Li, Y. Den *et al.,* "Open-book video captioning with retrieve-copy-generate network," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 9832–9841, 2021.

[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-Networks," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 3982–3992, 2019.

[39] H. Tang, S. Gan, A. A. Awan, S. Rajbhandari, C. Li *et al.,* "1-bit Adam: Communication efficient large-scale training with adam's convergence speed," in *Proc. of the 38th Int. Conf. on Machine Learning*, pp. 10118–10129, 2021.

[40] D. L. Chen and B. D. William, "Collecting highly parallel data for paraphrase evaluation," in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 190–200, 2011.

[41] J. Xu, T. Mei, T. Yao and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 5288–5296, 2016.

[42] K. Papineni, S. Roukos, T. Ward and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 311–318, 2002.

[43] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, USA, pp. 65–72, 2005.

[44] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branchesout*, Barcelona, Spain, pp. 74–81, 2004.

[45] R. Vedantam, C. L. Zitnick and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 4566–4575, 2015.

[46] Z. Lei and Y. Huang, "Video captioning based on channel soft attention and semantic reconstructor," *Future Internet*, vol. 13, no. 2, pp. 1–18, 2021.

[47] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 12479–12488, 2019.

[48] W. Ji, R. Wang, Y. Tian and X. Wang, "An attention based dual learning approach for video captioning," *Applied Soft Computing*, vol. 117, pp. 1–9, 2022.

[49] B. Wang, L. Ma, W. Zhang and W. Liu, "Reconstruction network for video captioning," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7622–7631, 2018.

[50] L. Gao, Z. Guo, H. Zhang, X. Xu and H. T. Shen, "Video captioning with attention based LSTM and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[51] X. Zhang, K. Gao, Y. Zhang, D. Zhang, J. Li *et al.,* "Task-driven dynamic fusion: Reducing ambiguity in video description," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii, USA, pp. 3713–3721, 2017.

[52] Y. Chen, S. Wang, W. Zhang and Q. Huang, "Less is more: Picking informative frames for video captioning," in *Proc. of the European Conf. on Computer Vision*, Munich, Germany, pp. 367–384, 2018.

[53] J. Vaidya, A. Subramaniam and A. Mittal, "Co-segmentation aided two-stream architecture for video captioning," in *IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, pp. 2442–2452, 2022.