Multimodal Feature Fusion Video Description Model Integrating Attention Mechanisms and Contrastive Learning

Wang Zhihao*, Che Zhanbin
School of Computer Science, Zhongyuan University of Technology, Zhengzhou, China

***Abstract*—To avoid the issue of significant redundancy in the spatiotemporal features extracted from multimodal video description methods and the substantial semantic gaps between different modalities within video data. Building upon the TimeSformer model, this paper proposes a two-stage video description approach (Multimodal Feature Fusion Video Description Model Integrating Attention Mechanism and Contrastive Learning, MFFCL). The TimeSformer encoder extracts spatiotemporal attention features from the input video and performs feature selection. Contrastive learning is employed to establish semantic associations between the spatiotemporal attention features and textual descriptions. Finally, GPT2 is employed to generate descriptive text. Experimental validations on the MAVD, MSR-VTT, and VATEX datasets were conducted against several typical benchmark methods, including Swin-BERT and GIT. The results indicate that the proposed method achieves outstanding performance on metrics such as Bleu-4, METEOR, ROUGE-L, and CIDEr. The spatiotemporal attention features extracted by the model can fully express the video content and that the language model can generate complete video description text.**

Keywords*—Multimodal feature fusion; video description; spatiotemporal attention; comparative learning

I. INTRODUCTION

Video description is a field of deep learning with practical value, and it has application value in areas such as assisting visually impaired individuals in accessing video content and video content analysis. When dealing with video content from real life, video description models require complex preprocessing steps, such as frame extraction and normalization operations, followed by feature extraction, and finally, the transformation of these features into linguistic descriptions. In this process, the model must not only delve into the semantic content of the video but also establish precise correspondences between visual and textual information to generate accurate descriptions [1]. However, due to the excessive redundancy in video data and the vast semantic gap between modalities, it is challenging for models to establish a unified representation of these modalities and to capture key information accurately for detailed descriptions.

To address the aforementioned issues, this paper proposes a two-stage multimodal feature fusion method. In the first stage, the TimeSformer encoder [2] is employed to extract spatiotemporal features from the input video. The TimeSformer is a transformer-based model for video action recognition that effectively captures spatiotemporal features within videos. The model generates video vectors rich in semantic features by dividing video frames into non-overlapping blocks and applying attention mechanisms in both temporal and spatial dimensions. After spatiotemporal feature extraction, these vectors undergo feature selection and are used as visual cues input into the GPT-2 model to generate video descriptions. This paper employs a contrastive learning approach in the second stage to align video embeddings with text embeddings in the latent space. Video-text contrastive learning is an effective training method that minimizes the semantic differences between different modalities by pulling closer the representations of the same entity across modalities and pushing apart the representations of different entities. This method enhances the similarity between the features output by TimeSformer and the corresponding textual descriptions. Experimental evidence suggests that the scheme incorporating contrastive learning is easier to train than the one that fine-tunes TimeSformer directly with video descriptions without contrastive learning. This ease of training may be attributed to the reduced involvement of generative methods, which typically require more extensive training time when contrastive learning is not employed.

This paper conducts comparative experiments to validate the effectiveness of the proposed model, and the results indicate that the proposed model achieves state-of-the-art results on the MSVD, MSR-VTT, and VATEX datasets. Compared to existing models, the text generated by this model is capable of providing a comprehensive description of video content and is straightforward to train. The contributions of this paper are as follows:

1) This paper proposes a two-stage multimodal feature fusion method that combines spatiotemporal attention with contrastive learning, efficiently integrating and utilizing temporal, visual, and textual features.

2) Within the training process, this paper conducts feature selection on spatiotemporal features to prevent redundant information from entering the language model, which could otherwise interfere with text generation.

3) Experimental evidence demonstrates that our method can effectively comprehend and describe the rich multimodal information within videos, achieving advanced results compared to similar models in the field.

This research was funded by Key Technology Research and Demonstration Application of News Intelligent Production (212102210417), Science and Technology Plan of Henan Province in 2021.

II. Related Work

When processing videos, video descriptions require the extraction of temporal information, image information, and other modal information from the video. These data are then multimodal fused to build a joint representation between modalities. The resulting representation serves as visual cues input into a text model to generate textual descriptions. This section discusses related work from spatiotemporal feature extraction and multimodal feature fusion perspectives.

A. Spatiotemporal Feature Extraction

In video description tasks, models must be capable of extracting temporal and spatial features from the video content. For temporal feature extraction, common methods include 3D Convolutional Neural Networks (CNN) and optical flow-based networks. As for spatial features that pertain to image characteristics, one can utilize popular pre-trained image feature extraction networks such as ResNet [3] and Vision Transformer [4].

DC-RL [5] employs a 3D CNN to model temporal features and concatenates these features with image features obtained from a pre-trained image encoder using an LSTM. However, this approach may yield little improvements over previous methods. This is because of the inherent locality of 3D neural networks, which limits their ability to learn long-term temporal dependencies. Moreover, LSTM are prone to vanishing or exploding gradients when dealing with long sequences of input temporal information, making them difficult to train effectively.

MA-Net [6] employs the Inflated 3D (I3D) [7] network to model temporal relationships and constructs semantic feature vectors from the textual descriptions of the video. These semantic feature vectors are then used alongside the video feature vectors for semantic detection, aiming to bridge the gap between the semantic video features and the actual semantic content of the video. I3D expands pre-trained 2D CNN into 3D CNN by "inflating" their 2D filters into 3D filters, allowing the network to capture spatiotemporal information in video data. Experimental results have shown improvements compared to DC-RL. However, it does not overcome the drawbacks of CNN, which fails to model long-range temporal information, and the extracted video features cannot fully represent the content of the video.

This paper addresses the issues above by employing TimeSformer. TimeSformer divides the video into non-overlapping spatial and temporal patches. It then applies attention mechanisms between patches that belong to the same spatial location but different time points and between patches from different spatial locations but the same time point. This approach enables efficient extraction of temporal and spatial features across the entire video, making it highly suitable for video description models.

B. Multimodal Feature Fusion

Videos are composed of multimodal data, including visual, audio, and textual components, each containing a vast amount of information accompanied by noise and uncertainty. To accurately describe the content of videos, it is necessary to employ multimodal feature fusion techniques to achieve complementarity and verification between different modalities. This approach facilitates a comprehensive understanding and analysis of video content, enhancing the accuracy and reliability of information processing. When performing multimodal feature fusion, common strategies include the following: (1) Data Fusion: This strategy involves combining information from multiple modalities through operations such as concatenation, addition, and multiplication and then passing the integrated data to subsequent processes. The advantages of this approach are its simplicity and the absence of the need for additional training methods. However, it may also lead to information redundancy or the inability to leverage complex modality information. (2) Neural Network Fusion: This strategy involves using neural network methods to jointly encode multimodal information based on data fusion or directly accepting multimodal inputs. For instance, the cross-attention mechanism in Transformer[8] utilizes self-attention to relate and fuse information from different modalities. The Feature Pyramid Network (FPN)[9] achieves fusion by constructing feature maps at different scales, which allows for information integration at various levels, thereby enhancing the performance of object detection tasks.

Fu et al. [10] have utilized the attention mechanism to query the relationship between object detection and action features, establishing a subject-verb relationship from a grammatical perspective and generating text accordingly. The primary benefit of this approach lies in the generation of coherent textual output. However, this grammar-based method has limitations in terms of text generation diversity. Since it focuses on establishing accurate grammatical structures, the generated text often follows similar sentence templates, which can lead to stiff and repetitive expressions in language. Moreover, in Fu's work, action features extracted by 3D neural networks fail to capture long-distance temporal information, which can lead to the model that is incapable of fully describing the semantics of the video.

The work by Ren et al. [11] is similar to the present study, as it also employs an attention mechanism to integrate spatiotemporal features. Additionally, they designed a semantic enhancement network to learn the latent semantic information of object features. However, in fusing visual-textual features, they utilized Long Short-Term Memory networks (LSTM). LSTM is a type of recurrent neural network (RNN) that is capable of handling sequence data and can remember long-term dependencies. In their work, LSTM networks were employed to process textual information, obviating the need to add position vectors within the attention mechanism. Position vectors are commonly used in attention mechanisms to ensure the model can understand the order of elements in the input sequence. Although LSTM networks have certain advantages in processing sequence data, it also have limitations, particularly when dealing with long-distance dependencies. LSTM networks may encounter issues with vanishing or exploding gradients, limiting its effectiveness in modeling long-distance dependencies between elements in long sequences.

To address the aforementioned issues, this paper employs a Visual-Text Contrastive (VTC) learning method to align the spatiotemporal features output by TimeSformer with textual

features. By analyzing the similarities and differences between data samples, VTC can discern the relationships and discrepancies between spatiotemporal and textual features, ensuring that semantically similar spatiotemporal features remain close to their corresponding textual features in the latent space. This facilitates the ability of feature mapping module to map spatiotemporal features to textual laten spaces and makes it possible to generate textual descriptions via GPT2[12]. This approach allows the model to learn a complete representation of spatiotemporal features and generates accurate textual descriptions.

III. METHODOLOGY

When processing videos, TimeSformer first applies an attention mechanism in both temporal and spatial dimensions to extract spatiotemporal features of videos. Subsequently, this paper utilizes a fully connected network to map the spatiotemporal features into a feature sequence, adapting them for input into the Transformer Encoder, the mapping network. In preparation for the input to the Transformer Encoder, a learnable vector of length τ was concatenated to the video feature sequence to screen the sequence and prevent the inclusion of redundant information. The learnable vector, which can be referred to as a visual prompt, is then input into the language model for decoding, yielding a textual description of the video. During the training process, the semantic gap between video and text features poses a significant challenge. To address this issue, the paper introduces a Video-Text Contrastive Learning module, which aligns video and text features in the semantic space. Fig. 1 depicts the structure of the entire model.

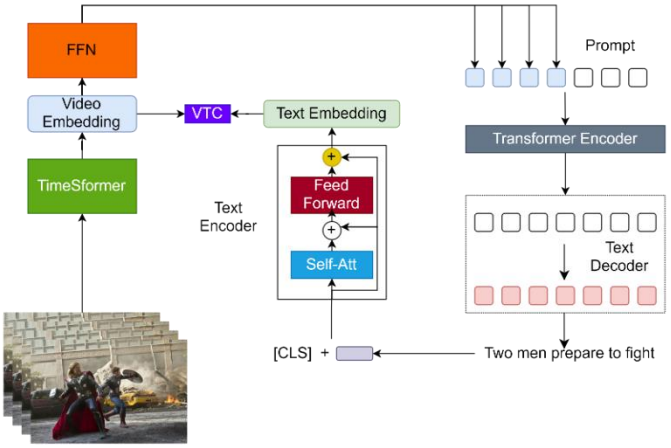

Fig. 1. Overall structure diagram.

A. Spatiotemporal Attention-based Feature Fusion

To fully comprehend video content, it is necessary to integrate both spatial and temporal information. In traditional 3D CNN, convolution operations are performed simultaneously across both temporal and spatial dimensions, allowing the model to capture spatiotemporal features within the video. However, 3D convolution operations are computationally intensive, leading to slower training and inference speeds for the model. To overcome this limitation, TimeSformer organizes the temporal and spatial dimensions into multiple video patches and performs attentional interactions on them separately, incorporating attention mechanisms into video understanding tasks. This approach has achieved outstanding results in the field of action recognition.

This paper employs TimeSformer to extract spatiotemporal features from videos. After preprocessing, a video is mapped into multiple non-overlapping patches through a linear layer, which is then input into the spatiotemporal feature extraction network, as illustrated in Fig. 2.

The output can be represented as $\{v_1, v_2, \ldots, v_\phi\}$, where $v_i \in R^{H_{video}}$ and ϕ denote the number of video patches. During the training of TimeSformer, the learnable vector CLS^{video} at the head of the video patch sequence is randomly initialized and incorporates the embedding representation of the entire video throughout the training process. In subsequent modules, the paper primarily uses CLS^{video} as the video feature for computation.

B. Visual Prompt Based on Contrastive Learning and Feature Selection

In the training process, the input video is first mapped using a linear layer to extract the embedded representation of the video. These representations capture the temporal and visual features between video frames. Subsequently, these video embeddings are concatenated with a learnable query vector of length τ and jointly input into a Transformer Encoder. The role of the query vector is to select the relevant features of the video that need to be described and filter out irrelevant information. Next, the query vector is used as a prompt and input into the GPT2 model.

Since the prompt already contains the necessary semantic information, the language model can generate readable text based on this prompt vector. This approach has been applied in

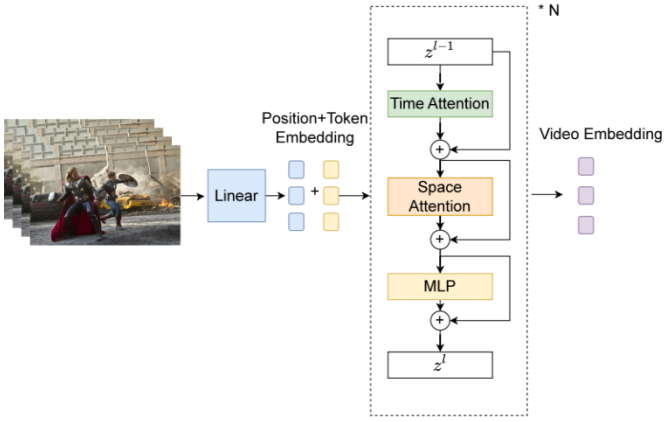

Fig. 2. TimeSformer architecture diagram.

the work by Zhou et al. [13]. However, a significant gap often exists between the vector spaces of video and text embeddings, even when they may be semantically related. This discrepancy makes it challenging to train directly using these embeddings. To address this issue, the paper introduces a contrastive learning approach. Contrastive learning aims to bridge the gap between visual and text embeddings in the

semantic space, even if they are far apart in the original space. By doing so, the model can learn how to map video content to relevant textual descriptions more easily, facilitating the training process.

Suppose the data consists of a dataset $\{V_i, T_i\}_{i=1}^{N}$ of N video-text pairs, where V represents the video and T corresponds to the associated textual description. Since GPT2 accepts input as a sequence of tokens, the paper maps the CLS^{video} obtained from TimeSformer to a sequence of video patches. Its formal representation is given in Eq. (1).

$$p_1, p_2, p_3, \ldots p_L = FFN(TimeSformer(CLS_{video})) \tag{1}$$

where, p_i is a one-dimensional vector of length H. In this paper, the hidden layer vectors for video and text are the same, set to 768.

During training, the paper concatenates the video patch vectors $\{p_1, p_2, p_3, \ldots p_L\}$ with randomly initialized learnable vectors $\{q_1, \ldots q_\tau\}$. It is then fed into the Transformer Encoder to obtain $\{p_1`, p_2`, p_3`, \ldots p_L`, q`_1, \ldots q`_\tau\}$, where L and τ are hyperparameters, with τ representing the length of the prefix.

The training objective is to predict the tokens autoregressively conditional on the prompt. The generation of the text loss objective can be described by the Eq. (2). The true distribution I_l is an indicator distribution that takes the value of 1 for the correct token t_l and 0 for all other tokens.

$$loss_{token} = -\sum_{i=1}^{L} \log p_\theta(t_1, \ldots, t_L \mid q`_1, \ldots q`_\tau) \tag{2}$$

In contrastive learning, the paper uses the CLS token as the video representation directly, which is paired with the text features extracted by BERT for contrastive learning. Similarly, the text feature utilizes BERT's CLS token, CLS^{text}, with the shape of R^H. The loss function is as follows in Eq. (3), where γ is the temperature parameter, taking the value of 0.07.

$$loss_c = -\log \frac{exp(CLS^{video} \cdot CLS^{text}{}_{+} / \gamma)}{\sum_{i=0}^{k} exp(CLS^{video} \cdot CLS^{text}{}_{i} / \gamma)} \tag{3}$$

IV. Experimental Verification

A. Dataset

In this paper, the experimental verification is carried out on three public datasets: MSVD, MSR-VTT, and VATEX

The Microsoft Research Video Description Corpus (MSVD) dataset [14] is widely used in video understanding models. Introduced by Microsoft Research in 2016, it is designed for video description generation, which automatically produces natural language descriptions for video clips. The dataset comprises 1,970 video segments, each accompanied by multiple English descriptions written by different individuals. The videos in the MSVD dataset are primarily sourced from YouTube and cover a variety of genres, including music videos, movie trailers, television shows, and more.

The Microsoft Research Video to Text [15] (MSR-VTT) dataset is another widely used dataset in video understanding models. Proposed by Microsoft Research in 2016 is also designed for video description generation, where the goal is to produce natural language descriptions for video clips automatically. This dataset comprises over 10,000 video segments, each with at least one English description. The videos in the MSR-VTT dataset are primarily sourced from YouTube and encompass a variety of genres, such as music videos, movie trailers, television shows, and more.

The Video-and-Text EXchange (VATEX) dataset[16] is a large-scale video description and subtitle dataset that contains multimodal information including video, audio, and text data. It is characterized by its vast scale, comprising over 250,000 pairs of videos and subtitle descriptions, covering multiple languages, with a particular focus on Chinese and English. The data is sourced from various scenes, including movies, TV series, news broadcasts, variety shows, and more, which results in a highly diverse dataset in both content and language.

B. Experimental Setup

All experiments were conducted using the PyTorch deep learning framework on two GTX-3090 GPUs. The model employed Adaptive Moment Estimation (Adam) as the optimization strategy and used the cross-entropy function as the loss function for back propagating gradients. The weight decay parameter was set to 0.009. The learning rate was scheduled using an inverse time scheme, as shown in the Eq. (4), where ò was set to 0.5, and the initial learning rate r was set to 0.01.

$$r_{new} = \frac{r}{1 + ò * step} \tag{4}$$

In the MSVD dataset, a total of 14,910 steps were trained with a batch size of 16; in the MSR-VTT dataset, a total of 25,320 steps were trained with a batch size of 8; and in the VATEX dataset, a total of 113,350 steps were trained with a batch size of 6.

C. Comparative Experiments

The proposed model is compared with state-of-the-art methods. MA-Net[6] attempts to construct semantic feature vectors from the textual description of videos, which are then used in conjunction with video feature vectors for semantic detection, to bridge the gap between semantic video features and the actual semantics of the video. However, the I3D network employed by the authors cannot model long-range temporal relationships.

MGRMP [17] proposes a recurrent regional attention module to extract diverse spatial features better and establish higher-order relationships between different regions across frames through motion-guided cross-frame message passing.

Uni-perceiver [18] attempts to create a unified model architecture that can flexibly handle multiple modalities of data without training separate models for each modality or task.

Fu Yan et al. [10] proposed a video description method based on the syntactic analysis of object features in scene representation. This method utilizes an object feature detector and constructs a grammar to generate textual descriptions.

Swin-BERT [19] successfully adapted the Swin Transformer [20] to the video description field and achieved promising results. However, it needed to address the significant semantic gap between video and text representations, leading to the protracted training process and failing to yield satisfactory results.

GIT [21] also attempted to model images, videos, and text using a unified network. However, video content merely relied on concatenating video frames without adequately learning the temporal features.

The evaluation metrics used are Bleu-4, METEOR, ROUGE-L, and CIDEr.

The performance of MFFCL on the MSVD dataset is presented in Table ITABLE I. In this work, we introduce a contrastive learning method in addition to Swin-BERT [16], which enables semantic alignment between videos and text. This method is easier to train than text generation tasks and has achieved promising results. Our approach outperforms Swin-BERT with 17%, 6%, and 18% improvements in Bleu-4, ROUGE-L, and CIDEr scores, respectively.

Unlike the MSVD dataset, the MSR-VTT dataset contains a richer set of scene information. TimeSformer is trained on human action recognition datasets. When the video has a lot of non-human behavior information, such as camera movements or birds flying, these details may act as noise for the MFFCL model, potentially leading to a degradation in performance. However, TimeSformer not only extracts features along the temporal dimension but also thoroughly learns video representations in the spatial dimension. Consequently, even in complex video scenes, MFFCL still achieves commendable results. In the work by Wang [22] et al. LSTM were employed for text generation. A limitation of this approach is that the length and coherence of the generated text are constrained. In contrast, our work utilizes GPT2 as the text generation model. During the text generation process, we randomly select from a set of p high-probability words, which allows for the generation of diverse and coherent texts. Compared to the method proposed by Wang et al., our approach demonstrates improvements of 4%, 3%, 9%, and 5% in Bleu-4, METEOR,

TABLE I. PERFORMANCE OF MFFCL ON THE MSVD DATASET

Model	Year	Bleu-4	METEOR	ROUGE-L	CIDEr
MA-Net[6]	2021	50.3	33.4	70.7	78.3
MGRMP[17]	2021	55.8	36.9	74.5	98.5
Uni-perceiver[18]	2022	56.7	38.7	70	88.2
Swin-BERT[19]	2022	58.2	41.3	77.5	120.6
FU et.al.[10]	2023	53.5	-	-	83.1
MFFCL	**2024**	**68.4**	**40.2**	**82.6**	**142.3**

TABLE II. PERFORMANCE OF MFFCL ON THE MSR-VTT DATASET

Model	Year	Bleu-4	METEOR	ROUGE-L	CIDEr
MGRMP[17]	2021	41.7	28.9	62.1	51.4
MA-Net[6]	2021	40.5	27.9	60.3	50.6
Swin-BERT[19]	2022	42.8	29.3	61.7	52.9
FU et al.[10]	2023	43.2	-	-	51.3
Wang et.al.[22]	2023	44.8	29.4	63.0	52.3
MFFCL	2023	**46.6**	**30.3**	**68.6**	**54.8**

TABLE III. PERFORMANCE OF MFFCL ON THE VATEX DATASET

Model	Year	Bleu-4	METEOR	ROUGE-L	CIDEr
GIT[21]	2021	41.6	28.1	55.4	91.5
Swin-BERT[19]	2022	38.7	26.2	53.2	73.0
MFFCL	2024	**50.2**	**35.3**	**65.6**	**100.2**

ROUGE-L, and CIDEr scores, respectively. The specific data are shown in Table II.

While constructing the VATEX dataset, the authors extensively reused videos from the Kinetics-600 dataset[7], resulting in a rich presence of human actions. TimeSformer can fully model spatiotemporal features and accurately recognize action information, thus achieving significant results on VATEX. As mentioned earlier, Swin-BERT lacks the contrastive learning module used in this paper, which may lead to insufficient training. The GIT model, which is not specialized for video data, is less effective in temporal feature extraction than ours. Compared to GIT, our approach shows improvements of 20%, 25%, 18%, and 9% in Bleu-4, METEOR, ROUGE-L, and CIDEr scores, respectively. The specific data are given in Table III.

D. Ablation Experiments

To validate the effectiveness of the modules, this paper conducts ablation studies on the model from three aspects: the contrastive learning module, the mapping module, and the prompt length.

In the learning process of the contrastive learning module, the TimeSformer was fine-tuned using only textual descriptions. During the experiment, it was observed that the improvement in evaluation metrics was very slow. Even doubling the training time on the MSVD dataset did not yield satisfactory results. This could be due to the fact that the text generation task, which relies on contrastive learning, requires more advanced GPUs, and our experimental setup may not meet the training requirements.

TABLE IV. ABLATION EXPERIMENTS WITH CONTRASTIVE LEARNING

Contrastive Learning	Prompt Length	Bleu-4	METEOR	ROUGE-L	CIDEr
√	10	32.6	19.6	39.3	52.6
√	30	60.8	36.6	42.1	82.3
√	50	68.4	40.2	82.6	142.3
×	30	52.3	34.1	69.8	80.3
×	50	52.3	34.1	69.8	80.3

The results of the ablation study are presented in Table IV, the contrastive learning module structure is shown in Fig. 4. The model structure without the contrastive learning module is shown in Fig. 3.

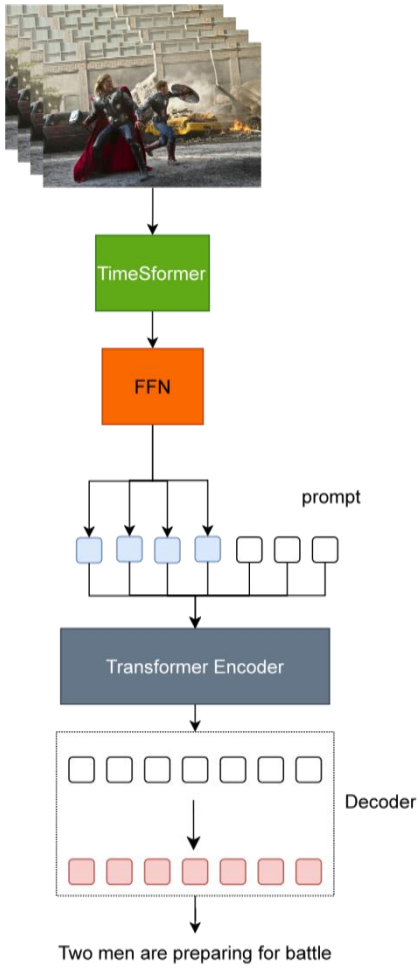

Fig. 3. Overall structure diagram without contrastive learning.

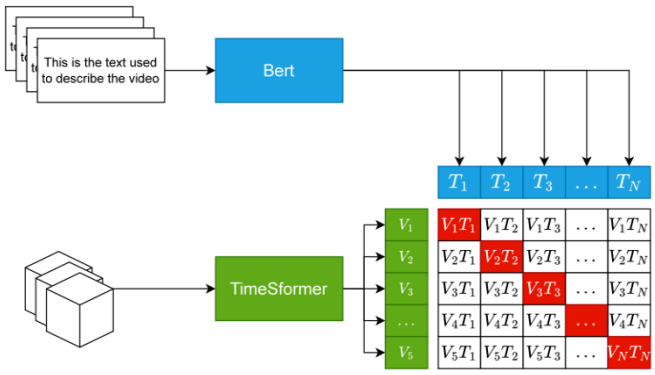

Fig. 4. Contrastive learning module structure.

Regarding the prompt length, it was noted that as the prompt length increased, the amount of information the model could accommodate increased, thereby enhancing its descriptive capabilities. It is believed that longer prompts can fully extract video representation information, which helps the model to understand and describe video content more accurately. However, excessively long prompts may complicate the training process and be limited by device performance and computational resources in practical applications. Experiments have demonstrated that the model can learn the correspondence between videos and texts through contrastive learning, bringing the representations of videos and texts closer in the feature space. This alignment aids the model in better understanding the video content and generates accurate descriptions based on the video embeddings.

To investigate the importance of the understanding mapping module, this paper replaces the Transformer Encoder with a Feedforward neural networks (FFN) to observe changes in the model's performance. Global Linguistic Evaluation Understudy (GLEU) was chosen as the activation function for this setup. Since feedforward neural networks do not contain attention mechanisms, the query vector is removed from the input, and the video block sequence is fed directly into the FFN.

The experimental results indicate that the model's performance decreases when the query vector is absent. This suggests that the query vector has learned a proficient video representation, which aids the model in focusing on the most relevant parts of the video and generates more accurate and coherent text descriptions. Furthermore, experimental results indicate that although feedforward neural networks can learn the mapping between video and text to some extent, their performance does not match that of the Transformer Encoder. The experimental results are presented in Table V.

TABLE V. ABLATION EXPERIMENTS OF THE MAPPING MODULE

Mapping Module	Bleu-4	METEOR	ROUGE-L	CIDEr
Transformer Encode	68.4	40.2	82.6	142.3
FFN	52.3	34.1	69.8	80.3

E. Qualitative Analysis

In this paper, a qualitative analysis of TimeSformer was conducted from the perspective of attention weights. During the execution of TimeSformer, the spatial attention weights were obtained and visualized after being weighted with the original images. The results are shown in Fig. 5. It can be observed that TimeSformer effectively extracts the spatiotemporal representation of the video and adjusts its focus to be similar to human visual attention through the two-stage learning tasks of video-text learning and text generation learning.

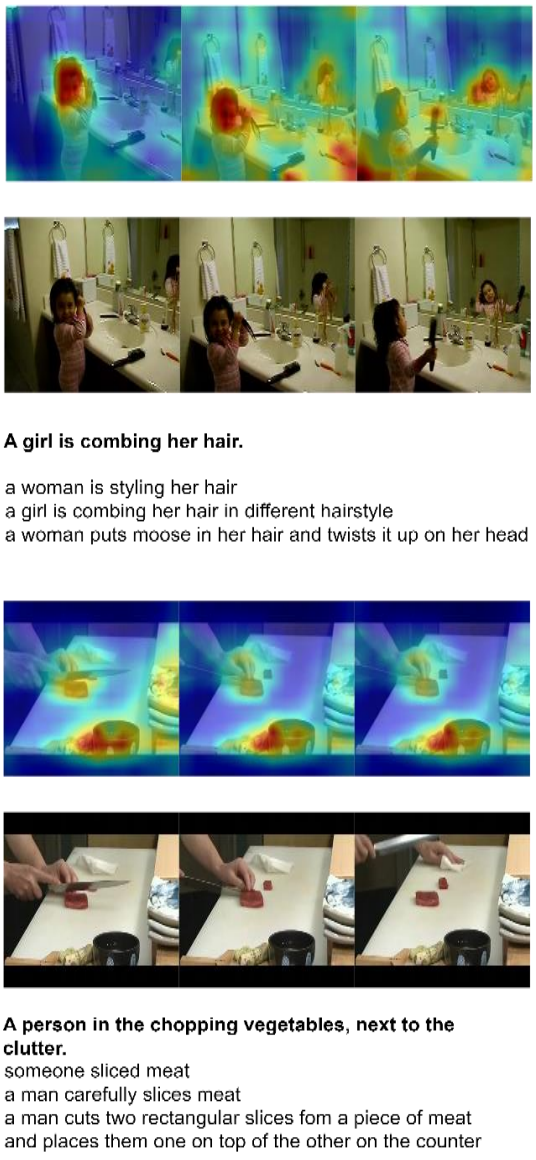

Fig. 5. The image illustrates the ablation results from the MSVD dataset. In each example, the top image represents a schematic of the attention weights, with darker colors indicate higher attention weights. The bottom image shows the original image. The bold part in the text corresponds to the model's output, while the non-bold part represents the dataset labels.

V. CONCLUSION

In the field of multimodal video description, effectively integrating temporal sequence information, visual imagery, and textual descriptions from videos is a worthwhile area of research. To address this challenge, this paper proposes a novel two-step fusion strategy designed to achieve a more precise and coherent understanding and description of video content through an efficient model architecture and training mechanism.

This paper employs TimeSformer, an advanced spatiotemporal feature extractor, in the first stage. Its unique network design for spatiotemporal feature extraction enables it to capture long-range temporal dependencies while preserving spatial details. In the second phase, the focus is on aligning video representations with text representations through contrastive learning. The core principle of contrastive learning is to minimize the distance between positive samples and maximize the distance between negative samples, thereby fostering similarity between video and text representations in the latent space. This study fine-tunes the TimeSformer through carefully designed contrastive tasks to produce video features that are more similar to textual features, prompt the model to generate more accurate video descriptions.

Compared to the Swin-BERT model on the MSVD dataset, our method achieves substantial improvements of 17%, 6%, and 18% on the critical evaluation metrics Bleu-4, ROUGE-L, and CIDEr, respectively. Experimental results confirm the efficacy of the method presented in this paper. TimeSformer is capable of fully representing video content in both temporal and spatial dimensions. The visual prompt also serve to filter out redundant features, while the contrastive learning module accelerates the training process of the TimeSformer.

Future researchers can focus on developing more advanced multimodal fusion techniques to enhance the model's understanding of context and long-term dependencies. Utilizing large-scale, diverse datasets and weak supervision learning can also be explored. Additionally, researching the field of dense video description are potential avenues for advancement.

ACKNOWLEDGMENT

This research is supported by the Key Technology Research and Demonstration Application of News Intelligent Production (212102210417), Science and Technology Plan of Henan Province in 2021.

REFERENCES

[1] P. Tang and H. Wang, “From Video to Language: A Review of Video Title Generation and Description,” Acta Automatica Sinica, vol. 48, no. 2, pp. 375-397, 2022.

[2] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?,” in ICML, 2021, p. 4.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[4] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv:2010.11929 [cs], Jun. 2021, Accessed: May 17, 2022. [Online]. Available: http://arxiv.org/abs/2010.11929

[5] Y. Lu and S. Chen, “Video Description Algorithm Based on Mixed Training and Semantic Association,” Journal of Huazhong University of Science and Technology (Natural Science Edition), vol. 51, no. 11, pp. 67-74, 2023.

[6] Y. Yan and X. Liu, “Research on Video Description Method Based on Multi-Attention and Semantic Detection,” Master’s Thesis, 10.27623/d.cnki.gzkyu.2021.000731, China University of Mining and Technology, 2021.

[7] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[8] A. Vaswani et al., “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[9] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017. doi: 10.1109/cvpr.2017.106.

[10] Y. Fu, M. Wang, and O. Ye, "Video Description Based on Object Feature Grammar Analysis in Scene Representation," Computer Engineering and Design, vol. 44, no. 2, pp. 488-493, 2023.

[11] J. Ren, Q. Zeng, X. Li, Z. Gong, and F. Liu, "Video Description Method Integrating Semantic Enhancement and Multi-Attention Mechanism," Journal of Nanchang University (Science & Technology Edition), vol. 47, no. 6, pp. 548-555, 2023.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners".

[13] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified Vision-Language Pre-Training for Image Captioning and VQA," Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13041–13049, Jun. 2020, doi: 10.1609/aaai.v34i07.7005.

[14] D. Chen and W. Dolan, "Collecting Highly Parallel Data for Paraphrase Evaluation," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds., Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 190–200. [Online]. Available: https://aclanthology.org/P11-1020

[15] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/

[16] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2019. doi: 10.1109/iccv.2019.00468.

[17] S. Chen and Y.-G. Jiang, "Motion Guided Region Message Passing for Video Captioning," International Conference on Computer Vision. Jan. 2021.

[18] X. Zhu et al., "Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16804–16815.

[19] K. Lin et al., "Swinbert: End-to-end transformers with sparse attention for video captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17949–17958.

[20] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.," international conference on computer vision, 2021.

[21] J. Wang et al., "GIT: A Generative Image-to-text Transformer for Vision and Language." 2022.

[22] L. Wang and Y. Bai, "Video Description Method Based on Feature Enhancement and Knowledge Supplementation," Journal of Computer Systems & Applications, vol. 32, no. 5, pp. 273-282, 2023.