Vision-Text Bidirectional Collaborative Image Captioning Algorithm

Mei-Qi Li and Zi-Wei Zhou*

***Abstract*—Image captioning is an interdisciplinary research hotspot at the intersection of computer vision and natural language processing, representing a multimodal task that integrates core technologies from both fields. This task requires the use of computer vision techniques to analyze and extract key visual features from images, followed by the application of natural language processing techniques to generate descriptive text that is syntactically and semantically aligned with human cognition. This process poses a significant challenge for computers. Existing models mostly ignore the relative positional information of visual objects and struggle to efficiently capture the complex relationships between visual and textual data. To address these challenges, we propose a vision-to-text bidirectional collaborative image captioning method. This approach extracts both visual features and positional information of objects, allowing the model to better understand the spatial relationships between objects. The CEW word embedding approach encodes textual information more profoundly, enhancing semantic expression and contextual understanding. In the decoding phase, a bidirectional cross-attention mechanism strengthens the interaction between vision and text, leading to improved accuracy in image understanding. The model is trained and tested on the MSCOCO 2014 dataset and compared with several popular models. Experimental results demonstrate that the proposed method achieves significant improvements on the CIDEr and BLEU-1 evaluation metrics with an increase of 1.5 and 1.1, respectively. In addition, we conduct ablation experiments, quantitative analysis, and qualitative analysis to comprehensively validate the effectiveness and stability of the proposed algorithm.**

***Index Terms*—Image Captioning, Transformer, Cross Attention, Spatially-aware Embedding, CEW Word Embedding**

I. Introduction

With the rapid development of machine translation technologies in the field of natural language processing, the encoder-decoder framework has been widely applied in the image captioning task. In this framework, the encoder uses convolutional neural networks (CNN) or object detection networks, such as VGG [1], ResNet [2], and Fast R-CNN [3], to extract image feature information. The decoder then utilizes recurrent neural networks (RNN), such as LSTM [4] or GRU [5], to generate descriptive text. This approach can handle more complex image semantics; however, differences in data types between the pre-trained visual feature extractors and the subsequent image captioning task limit the model's performance improvement. Additionally, the process of extracting visual features is time-consuming, making it unsuitable for real-time image captioning applications. To address these challenges, some researchers have proposed Transformer-based models for image captioning. The main advantages of this approach include: first, it ensures consistency in the architecture of the encoder and decoder; second, it allows for simultaneous optimization of parameters in both the encoder and decoder. Although this method demonstrates strong performance, effectively aligning the powerful visual features obtained from pre-trained models with the textual descriptions in the dataset remains a problem worth additional exploration.

In recent years, researchers have proposed using Faster R-CNN [6] technology to extract visual features from images. Unlike previous global feature extraction methods, this approach captures fine-grained object feature information within images, rather than including a large amount of irrelevant information, thereby enabling more accurate descriptions. However, merely using a better visual feature extractor to capture surface information (image features) from images is not sufficient; it is also necessary to extract deeper information (object relationships) within the images. Moreover, most popular models currently use standard Transformer word embedding methods [7], which can meet certain task requirements but lack strong contextual awareness in the generated embeddings and do not effectively facilitate the interaction between visual and textual information. This restriction prevents existing word embedding methods from fully utilizing image features, thereby affecting the accuracy and vividness of the descriptions.

To address the aforementioned issues, this paper proposes a vision-text bidirectional collaborative image captioning method, with the following three main contributions:

1. A Spatially-aware embedding module (SAEM) is proposed to encode visual objects and their corresponding positional information, transforming absolute positional relationships into relative positional relationships between visual objects.

2. A CEW word embedding encoding module is introduced, which combines the subword-level embedding method, WordPiece [8], and applies special encoding to word embeddings. This approach enhances the contextual

Manuscript received September 10, 2024; revised December 17, 2024.

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

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

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

awareness of word embeddings and improves their generalization ability when handling new and unseen words.

3. A dual-layer cross-attention mechanism for vision-text collaboration is incorporated into the decoder. First, the masked attention mechanism is improved by introducing visual information to guide the attention distribution between generated words. Second, the mechanism effectively aligns image and text information during the decoding process, allowing the model to consider the most relevant visual information for each word at every step of generation.

II. Relates Jobs

A. Image captioning

Image captioning plays a crucial role in fields such as assisting visually impaired individuals and human-computer interaction. In recent years, with the continuous development of deep learning and the in-depth research by scholars in the field of image captioning, numerous different methods have emerged. Currently, existing image captioning models can be categorized into three types: template-based methods, retrieval-based methods, and generation-based methods.

Template-based methods [9], [10], [11] were widely used in early image captioning tasks. These methods first create a predefined language template, then use object detectors and attribute detectors to extract features from the image, identifying entities and their associated attributes. Finally, the identified entities and attribute information are populated into the language template to form a textual description of the image. While this method can effectively identify and describe entities and their attributes in an image, it tends to generate monotonous descriptions due to the use of fixed templates and may introduce grammatical errors.

Retrieval-based image captioning methods [12], [13], [14] primarily focus on constructing and maintaining a large corpus containing a variety of image descriptions. This approach filters a set of candidate descriptions by comparing the similarity between the input image and the descriptions in the corpus. Ultimately, the description with the highest similarity to the image is selected as the final description. Although this method can enrich the textual description of an image and provide relatively diverse options, one of its main limitations is its inability to create new descriptions, thus failing to fundamentally address the issue of diversity in descriptions.

With the rise of deep learning, Vinyals et al. [15] proposed a generation-based method using a deep recurrent architecture that combines Convolutional Neural Networks (CNNs) from computer vision with Long Short-Term Memory (LSTM) Networks from machine translation. This method generates diverse image descriptions in an end-to-end manner, fundamentally solving the issue of generating fixed-pattern descriptive sentences inherent in template-based methods. The deep neural network-based architecture does not rely on predefined textual rules, allowing for the generation of grammatically flexible image descriptions.

B. Transformer

The Transformer model based on the encoder-decoder architecture was first introd uced by Google in 2017 [7], quickly becoming one of the core technologies in the field of natural language processing (NLP). The key innovation of the Transformer lies in its self-attention mechanism, which enables the model to process input data in parallel and capture the relationships between different parts of the data, without relying on traditional convolutional neural networks (CNNs) or recurrent neural networks (RNNs). This parallel processing approach significantly improves the model's training efficiency and performance, allowing the Transformer to achieve remarkable results in tasks such as machine translation, text generation, and sentiment analysis. However, despite its remarkable success in NLP, the application of Transformer models in computer vision did not make breakthrough progress. It was not until the introduction of the Vision Transformer (ViT) [16] that Transformer-based models for visual tasks gained renewed attention and demonstrated superior performance in certain tasks compared to traditional CNNs.

To make the encoder and decoder structure of the Transformer more suitable for image captioning tasks, researchers have proposed various optimization methods. In 2022, Yang et al. [17] enhanced the model's ability to describe image semantics by jointly modeling intra-modal and inter-modal attention, allowing attention layers to stack more profoundly. In 2023, Song et al. [18] proposed a bidirectional synergistic Transformer model that effectively aligns regional features with grid features to obtain more representative visual-semantic features. In the same year, Heng et al. [19] designed two types of encoders to separately encode object features and relational features in images. By concatenating these two encoded features, the model achieves a fusion of relational features and local object features within the image.

Long-term practice and research have demonstrated that Transformer models perform exceptionally well in handling visual tasks, significantly outperforming traditional Convolutional Neural Network (CNN) models. Moreover, the scalability and flexibility of the Transformer model allow it to perform better on large-scale image datasets, adapting to diverse and complex visual task requirements. As a result, Transformers are gradually becoming the mainstream models in the field of computer vision.

III. Model Design

The proposed model consists of four modules: (1) Spatially-aware Embedding Module (SAEM); (2) CEW Word Embedding Module; (3) Standard Transformer Encoder Module; (4) Bidirectional Collaborative Cross Visual-Text Decoder Module. The overall structure of the model is shown in Figure 1. The input image first undergoes visual and positional feature extraction using Faster R-CNN. Subsequently, the extracted features are processed by the SAEM to obtain spatially-enhanced embedding features, which are then fed into the Transformer encoder for encoding. The CEW word embedding module processes the textual information, and the text cross-masking attention layer applies masking operations to it. The decoder then uses the embedding features output by the encoder and the word generated at the previous time step to interactively attend to both textual and visual information at each time step to predict the next word.

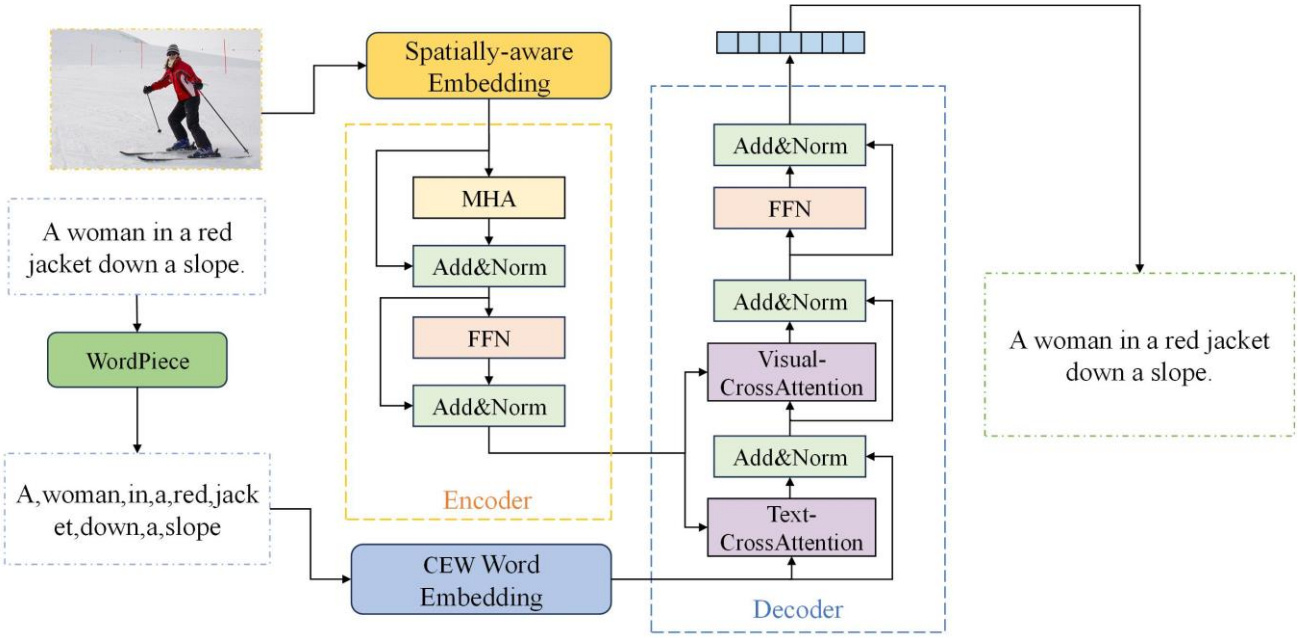

Fig. 1. Text-visual bidirectional collaboration image description model

A. Spatially-aware Embedding Module

Previous approaches used Faster R-CNN for object detection in images, directly feeding the extracted visual features into the encoding layer of the Transformer. However, this approach is not effective in helping the model recognize spatial relationships between objects. Therefore, ResNet-101 is used as the backbone network of Faster R-CNN to extract the visual features of the target objects and a region proposal network is used to extract the positional features of these objects. These two types of features are then fused to generate spatially aware embedding features, as illustrated in Figure 2. This fusion transforms absolute positional information into relative positional relationships between objects, enhancing the model's understanding of spatial relationships between objects.

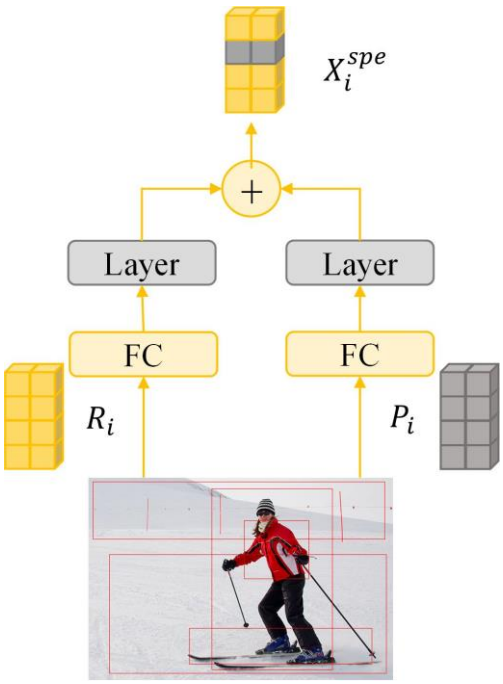

Fig. 2. Spatially-aware embedding encoding architecture

Specifically, for an input image I, Faster R-CNN detects m objects, denoted by $o_i = \{o_1, o_2, \cdots, o_m\}$. Each object is characterized by its region of interest feature R_i and its positional feature P_i (i.e., the bounding box coordinates). The width and height of the image are W and H, respectively, and the bounding box coordinates are given by $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ (representing the x and y coordinates of the top-left corner and the x and y coordinates of the bottom-right corner of the bounding box, respectively). The coordinates are transformed into normalized position eigenvectors using an absolute coordinate normalization function. The formula for this calculation is:

$$P_i = PositionEncode(x_{\min}, y_{\min}, x_{\max}, y_{\max}) = (\frac{x_{\min}}{W}, \frac{y_{\min}}{H}, \frac{x_{\max}}{W}, \frac{y_{\max}}{H}) \quad (1)$$

Before fusing the visual feature vector R_i and the positional feature vector P_i, a fully connected operation is first applied to each of the two features to map them to the same dimension, followed by normalization. The formula for the calculation is as follows:

$$R_i' = FC(R_i) \quad (2)$$

$$P_i' = FC(P_i) \quad (3)$$

$$\hat{R}_i = LayerNorm(R_i') \quad (4)$$

$$\hat{P}_i = LayerNorm(P_i') \quad (5)$$

Subsequently, the processed region features $\hat{R}_i$ and positional features $\hat{P}_i$ are concatenated using a *concat* operation to obtain the spatially-aware embedding features X_i^{spe}. The formula is as follows:

$$X_i^{spe} = Encoder(concat(\hat{R}_i, \hat{P}_i)) \quad (6)$$

B. CEW Word Embedding Module

The difference between the CEW word embedding encoder and traditional word embedding methods lies in its ability to generate a context-based feature embedding matrix by considering contextual information, rather than just encoding individual words. The CEW word embedding module is shown in Figure 3.

Before decoding, the descriptive sentences are first processed by adding a start token <s> and an end token <e> to each sentence. Following the approach of Tan et al. [20], the WordPiece tokenizer is used to tokenize the sentences, assigning an index to each word based on its order in the sentence. The tokenized words are represented as $w_j = \{w_1, w_2, \ldots, w_n\}$. Finally, each word w_j and its positional index j are converted into embedding vectors through an embedding sublayer and normalized using $LayerNorm$ normalization to obtain the word embedding features X_j^{word}. The specific formula is as follows:

$$\hat{w}_j = WordEmbed(w_j) \tag{7}$$

$$\hat{j} = IndexEmbed(j) \tag{8}$$

$$X_j^{word} = LayerNorm(\hat{w}_j + \hat{j}) \tag{9}$$

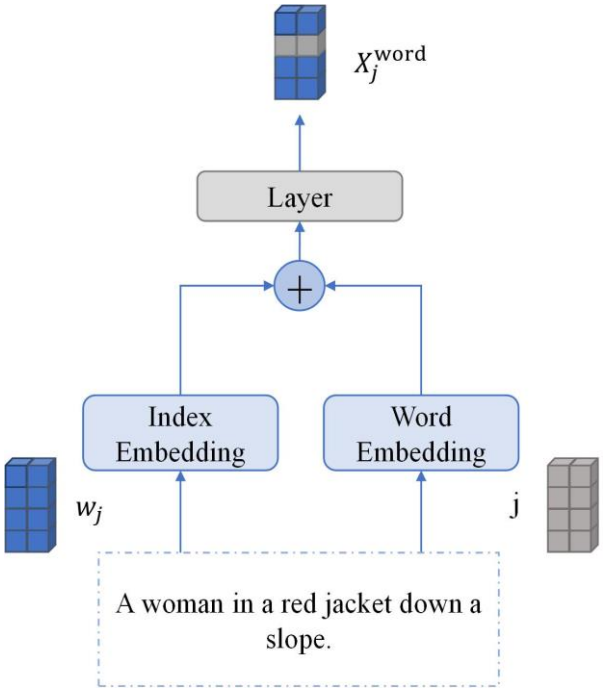

Fig. 3. CEW word embedding architecture

C. Visual-Text Bidirectional Collaborative Decoder

1) *Text Cross-Attention Layer:* During the training process, the generated word embedding features undergo masking. The masked multi-head attention mechanism ensures that the model does not rely on future information, thereby maintaining consistency during description generation. In previous studies, the Masked Self-Attention module primarily focused on the interaction of information between words and did not fully leverage the influence of visual information on the calculation of intra-modal attention weights. To allow visual features to more effectively guide the generation of descriptive sentences, this paper introduces a cross-attention sublayer from vision to text. This sublayer integrates visual information to optimize the attention distribution between words within the sequence, enabling the visual features to have a significant impact on the decoding process and improving the overall quality and semantic relevance of the generated descriptions. A diagram of the text cross-attention layer structure is shown in Figure 4.

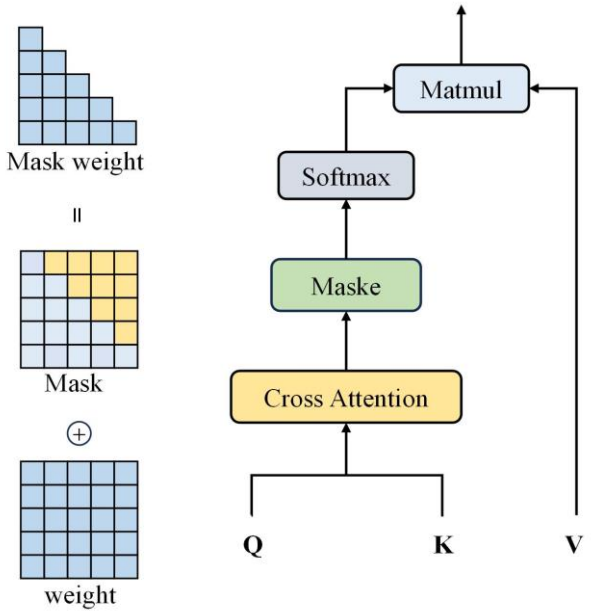

Fig. 4. Text cross-attention layer

The word embedding sequence $X^{word} = \{X_1^{word}, X_2^{word}, \ldots, X_n^{word}\}$ is used as the query vector Q and the value vector V in the text cross-attention layer, while the encoded spatially-aware embedding sequence $X^{spe} = \{X_1^{spe}, X_2^{spe}, \cdots, X_m^{spe}\}$ is used as the key vector K. First, a dot product is computed between Q and K to obtain the attention weights A_{v2l}. Next, a mask is applied to A_{v2l} so that it only relies on the word sequence information generated before time step t. Finally, the output of the $softmax$ layer is applied to V to obtain the visually sensitive word embedding feature sequence X^{word*}. The specific calculation process is as follows:

$$[Q, K, V] = \left[W^Q X^{word}, W^K X^{spe}, W^V X^{word}\right] \tag{10}$$

$$A_{v2l} = \frac{QK^T}{\sqrt{d}} \tag{11}$$

$$\hat{A}_{v2l} = Mask(A_{v2l}) \tag{12}$$

$$X^{word*} = softmax(\hat{A}_{v2l})V \tag{13}$$

Where W^Q, W^K and W^V are the weight matrices of the linear transformations, and $Mask$ represents the masking operation.

2) *Visual Cross-Attention Layer:* Word prediction relies on the fusion of multimodal information. In a standard Transformer decoder, a multi-head attention module is used to unify information from different modalities, achieving inter-modal interaction and fusion through the attention mechanism, aligning textual information with image features. However, cross-attention allows the decoder to access the

entire output of the encoder while generating each word, enabling the generated descriptions to more accurately reflect the details and spatial layout in the image. Therefore, a visual cross-attention layer is introduced in the Transformer decoder to facilitate the interaction between visual and textual information.

The word embedding sequence output from the text cross-attention layer X^{word*} is used as the query Q, while the spatially-aware embedding sequence output from the encoder X^{spe} is used as the key K and value V. This setup enables attention weighting between the visual and textual modalities, resulting in an attention-weighted matrix for text descriptions guided by image information. The calculation formula is as follows:

$$[Q,K,V]=\left[W^{Q}X^{word*},W^{K}X^{spe},W^{V}X^{spe}\right] \quad (14)$$

$$CrossAttention(Q,K,V)=softmax(\frac{QK^{T}}{\sqrt{d}}) \quad (15)$$

Where W^{Q}, W^{K} and W^{V} are the weight matrices of the linear transformations. To address the problem of vanishing gradients as the network depth increases, residual connections and normalization operations are added between the visual cross-attention sub-layer and the Feed Forward Network (FFN). Finally, the output of the decoder is passed through a $softmax$ layer to compute the word generated at the current time step.

IV. Experiments and Results Analysis

A. Datasets and Experimental Environment

The MSCOCO 2014 dataset [21] is used for experiments in this paper. This dataset is a commonly used large-scale English-annotated dataset for image captioning tasks, containing 123,287 images, each with five manually annotated sentences.

Before training, the dataset is preprocessed using a random partitioning method to handle the training and validation sets of the MSCOCO 2014 dataset. The dataset is shuffled, and 5,000 images are randomly selected as the validation set, another 5,000 images are randomly selected as the test set, and the remaining 113,287 images are used for training the model. This ensures the diversity and randomness of the dataset. The commonly used evaluation metrics BLEU (1-4) [22], METEOR [23], ROUGE-L [24], and CIDEr [25] are employed to assess the effectiveness and advancement of the model.

The experimental environment for this research is built on the Linux operating system, Ubuntu version 20.04, using PyTorch as the main deep learning framework for conducting experiments. The hardware configuration includes an Intel(R) Core(TM) i9-13900KF CPU and a Geforce RTX 4060ti GPU (16GB VRAM, 32GB RAM).

B. Experimental Parameters

The features extracted by ResNet-101 are used as inputs to Faster R-CNN, and the Region Proposal Network (RPN) generates bounding boxes. The anchor sizes are set to {128, 256, 512}, and the aspect ratios are set to {0.5, 1.0, 2.0}. Overlapping bounding boxes with an Intersection over Union (IoU) threshold greater than 0.7 are discarded. The word vector dimension and the hidden layer dimension are both set to 1024. The number of heads in the multi-head attention mechanism is 8, with each head having a feature dimension of 128. Both the encoder and decoder consist of 6 layers to balance model complexity and performance.

The training process runs for 40 epochs. In the first 20 epochs, the cross-entropy loss function is used for training with a batch size of 20. In the latter 20 epochs, the CIDEr-D optimized reinforcement learning method is employed, with a batch size of 10. The learning rate follows a scheduling strategy with a warm-up phase, starting at an initial value of 1e-4, and the Adam optimizer is used ($\beta1 = 0.9$, $\beta2 = 0.999$, $\varepsilon = 1e\text{-}8$). To further improve the model's generalization ability, Dropout is applied to the attention layers and fully connected layers, with the Dropout rate set to 0.1. Beam search is used during model training, with the beam size set to 5, and the hyperparameter k in relative position encoding is set to 8.

C. Experimental Results and Analysis

1) *Ablation Experiment:* This paper conducts ablation experiments to validate the rationality and effectiveness of the model, using a control variable method to demonstrate the impact of each module on the overall model. The experimental results are shown in Table I. The experiment uses the standard Transformer structure as the baseline model, where SAE represents the introduction of the Spatially-aware Embedding Module, CEW represents the introduction of the CEW Word Embedding Module, V2TC represents the introduction of the Visual-Text Bidirectional Collaborative Decoder, and SCV2TC represents the complete model proposed in this paper. By comparing the performance of each model, it is clear that each module contributes to the improvement of the model, and the combination of all three modules in the SCV2TC model achieves the best performance, demonstrating that these modules work synergistically to achieve the maximum performance improvement.

TABLE I
Comparison of Ablation Experiment Results

Methods	B1	B2	B3	B4	M	R	C
Baseline	76.0	60.0	46.5	35.4	27.8	56.8	119.6
SAE	78.2	62.3	48.8	38.6	27.8	56.9	124.9
CEW	76.5	62.8	47.6	37.2	28.6	58.0	122.3
V2T	78.7	64.2	51.6	38.7	29.0	58.5	126.4
SCV2TC	82.8	67.1	52.9	40.9	30.8	59.9	136.8

As shown in Table I, both the SAE model and the CEW model are compared with the baseline model, with multiple metrics (such as BLEU, METEOR, ROUGE-L, etc.) showing varying degrees of improvement. In particular, the CIDEr (C) score shows a significant increase, highlighting the key role of the Spatially-aware Embedding Module (SAE) in helping the model better understand the spatial structure of the image and the relative positioning of objects. The CEW Word Embedding Module (CEW) enhances the semantic representation of words, making the generated text more contextually relevant and coherent. The V2TC model, which introduces a bidirectional cross-attention mechanism, further enhances the complementarity between image and text, allowing the model to more precisely capture the subtle

relationships between visual information and language. This improvement is reflected in the performance of all metrics, with the CIDEr (C) score improving by as much as 6.8%. This result shows that the bidirectional cross-attention mechanism effectively combines visual and textual information, enhancing the accuracy and contextual consistency of the generated text. Further comparison between the SCV2TC model and the baseline model reveals that the introduction of SCV2TC significantly improves the model's ability to understand and describe complex scenes. Specifically, SCV2TC combines the Spatially-aware Embedding, CEW Word Embedding, and Visual-Text Bidirectional Collaborative Decoder in a well-integrated manner, improving performance across various metrics while significantly enhancing the richness and accuracy of the generated descriptions. This demonstrates that the combination of SAE, CEW, and V2TC modules can mutually reinforce each other, leveraging their individual advantages to help the model better understand relationships between visual objects and more delicately and precisely capture relationships between different regions of the image. This results in feature enhancement, which improves the quality and accuracy of the generated descriptions. The experimental results validate the superiority and effectiveness of the proposed SCV2TC model in image captioning tasks.

2) *Quantitative Analysis:* This paper validates the effectiveness of the model through quantitative analysis, comparing it with representative mainstream algorithms such as Soft-Attention [26], Hard-Attention [26], MSM [27], ELMo-MCT [18], Up-Down [6], ORT [28], AOA [29], DLCT [30], and X-Transformer [31] on the MS COCO 2014 dataset. As shown in Table II, the proposed SCV2TC model achieves favorable results across various evaluation metrics, with BLUE-1 (B1), BLUE-2 (B2), BLUE-3 (B3), and BLUE-4 (B4) reaching 82.8%, 67.1%, 52.9%, and 40.9%, respectively. Additionally, METEOR (M), ROUGE-L (R), and CIDEr (C) reach 30.8%, 59.9%, and 136.8%, respectively.

Models like Soft-Attention and Hard-Attention are based on attention mechanisms, generating descriptions by calculating the attention of each word to image regions. In contrast, SCV2TC not only relies on a single attention mechanism but also introduces a Bidirectional Visual-Text Collaborative Decoder (V2TC), enabling bidirectional information flow between the image and text, thereby more accurately capturing multimodal relationships in complex scenes. The MSM model uses multi-scale learning to capture features at different levels in the image, while SCV2TC further enhances spatial understanding through Spatially-aware Embedding (SAE), improving the precision of description generation. ELMo-MCT focuses on text processing by incorporating context-aware word embeddings and multimodal contrastive learning to understand linguistic context, while SCV2TC strengthens the complementarity between vision and text through the V2TC model, improving performance in image captioning tasks. The Up-Down model uses a hierarchical structure to jointly model image and text features, while SCV2TC further refines visual-text collaboration, especially in complex scenes, to precisely handle spatial relationships between objects. The ORT model focuses on object and action recognition, while SCV2TC enhances spatial relationship modeling through a bidirectional cross-attention mechanism, capturing more detailed image features and generating more natural, fluent descriptions. The X-Transformer model incorporates the Transformer architecture for image captioning, and SCV2TC further optimizes this by introducing bidirectional visual-text interactions, generating more accurate text descriptions that align with visual information, particularly in complex scenes with multiple objects.

As can be seen, the proposed Spatially-aware Embedding module, CEW Word Embedding module, and Visual-Text Bidirectional Collaborative Decoder demonstrate significant advantages over traditional image captioning methods. First, the Spatially-aware Embedding module captures spatial information in the image accurately, allowing the model to better understand the spatial relationships between objects when processing complex scenes, providing a solid foundation for generating more natural and realistic text descriptions. Second, the CEW Word Embedding module optimizes the representation of words by incorporating contextual information, further enhancing the model's understanding of word meanings in different contexts, resulting in more semantically and syntactically accurate and fluent generated text. Finally, the Visual-Text Bidirectional Collaborative Decoder effectively integrates bidirectional information flow between vision and language, not only enhancing the complementarity between the image and the text but also ensuring more contextually consistent text generation. The organic combination of these three modules fully leverages the synergistic effect of visual and textual information, significantly improving the quality and effectiveness of image caption generation.

3) *Qualitative Analysis:* To qualitatively analyze the visual features extracted by the model, this study employs visualization techniques, such as heatmaps, to illustrate the importance of visual features in generating output words. This visualization reveals how different visual inputs influence the text generation process, providing deeper insights into the model's decision-making mechanism. Specifically, we visualize the visual features by utilizing the cross-attention weights in the final layer of the Transformer decoder, which helps clarify the relationship between each generated word and the key regions of the image. This method not only highlights the role of different parts of the image during the generation process but also significantly contributes to the transparency and interpretability of the model.

The detailed visualization results are presented in Figure 5, which includes the original image for each test sample, the words generated by the Baseline model, and those generated by the proposed SCV2TC model at various time steps, along with their corresponding attention heatmaps. As shown in Figure 5, both the SCV2TC and Baseline models are able to focus on image regions relevant to the generated words. However, compared to the Baseline model, the SCV2TC model is better at precisely focusing on the relevant visual regions when generating descriptions, especially in complex or detail-rich scenes. Specifically, the SCV2TC model demonstrates significant advantages in terms of attention to object content, detailed descriptions, and overall contextual coherence. It is better equipped to handle the spatial

relationships between different objects in the image and capture finer details, resulting in more accurate and natural text descriptions. This indicates that the SCV2TC model has higher efficacy and stronger contextual awareness when understanding and describing complex scenes.

To provide a more intuitive comparison of the experimental results, this study selects several results generated by both the SCV2TC model and the Baseline model, comparing them with five human-annotated reference sentences from the dataset, as shown in Figure 6. In these comparison results, the SCV2TC model more accurately captures key details in the image and the relationships between objects in the generated text descriptions. For example, in some complex scenes, the SCV2TC model not only correctly identifies the objects but also accurately describes details such as the relative position, color, and actions of the objects. In contrast, the Baseline model may omit or misidentify certain aspects. Comparing the generated text with the human-annotated sentences, the SCV2TC model significantly outperforms the Baseline model in terms of accuracy, naturalness, and contextual consistency, further validating the effectiveness and advantages of the SCV2TC model in multimodal understanding and text generation tasks.

It can be seen that the descriptions generated by the baseline model are logically correct and have some association with the image content. However, the results produced by the proposed SCV2TC model can better capture the spatial relationships between visual objects and provide more accurate and vivid descriptions of image details. For example, in the first example, the baseline model generates "running on the road," which does not align well with the visual information in the image. In contrast, the description generated by the proposed model, "lying on the side of the road with a bicycle parked next to it," more accurately reflects the visual information in the image, vividly describing the relationships between objects within the image. Through comparative analysis, it is evident that the SCV2TC model is better at handling complex image scenes, especially those involving multiple objects and complex backgrounds. It can better understand and express the interactions and spatial layouts between objects, thereby generating descriptive texts that are more consistent with the actual situation.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS OF DIFFERENT MODELS

Methods	B1	B2	B3	B4	M	R	C
Soft-Atten	70.7	49.2	34.4	24.3	23.9	—	—
Hard-Atten	71.8	50.4	35,7	25.0	23.0	—	—
MSM	73.0	56.5	42.9	32.5	25.1	53.8	98.6
ELMo-MCT	76.2	59.6	45.6	34.2	—	56.0	111.5
Up-Down	79.8	—	—	36.3	27.7	56.9	120.4
ORT	80.5	—	—	38.6	28.7	58.4	128.3
AOA	81.0	65.8	51.4	39.4	29.1	58.9	126.9
DLCT	81.4	—	—	39.8	29.5	58.9	133.8
X-Trans	81.7	66.8	52.6	40.7	29.9	59.7	135.3
SCV2TC	**82.8**	**67.1**	**52.9**	**40.9**	**30.8**	**59.9**	**136.8**

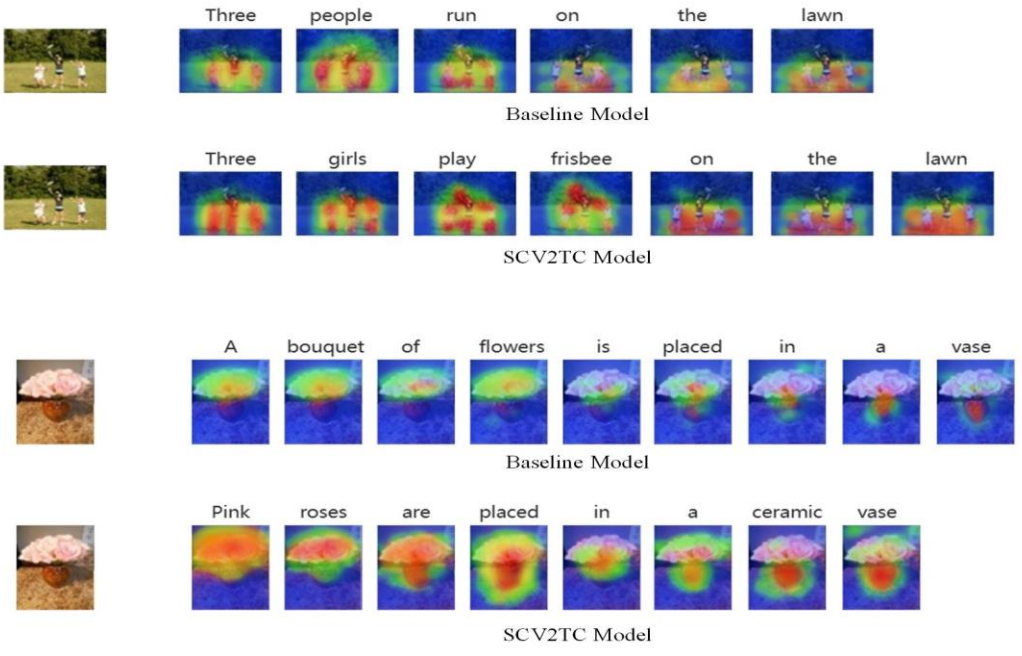

Fig. 5. Visualization of visual attention

Manual Captions1: A picture of a dog laying on the ground.
Manual Captions2: Dog snoozing by a bike on the edge of a cobblestone street.
Manual Captions3: The white dog lays next to the bicycle on the sidewalk.
Manual Captions4: a white dog is sleeping on a street and a bicycle.
Manual Captions5: A puppy rests on the street next to a bicycle.
Baseline Model: A dog running on the road.
SCV2TC Model: A white dog is resting next to a bicycle on the sidewalk.

Manual Captions1: This table is filled with a variety of different.
Manual Captions2: A variety of food in dishes is displayed on a table.
Manual Captions3: Table of food including carrots, peas, salad, corn, gravy, pies, bread.
Manual Captions4: A table layed out with food such as, salad, steamed peas and carrots, steamed corn, and bread rolls.
Manual Captions5: A table full of food such as peas and carrots, bread, salad and gravy.
Baseline Model: Some fruits and vegetables are in the bowl.
SCV2TC Model: A puppy rests on the street next to a bicycle

Manual Captions1: A man using a cell phone amongst a crowd of people.
Manual Captions2: A man talking on his phone in the public.
Manual Captions3: A man in a crown wears a red jacket and is on the phone.
Manual Captions4: A man standing on a busy sidewalk while talking on his cellphone.
Manual Captions5: a man with a white beard and hat on a cellphone.
Baseline Model: A man dressed in red is talking with others.
SCV2TC Model: An elderly man wearing a red jacket and a black hat is talking on the phone.

Fig. 6. Comparison of image description results

V. Conclusion

This paper proposes a Transformer-based visual-text bidirectional collaborative image captioning model, aimed at improving the quality and diversity of generated image captions. The model innovatively integrates spatial perception feature embeddings and CEW word embedding, which play a crucial role in image captioning. Specifically, the spatial perception feature embeddings are used to extract deep spatial relationships between different objects in the image, while CEW word embeddings enhance the contextual understanding of words, enabling the generation of more accurate and rich descriptions. By introducing a bidirectional cross-attention mechanism, the model significantly strengthens the interaction and collaboration between visual and textual information, allowing image and text features to complement each other more effectively, ultimately generating higher-quality descriptive sentences.

In terms of experiments, extensive evaluations were conducted on the MS COCO 2014 dataset to validate the effectiveness of the proposed method. Through a series of ablation experiments, the paper further demonstrates the contribution of each module and its role in improving overall performance. Compared to current state-of-the-art image captioning models, the proposed model achieves significant improvements across all evaluation metrics, with the CIDEr score reaching 1.368, indicating superior caption quality. Moreover, the model is able to better capture details and semantic relationships in the image, and the generated captions are more natural, fluent, and accurate than those produced by traditional methods.

For future work, we plan to further optimize the model's encoder by exploring the introduction of a Gated Attention Unit in the encoding process. This innovative design will allow for more flexible adjustment and control of the interaction between the multi-head attention layers and feed-forward neural network layers in the encoder, enabling the model to more accurately allocate attention based on task requirements. Additionally, the introduction of the gating mechanism will help reduce computational complexity and improve training efficiency. To further reduce model complexity and enhance training speed, we will also consider parameter sharing across multiple modules to minimize redundant calculations, thus accelerating both the model's training process and inference speed.

References

[1] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–14.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 26–July 1, 2016, pp. 770–778.

[3] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 7–13, 2015, pp. 1440–1448.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 25–29, 2014, pp. 1724–1734.

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, June 18–22, 2018, pp. 6077–6086.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, Dec. 4–9, 2017, pp. 5998–6008.

[8] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, and Y. Cao, "Google's neural machine translation system: bridging the gap between human and machine translation," *arXiv preprint*, arXiv:1609.08144, 2016.

[9] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, and A. C. Berg, "BabyTalk: Understanding and generating simple image descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, June 20–25, 2011, pp. 1601–1608.

[10] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, et al., "Midge: Generating image descriptions from computer vision detections," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012, pp. 747–756.

[11] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, WA, USA, 2013, pp. 1292–1302.

[12] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "TreeTalk: Composition and compression of trees for image descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 351–362, 2014.

[13] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, et al., "Every picture tells a story: Generating sentences from images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Crete, Greece, Sept. 5–11, 2010, pp. 15–29.

[14] Y. Ushiku, M. Yamaguchi, Y. Mukuta, and T. Harada, "Common subspace for model and similarity: Phrase learning for caption generation from images," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2668–2676.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 7–12, 2015, pp. 3156–3164.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, May 3–7, 2021.

[17] Wen-Rui Yang, Tao Shen, Yan Zhu, Ying-Li Liu, "Image caption with ELMo embedding and multimodal transformer," *Computer Engineering and Applications*, vol. 58, no. 21, pp. 223–231, 2022.

[18] Jing-Kuan Song, Peng-Peng Zeng, Jia-Yang Gu, Jin-Kuan Zhu, and Lian-Li Gao, "End-to-end image captioning via visual region aggregation and dual-level collaboration," *Ruan Jian Xue Bao/Journal of Software*, vol. 34, no. 5, pp. 2152–2169, 2023.

[19] Hong-Jun Heng, Yi-Chen Fan, and Jia-Liang Wang, "Multifaceted feature coding image caption generation algorithm based on Transformer," *Computer Engineering*, vol. 49, no. 2, pp. 199–205, 2023.

[20] Hao Tan, and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 5100–5111.

[21] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, et al., "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, Sept. 6–12, 2014, pp. 740–755.

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

[23] M. Denkowski and A. Lavie, "Meteor Universal: Language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, 2014, pp. 376–380.

[24] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.

[25] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 7–12, 2015, pp. 4566–4575.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, July 6–11, 2015, pp. 2048–2057.

[27] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Oct. 22–29, 2017, pp. 4894–4902.

[28] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, Dec. 8–14, 2019, pp. 11135–11145.

[29] Lun Huang, Wen-Min Wang, Jie Chen, and Xiao-Yong Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, Oct. 27–Nov. 2, 2019, pp. 4633–4642.

[30] Yun-Peng Luo, Jia-Yi Ji, Xiao-Huai Sun, Liu-Jun Cao, Yong-Jian Wu, Fei-Yue Huang, et al., "Dual-level collaborative transformer for image captioning," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, Virtual, Feb. 2–9, 2021, pp. 2286–2293.

[31] Ying-Wei Pan, Ting Yao, Ye-Hao Li, and Tao Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 14–19, 2020, pp. 10971–10980.