Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060788

ARTICLE

A Novelty Framework in Image-Captioning with Visual Attention-Based Refined Visual Features

Alaa Thobhani[1,*], Beiji Zou[1], Xiaoyan Kui[1,*], Amr Abdussalam[2], Muhammad Asim[3], Mohammed ELAffendi[3] and Sajid Shah[3]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2]Electronic Engineering and Information Science Department, University of Science and Technology of China, Hefei, 230026, China
[3]EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia

*Corresponding Authors: Alaa Thobhani. Email: althobhanialaa@gmail.com; Xiaoyan Kui. Email: xykui@csu.edu.cn

Received: 10 November 2024; Accepted: 24 December 2024; Published: 06 March 2025

ABSTRACT: Image captioning, the task of generating descriptive sentences for images, has advanced significantly with the integration of semantic information. However, traditional models still rely on static visual features that do not evolve with the changing linguistic context, which can hinder the ability to form meaningful connections between the image and the generated captions. This limitation often leads to captions that are less accurate or descriptive. In this paper, we propose a novel approach to enhance image captioning by introducing dynamic interactions where visual features continuously adapt to the evolving linguistic context. Our model strengthens the alignment between visual and linguistic elements, resulting in more coherent and contextually appropriate captions. Specifically, we introduce two innovative modules: the Visual Weighting Module (VWM) and the Enhanced Features Attention Module (EFAM). The VWM adjusts visual features using partial attention, enabling dynamic reweighting of the visual inputs, while the EFAM further refines these features to improve their relevance to the generated caption. By continuously adjusting visual features in response to the linguistic context, our model bridges the gap between static visual features and dynamic language generation. We demonstrate the effectiveness of our approach through experiments on the MS-COCO dataset, where our method outperforms state-of-the-art techniques in terms of caption quality and contextual relevance. Our results show that dynamic visual-linguistic alignment significantly enhances image captioning performance.

KEYWORDS: Image-captioning; visual attention; deep learning; visual features

1 Introduction

Image-captioning is a complex process that forms a connection between two fundamental artificial intelligence pillars, computer vision [1–3] and natural language processing. It involves producing detailed textual descriptions for images, requiring a deep comprehension of the visual elements and the ability to articulate them clearly in natural language. Recent developments in deep learning have greatly impacted this area. Convolutional Neural Networks (CNNs) are particularly effective at extracting visual features, allowing systems to thoroughly interpret the contents of an image. On the other hand, Recurrent Neural Networks (RNNs) or their advanced derivatives, such as Long Short-Term Memory (LSTM), are particularly suited to handle sequential data, which makes them an ideal option for generating the flow of text in captions. One of the key developments that have driven progress in image-captioning is integrating attention mechanisms.

Such mechanisms have enabled models to put more emphasis and focus on specific and selective parts or regions of the image during the process of generating their relevant captions. This helps the model produce more accurate and coherent descriptions, with highly contextual relevance by correlating the image's critical visual cues with the corresponding words. Consequently, significant improvements have been shown in the generated captions of the models that incorporated attention mechanisms whether in their accuracy or in the overall results.

Despite the advancements, challenges, such as generating detailed captions that accurately capture complex interactions within an image, remain. The traditional captioning models use fixed visual features, which do not adapt dynamically based on the linguistic context, limiting their ability to form strong, relevant connections between the visual input and the generated captions. This static approach often results in less accurate or descriptive captions. The proposed model addresses this by introducing a dynamic interaction where visual features are influenced by the linguistic context, allowing the visual features to adapt and align with the evolving linguistic content. This strengthens the relationship between the visual and linguistic elements, leading to more appropriate captions in both coherence and contextualization. Contrary to conventional models, in which visual features remain static at each time step, the proposed method dynamically adjusts them using partial attention. This attention generates weights that modify the corresponding local visual features in real time, better aligning them with the linguistic context. These adjusted visual features are then processed by further refining their alignment with the linguistic content. This continual adaptation ensures stronger visual-linguistic alignment, enhancing the relevance of the visual features and generating captions with high overall quality.

In this paper, a novelty in images' caption generation is proposed; called Visual Attention-based Refined Visual Features (VARVF), which enhances image-captioning by aligning visual features more closely with the linguistic context. Specifically, we introduce a method where, at each time step, we dynamically reweight the visual features to create new, context-aware visual representations. These updated visual features are more aligned with the evolving linguistic context, allowing for a more seamless interaction between the image and the language. We introduce two modules: the Visual Weighting Module (VWM), which dynamically adjusts visual features at each time step depending on the linguistic context using partial attention, and the Enhanced Features Attention Module (EFAM), which further refines these features through an additional attention layer. This continuous adaptation strengthens the alignment between visual and linguistic elements, enhancing both the accuracy and relevance of the produced captions. The proposed model leverages this reweighted visual information, using it in subsequent attention mechanisms to generate captions that are both semantically and visually coherent. By continuously adapting the visual features throughout the captioning process, we produce captions that have more contextual relevance and higher image content visual consistency. This contribution bridges the gap between static visual features and dynamic lingual generation, resulting in captions with higher meaningfulness and accuracy.

This study offers several important contributions to the field, which can be summarized in the following key points:

- We explore the enhancement of visual features in image-captioning by making them more closely integrated with the linguistic context.
- One of the features of the proposed model is the reweighing process of the visual features at each time step, dynamically creating new visual representations that are more contextually aligned with the language.
- These updated visual features are used in an attention mechanism to produce coherent captions in terms of visualization and semantics, resulting in more meaningful and connected descriptions.

- We introduce two novel modules: the Visual Weighting Module (VWM), which dynamically adjusts visual features at each step of time, depending on the linguistic context, using partial attention, and the Enhanced Features Attention Module (EFAM), which further refines these features through an additional attention layer. This continuous adaptation strengthens the alignment between visual and linguistic elements, enhancing the generated captions in terms of accuracy and relevance.
- We implemented the suggested novel VARVF framework on the MS-COCO dataset, and the resulting performance metrics demonstrated that our method was competitive with the most advanced techniques currently available, based on the evaluation criteria.

2 Related Work

Throughout the years, various leaps have been made in attention mechanisms and caption generation. One significant development is integrating dynamic visual attention mechanisms into image-captioning, as seen in works like [4,5]. In a related vein, Reference [6] introduced the concept of integrating attention directly into the captioning process, while Reference [7] proposed an adaptive attention mechanism that dynamically adjusts throughout the caption generation process. Further refinements may include a combination of bottom-to-up and top-to-down attention strategies. Reference [8] and memory-enhanced attention mechanisms to improve caption quality [9]. Innovations like dual attention on pyramid image features [10] and cluster-based grounding networks [11] have been employed to further improve the coherence and relevance of the captions. A notable contribution to evaluation is the Proposal Attention Correctness (PAC) metric [11], which bridges the gap between performance assessment and visual grounding. The advent of Transformer-based architectures has profoundly influenced the domain, particularly with the multimodal Transformer model [12], which improves caption generation by leveraging multi-view feature extraction and learning interactions that are sensitive to specific regions [13]. In terms of novel techniques, Reference [14] introduced a method that combines wavelet decomposition with convolutional neural networks to generate captions, while Reference [15] presented the JRAN model, which focuses on feature relationships through the use of both regional and semantic features. Most recently, the Guided Visual Attention (GVA) technique [16] has been introduced to enhance caption generation by dynamically updating and readjusting attentional weights, showcasing how the field continues to evolve with increasingly sophisticated mechanisms that align captions more closely with visual input.

Earlier studies have incorporated a range of information sources, such as visual Features, descriptive tags, and part-of-speech (PoS) topics, along with saliency mechanisms, to examine a variety of methods designed to improve image-captioning. For example, the inclusion of image attributes has been shown to significantly improve caption quality, with some models incorporating multimodal attribute detectors trained in parallel to captioning systems to provide more accurate and highly contextually relevant descriptions [17,18]. PoS information has also been instrumental in guiding the structure of captions, ensuring grammatical coherence and relevance [19,20]. In addition to these linguistic components, topics derived from caption corpora have influenced sentence generation, allowing for better alignment between the visual content and thematic context [21–24]. Furthermore, saliency mechanisms have been used to enhance the representation of images by emphasizing key visual elements based on their importance [25]. Advanced attention mechanisms, such as semantic-guided and text-guided attention, have been developed to create stronger correlations between semantic attributes and image representations, resulting in accurate and contextually suitable captions [26]. Methods like Stack-VS leverage multistage image descriptors, efficiently combining visual and semantic information through attention layers to further improve captioning performance [27]. Additionally, models like FUSECAP [28] enrich captions by integrating visual expert insights, enhancing both captioning accuracy and image retrieval performance. Another approach,

the Face-Att model [29], focuses on generating captions that prioritize facial features, emphasizing the potential of attribute-specific models to produce more detailed and relevant descriptions. These techniques collectively illustrate the advancements in using varied information sources to generate more accurate, and coherent image captions, with highly contextual relevance. Reference [30] introduces two key modules, the Independent Attribute Predictor (IAP) and the Enhanced Attribute Predictor (EAP), which significantly improve fine-grained image-captioning. The IAP focuses on accurately predicting image-related attributes, while the EAP rebalances visual and linguistic attributes to generate more contextually accurate captions. Reference [31] introduces a dynamic Attribute Selector Module (ASM) that selects relevant attributes at each time step based on the visual and the lingual contexts. This ensures that only the most pertinent attributes contribute to generating fine-grained captions. Additionally, the work integrated a combination of attribute information and guided visual attention, assembled in a fusion mechanism.

To effectively capture a wide range of visual content, several image-captioning techniques are specifically designed to produce multiple descriptions for each image. Our proposed approach aims to enhance the representation of complex scenes and provide a richer understanding of the visual content. For instance, one notable method employs conditional Generative Adversarial Networks (GAN) [32] to create diverse captions. This technique involves the simultaneous training of both a generator, which creates the captions, and an evaluator, which assesses the quality and variation of the generated descriptions. By conditioning the generation process, the model can yield a variety of captions that reflect different perspectives or aspects of the same image. Another model suggests a topic-based multi-caption network [33], which generates coherent and relevant captions that are directly related to specific topics. This method works by integrating both an image and its associated topic, allowing the model to produce captions that maintain a consistent thematic focus. By emphasizing topic relevance, this strategy enhances the clarity and utility of the generated captions, making them more informative for users seeking specific insights from the images. Collectively, these strategies address the inherent complexity and diversity present in image content, ensuring that various facets of the visual scene are captured in the captions. Additionally, a recent model [34] further advances this field by considering the number of reference captions available throughout the training process. This innovative approach uses associated numerical data to generate diverse image captions that reflect the quantitative aspect of caption availability. By doing so, this model not only relies on semantic information but also utilizes the richness of multiple captions, achieving a notable progression toward the development of image-captioning techniques and their ability to provide comprehensive and nuanced descriptions of visual content. In recent years, diffusion models have emerged as a powerful approach in computer vision, particularly for tasks involving the generation and refinement of visual features. Diffusion models are a type of generative model that have gained popularity for their impressive ability to generate data, surpassing traditional methods like GANs and autoregressive transformers. They excel not only in image generation but also in tasks like image captioning, where they generate textual descriptions from visual content. Diffusion models handle complex dependencies between words and image features, offering diverse, contextually relevant captions. Recent advances in this area focus on enhancing these models for more coherent and meaningful image-to-text generation [35].

Unlike traditional models that often depend on static visual features, which fail to adapt to the changing linguistic context, the proposed model takes a different approach. This research introduces an innovative method for improving image-captioning by dynamically aligning visual features with the evolving linguistic context. We present two key modules: the Visual Weighting Module (VWM), which modifies visual features through partial attention, and the Enhanced Features Attention Module (EFAM) further enhances these features. This dynamic reweighting mechanism improves the interaction between visual and linguistic elements, resulting in accurate and contextually suitable captions. The proposed model effectively

bridges the divide between static visual features and dynamic lingual generation, leading to enhanced captioning performance.

3 Methodology

The proposed model is specifically designed to enhance the utilization of visual features that are extracted from the provided images, thereby driving significant improvements in image-captioning models and generating higher-quality captions. The overall workflow of the model is depicted in Fig. 1, illustrating the various components and their interactions. Initially, a Faster RCNN network is employed to extract key visual features from the image, capturing essential visual information such as objects, scenes, and contexts. These extracted features are then passed through the Visual Weighting Module (VWM), which plays a crucial role in dynamically adjusting the visual representations at each time step. The adjustments are made based on the evolving linguistic context, achieved through the use of partial attention mechanisms. This allows the model to focus on different aspects of the image at different points in the caption generation process. To further enhance the extracted visual features, these are processed by the Enhanced Features Attention Module (EFAM), which introduces an additional attention layer to refine and optimize the features. This continuous process of adaptation and refinement improves the alignment between the visual components of the image and the linguistic components of the caption, ultimately resulting in captions that are more accurate, contextually relevant, and semantically aligned with the image content.

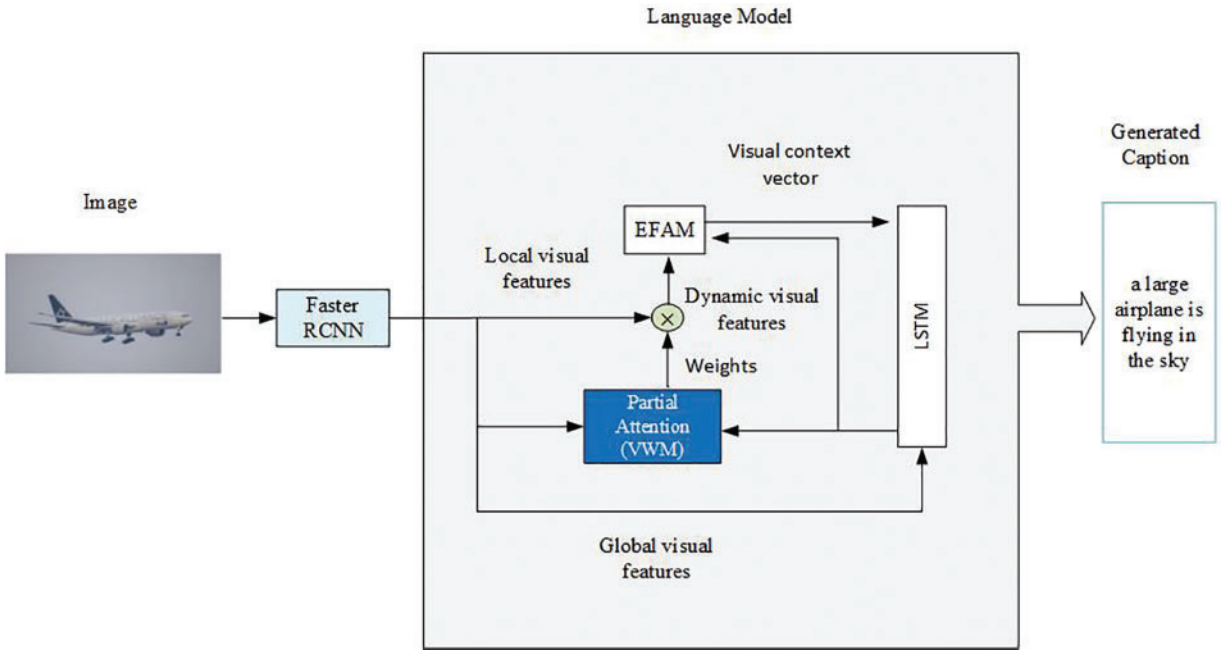


Figure 1: The structural pipeline of the proposed VARVF image annotation network

3.1 Visual Features of the Input Image

The first step to generate a description for some image, using the proposed captioning network, is to grab the visual features out of the input image. This preliminary stage includes extracting Features of objects through a Faster-RCNN architecture that utilizes ResNet-101 as its backbone. The features extracted from

the object are then organized into a feature matrix, denoted as V, which contains N objects representing the feature vectors. These visual features are essential for further processing through the language model.

$$V = \{v_1, v_2, \ldots, v_N\} \tag{1}$$

As another confirmation measurement to gauge the proposed image annotation system, we used the input image's mean-pooled object feature $\bar{v}$. The individual object feature vector is denoted as $v_i \in \mathbb{R}^h$, where i represents an index that varies from 1 to N. The collection of these object features is encapsulated in the matrix $V \in \mathbb{R}^{h \times N}$.

$$\bar{v} = \frac{1}{N} \sum_{i=1}^{N} v_i \tag{2}$$

Here, $\bar{v} \in \mathbb{R}^h$.

3.2 Visual Weighting Module (VWM)

The object features extracted by the Faster-RCNN network serve as the input for the Visual Weighting Module (VWM), a key component in the proposed model's architecture. The VWM plays a crucial role by dynamically adjusting these visual features at each time step, utilizing a partial attention mechanism. This process assigns context-aware weights to the extracted object features, which are then applied to the corresponding set of N local visual features. These weights are recalculated at each step, allowing the visual features to be modified in real-time based on the evolving linguistic context. By continuously aligning the visual features with the language model, and the use of VWM, the accuracy and relevance of the produced captions were ensured, both semantically and visually. The internal structure and functioning of the VWM, which is central to this real-time adaptation, is illustrated in Fig. 2, highlighting the intricate process of feature refinement that occurs throughout the captioning generation.

$$\alpha_t^i = W_c \cdot \tanh\left(W_a \cdot h_t^a + W_b \cdot v_i\right) \tag{3}$$

$$\beta_t = \text{softmax}\left(\alpha_t\right) \tag{4}$$

$$r_t^i = \beta_t^i \odot v_i \tag{5}$$

$$R_t = \left\{r_t^1, r_t^2, \ldots, r_t^N\right\} \tag{6}$$

where $W_b \in \mathbb{R}^{h \times e}$, $W_a \in \mathbb{R}^{g \times e}$, and $W_c \in \mathbb{R}^e$ are learnable weights. $\beta_t \in \mathbb{R}^N$ is the attention weights and $\alpha_t \in \mathbb{R}^N$. $R_t \in \mathbb{R}^{h \times N}$ represents the adjusted output features (dynamically adjusted local visual features) and $r_t^i \in \mathbb{R}^h$ signifies an individual adjusted visual vector. Attention LSTM's hidden state is $h_t^a \in \mathbb{R}^g$.

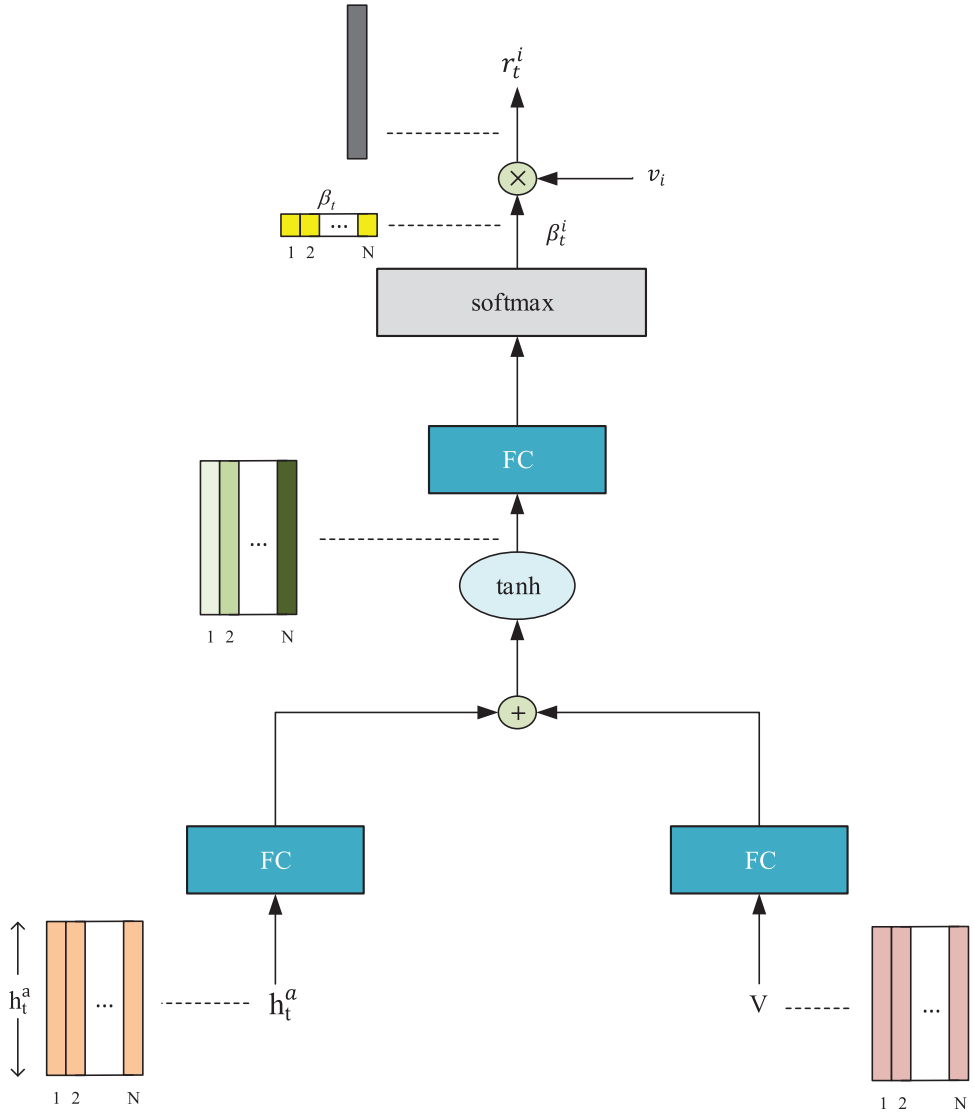


Figure 2: The inner architecture of the VWM Module

3.3 Enhanced Features Attention Module (EFAM)

For more feature refining, an additional subsequent attention layer is added, named the Enhanced Features Attention Module (EFAM), where the adjusted output features are passed to, as shown in Fig. 3. This ongoing refinement improves the alignment between visual and linguistic components, thereby increasing the accuracy and relevance of the produced captions. EFAM employs a standard visual attention mechanism, and its operations are described by the following equations:

$$\delta_t^i = W_d \cdot \tanh\left(W_e \cdot h_t^a + W_f \cdot r_t^i\right) \tag{7}$$

$$\gamma_t = \text{softmax}\left(\delta_t\right) \tag{8}$$

$$\hat{v}_t = \sum_{i=1}^{N} \gamma_t^i \odot r_t^i \tag{9}$$

where $\hat{v}_t \in \mathbb{R}^h$ is the visual context vector. $\delta_t \in \mathbb{R}^N$, and $\gamma_t \in \mathbb{R}^N$ represents the attention weights. $W_d \in \mathbb{R}^e$, $W_e \in \mathbb{R}^{g \times e}$, and $W_f \in \mathbb{R}^{h \times e}$ are trainable weights.

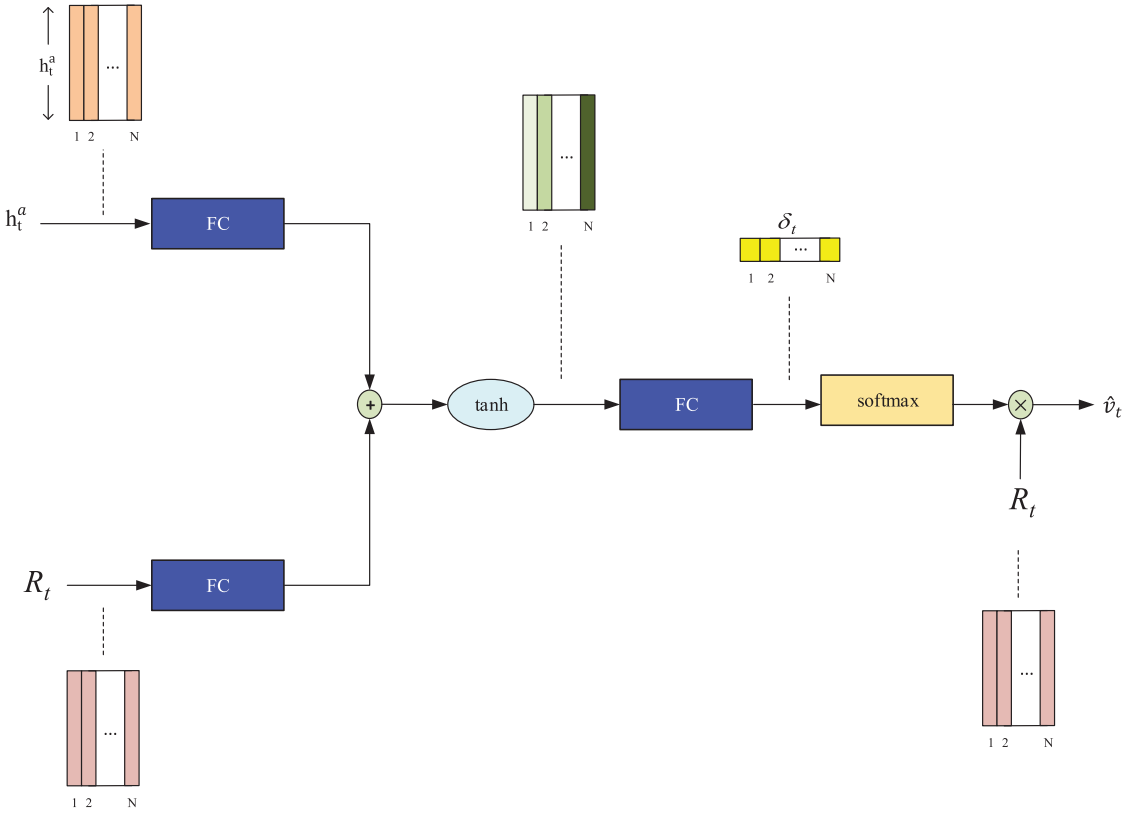


Figure 3: The inner architecture of the EFAM Module

3.4 Language Model

The language model architecture is shown in Fig. 4. The proposed approach uses the up-down framework as the baseline structure, which includes dual LSTM layers: the LSTM of attention ($LSTM_{att}$) and the language LSTM ($LSTM_{lan}$) whose hidden states are denoted by $h_t^a \in \mathbb{R}^g$ for the attention LSTM and $h_t^l \in \mathbb{R}^g$ for the language LSTM, respectively, and described by the following equations:

$$h_t^a = LSTM_{att}\left(h_{t-1}^a; \left[h_{t-1}^l, E \cdot y_{t-1}, \bar{v}\right]\right) \tag{10}$$

$$h_t^l = LSTM_{\text{lan}}\left(h_{t-1}^l; \left[h_t^a, \hat{v}_t\right]\right) \tag{11}$$

In this scenario, $E \in \mathbb{R}^{c \times m}$ represents the matrix of the word embeddings, while $y_{t-1} \in \mathbb{R}^c$ signifies the token produced from the preceding time step. The hidden state of the language LSTM, denoted as h_t^l, is subsequently directed to a fully connected layer that employs softmax as its activation function. This process results in the generation of the probability distribution p_t for predicting the next token across the entire vocabulary, which can be expressed as follows:

$$p_t = \text{softmax}\left(h_t^l \cdot W_p\right) \tag{12}$$

where the variable $p_t \in \mathbb{R}^c$. The matrix $W_p \in \mathbb{R}^{g \times c}$ denotes trainable parameters. In both the initial training phase and the subsequent testing phase, the process commences with an input of a predefined token referred to as the beginning-of-sequence (BoS). This token serves as a placeholder that initiates the operation of the attention-based Long Short-Term Memory (LSTM) network during the first time step. However, after this initial step, the process of providing input words to the attention LSTM diverges between the training and testing phases. During the training phase, at each subsequent time step, the attention LSTM receives input words directly from the ground truth annotations. This ensures that the model is consistently exposed to the correct sequence of words, allowing it to learn from actual examples. In contrast, during the testing phase, the model operates differently. Rather than using ground truth words as inputs, the attention LSTM takes its input word from the word predicted at the previous time step, as described in Eq. (10). This introduces an element of variability, as the model must rely on its own predictions to generate the subsequent word in the sequence. The word generation process continues step by step, until either an end-of-sequence (EoS) token is predicted by the model, or it reaches the maximum allowable description length. This setup ensures that, during testing, the model simulates real-world usage where ground truth information is not available and must generate descriptions based on its own predictions.

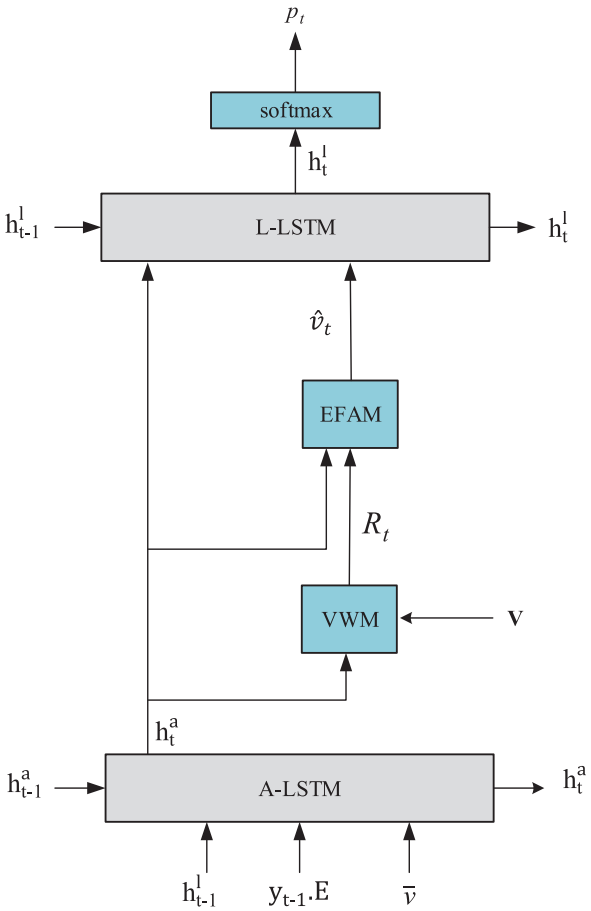


Figure 4: The inner framework of VARVF linguistic architecture for generating image captions

3.5 Loss Functions Optimization

The training of the proposed image annotation network occurs in two distinct phases. Initially, cross-entropy (XE) is employed, which is followed by a subsequent phase that emphasizes optimization through CIDEr. During the first phase, we implement the standard cross-entropy loss, which can be expressed mathematically as:

$$\text{Loss}_{XE} = \frac{1}{T}\sum_{t=1}^{T} -\log\left(p_t\left(y_t \mid y_{1:t-1}, V\right)\right) \tag{13}$$

We utilized Self Critical Sequence Training (SCST), alongside the CIDEr-D evaluation metric during the second stage to enhance the training process and eventually optimize the image annotation network. The loss function used in this stage is described as follows:

$$\text{Loss}_{RL} = -E_{w_{1:T}\sim\theta}\left[r\left(w_{1:T}\right)\right] \tag{14}$$

In this context, the sampled annotation and its CIDEr-D score are denoted by $w_{1:T}$, and r respectively, whilst the gradient approximation of Loss_{RL} "$\nabla_\theta \text{Loss}_{RL}$" is defined in Eq. (15), in which $r\left(\hat{w}_{1:T}\right)$ corresponds to the reward (yield by CIDEr) for the sampled annotation max, whereas $r\left(w^s_{1:T}\right)$ represents the reward of the random sample of the annotation:

$$\nabla_\theta \text{Loss}_{RL} = -\left(r\left(w^s_{1:T}\right) - r\left(\hat{w}_{1:T}\right)\right)\nabla_\theta \log\left(p\left(w^s_{1:T}\right)\right) \tag{15}$$

In Algorithm 1, we provide a detailed description of training procedure for our VARVF model. The process begins with the extraction of feature from the images, as defined in Eq. (1). In every word position in the caption, the algorithm calculates the refined visual features R_t through the Visual Weighting Module (VWM) (Eq. (6)) and the output generated by the Enhanced Features Attention Module (EFAM) (Eq. (9)). These refined features are then used to calculate the probability distribution for predicting the next word in the caption, as outlined in Eq. (12). Following this, the model updates the cross-entropy loss ($loss_{XE}$) and the reinforcement learning loss ($loss_{RL}$) according to Eqs. (13) and (14), respectively. This process is repeated iteratively until convergence is achieved. Upon completion, the algorithm returns the generated word probabilities and the model parameters, which reflect the knowledge and insights gained during the training process, thus enabling the model to generate more accurate captions.

Algorithm 1: VARVF model training procedure

Require: Training dataset consisting of image-caption pairs *(I, C)*, where I is the input image and $C = \{y_1, y_2, \ldots, y_T\}$ represents the corresponding ground truth caption.
Require: Initialize the batch size Z and learning rate ψ.
Ensure: The predicted word probabilities y_t, and the VARVF model parameters W

```
1: repeat
2:     Calculate feature representations V as per Eq. (1)
3:     for t = 1 to T do
4:         Compute the attention weights from the VWM module and adjust the visual features to get R_t as per Eq. (6);
5:         Obtain the output of the EFAM module v̂_t using Eq. (9);
6:         Predict the word y_t using Eq. (12);
```

(Continued)

Algorithm 1 (continued)

7: **end for**
8: Update the cross-entropy loss $loss_{XE}$ and reinforcement learning loss $loss_{RL}$ based on Eqs. (13) and (14);
9: **until** Convergence criterion is met
10: **return** Predicted words y_t, model parameters W

4 Experiments and Results

This section provides an in-depth discussion of several critical elements associated with the experiments we conducted. It covers the evaluation metrics utilized to assess the performance of our models, details the dataset that was chosen for experimentation, and outlines the specific configurations of the model used in the study. Additionally, the section elaborates on the training process of the proposed image annotation networks, highlighting the key stages and parameters involved in model optimization. We also present a comprehensive analysis of the experimental results, including a comparative evaluation of the proposed networks against other established models in the field. Furthermore, a detailed quality assessment of the generated captions is provided, offering insights into the accuracy, relevance, and fluency of the textual output produced by the models under different conditions.

4.1 Dataset and Evaluation Methods

In this study, extensive evaluations were carried out, using the well-established MS-COCO dataset [36], which is highly regarded for its richness and variety in image-caption pairs. This dataset has become a cornerstone in image-captioning research, offering a vigorous framework, whether for the evaluation or in the training of the model's performance. Its widespread use in the field is due to its diversity and the large volume of annotated data it provides, and this in turn made it a perfect option for gauging and assessing the effectiveness of the proposed model. The MS-COCO dataset, consisting of 123,287 images, is frequently used in various image annotation tasks, as noted in [36]. As part of ensuring consistent evaluation measures, we followed the data split method suggested by Karpathy et al. [37], which emphasizes the use of a training/validation ratio of 113,287/5000 images, besides another 5000 additional images are allocated for testing. This partition provides a balanced framework for model development and testing, which allows thorough evaluation and greater generalizability for the proposed model. The proposed approach has been evaluated using several widely recognized metrics, including CIDEr [38], BLEU [39], ROUGE-L [40], METEOR [41], and SPICE [42]. CIDEr calculates similarity using TF-IDF-weighted n-grams, balancing precision and recall, which is particularly suitable for image-captioning. BLEU, a precision-based metric designed originally for the translation of machine, measures the n-gram overlap, and compares the difference between the generated captions and reference captions, with BLEU-n (where n = 1, 2, 3, 4) evaluating precision at different levels. ROUGE-L, focused on recall, compares the generated captions with human references to assess content overlap. METEOR improves on BLEU by incorporating both precision and recall, taking synonyms and stemming into account. Lastly, SPICE evaluates the extent to which the semantic relationships, the objects, and the attributes were captured within the images by the captions. For brevity, we refer to symbolized these metrics using their first characters, i.e., C, B-n (where n =1, 2, 3, 4), R, M, and S, respectively.

4.2 Experimental Settings

In this study, we employ the Faster-RCNN model with a ResNet-101 backbone to extract object features from images, producing feature vectors of size 36×2048. For each image, $N = 36$ object features are

processed. Our vocabulary consists of $c = 9487$ unique words, where only words that occur more than five times in the MS-COCO dataset are included. Sentences that exceed 16 tokens are truncated. Each word in the vocabulary is represented by a vector of size $m = 1000$, and both LSTMs use 1000-dimensional embeddings with a 1000 hidden state size. The input image is characterized by a visual feature vector with dimensions $h = 2048$, which plays a crucial role in the image-captioning process. The hidden state size of the LSTM is configured to $g = 1000$, enabling it to adeptly capture intricate linguistic patterns during the generation of captions. In order to enhance the model's ability to concentrate on significant regions within the image, we incorporate an attention mechanism with a dimensionality of $e = 512$.

In our training process, we utilize the Adam optimizer and execute the model for 50 epochs during the cross-entropy training phase, followed by an additional 100 epochs dedicated to CIDEr optimization. The initial learning rate is set at 0.0005 which is decreased by a factor of 0.8 every 5 epochs throughout the cross-entropy phase and subsequently reduced every 10 epochs during the CIDEr optimization phase. We maintain a consistent batch size of 40, while scheduled sampling is incrementally increased by 5 For testing, a beam search strategy was employed, with a beam size = 3 to generate the final captions. All models are developed using the PyTorch framework, ensuring robust and efficient implementation.

Tables 1 and 2 provide a representation of parameter symbols with their corresponding values, and training parameters with their corresponding values, respectively.

Table 1: Parameter symbols and their corresponding values

Name	Symbol	Value
Visual features vector size	h	2048
LSTM hidden state size	g	1000
Word embedding vector length	m	1000
Number of object features	N	36
Vocabulary size	c	9487
Internal hidden attention size	e	512

Table 2: Experimental setup: corresponding values for key parameters and configurations

Name	Value
Batch size	40
Total number of training epochs	150
Total number of training epochs (XE)	50
Total number of training epochs (RL)	100
Learning rate	0.0005
Learning rate decay rate	0.8
Learning rate decay every (XE)	5 epochs
Learning rate decay every (RL)	10 epochs
Scheduled sampling increases by	5%

(Continued)

Table 2 (continued)

Name	Value
Scheduled sampling increases every	5 epochs
Scheduled sampling maximum limit	25%
Dropout ratio	0.5
Gradient clipping maximum absolute value	0.1
Beam size	3

4.3 Quantitative Scores

Table 3 presents an extensive overview of the performance metrics associated with the proposed model on the MS-COCO dataset, assessed during both the cross-entropy and CIDEr optimization phases. The image-captioning model demonstrated consistently enhanced performance in the cross-entropy phase when compared to the baseline across all evaluation metrics. In particular, the model achieved remarkable improvements in key metrics such as CIDEr, SPICE, BLEU-4, and METEOR. These results indicate the proposed model acquires a strong ability to produce high-quality captions even at the early stages of training. The cross-entropy phase thus serves as a solid foundation for further optimization. Furthermore, the results from the CIDEr optimization phase highlight even greater advancements. The proposed model exhibited substantial improvements over the baseline, particularly in CIDEr, BLEU-1, and ROUGE scores. These metrics underscore the model's enhanced ability to produce not only accurate but also more diverse captions. Notably, the CIDEr phase focuses on optimizing caption quality, ensuring that the generated descriptions are both meaningful and varied. These findings clearly indicate the effectiveness of the proposed framework, VARVF, and its ability to refine captions throughout training. In this phase, the proposed model showed a marked improvement in generating captions that align more closely with human evaluation metrics, reflecting the framework's potential in practical applications. Integrating the Visual Weighing Mechanism (VWM) and the Enhanced Feature Attention Mechanism (EFAM) within the proposed network architecture has played a critical role in boosting image-captioning performance. These modules work in tandem to provide continuous adaptation and alignment between the visual and linguistic elements of the model. The VWM dynamically adjusts visual features in real-time, based on the evolving linguistic context, ensuring that the visual information remains relevant throughout the caption generation process. Meanwhile, the EFAM further refines the visual features through an additional attention layer, which places emphasis on the most contextually relevant parts of the image. This ongoing reweighting and refinement process greatly enhances the model's ability to produce captions that are semantically accurate and visually aligned with the image content. As a result, the captions produced are not only more accurate but also contextually richer, offering a coherent description that closely matches the visual elements of the image. The combination of VWM and EFAM contributes significantly to the model's overall capability to create detailed captions with high contextual relevance, and that are better aligned with human expectations. This demonstrates the model's strength in balancing linguistic and visual information, leading to more natural and meaningful captioning outcomes.

Table 3: XE (Cross-Entropy) & RL (the proposed model's scores of CIDEr optimization), evaluated on the MS-COCO dataset

Model	B1	B4	M	R	C	S
VARVF(XE)	76.9	36.8	27.8	56.8	114.6	20.8
Baseline (XE) [8]	76.6	36.2	27.0	56.4	113.5	20.3
VARVF(CIDEr)	80.9	36.8	28.2	57.8	122.1	21.6
Baseline (CIDEr) [8]	79.8	36.3	27.7	56.9	120.1	21.4

4.4 Comparison Results

Tables 4 and 5 provide a detailed comparison between the performance of the proposed model and various state-of-the-art image-captioning approaches on the MS-COCO dataset. The comparative analysis is conducted over two distinct phases: the cross-entropy training and the CIDEr optimization. Table 4 illustrates the cross-entropy results, while Table 5 focuses on the scores obtained during CIDEr optimization. These tables offer a comprehensive view of how the proposed model stands in relation to other leading models in the field, across multiple evaluation metrics that are common in measuring image-captioning tasks. In the cross-entropy results displayed in Table 4, the proposed model consistently surpasses most of the other methods. Specifically, the proposed model achieves the highest scores in BLEU-3, BLEU-4, CIDEr, SPICE, and ROUGE-L, showing its strong ability to generate captions that are both accurate and descriptive. Furthermore, the proposed model ranks second in BLEU-1 and BLEU-2, demonstrating its robustness across different evaluation metrics. One notable exception is the r-GRU model, which secures the top position in the METEOR metric. Despite this, the overall performance of the proposed model in this phase is highly competitive, with significant differences in performance when compared to the other models. These performance gaps highlight the strength of the proposed approach, particularly during the training of the cross-entropy phase, where the proposed model effectively captures the relationship between visual features and linguistic elements to produce high-quality captions. Table 5 presents the results from the CIDEr optimization phase, where the proposed model continues to demonstrate exceptional performance. In this phase, the proposed model leads in multiple metrics, including BLEU-1, BLEU-2, BLEU-3, METEOR, SPICE, and ROUGE-L. These results further emphasize the model's capacity to generate diverse and high-quality captions. Although the proposed model ranks second in the CIDEr metric, the overall performance remains strong and competitive, showcasing its ability to optimize caption generation not only for quality but also for relevance and diversity. The improvement in CIDEr optimization results indicates that the proposed model effectively balances caption accuracy with fluency and contextual richness.

Table 4: Comparison of the performance of VARVF and the other models trained using the MS-COCO dataset, harnessing cross-entropy (XE) optimization. The second-highest and highest scores were underlined and highlighted in bold, respectively

Model	B1	B2	B3	B4	M	R	C	S
UpDown [8]	**77.2**	–	–	36.2	27.0	56.4	113.5	20.3
RFNet [43]	76.4	60.4	46.6	35.8	27.4	56.5	112.5	20.5
RecallNet [44]	73.4	–	–	32.2	25.9	53.9	101.6	–
SCST [45]	–	–	–	30.0	25.9	53.4	99.4	–

(Continued)

Table 4 (continued)

Model	B1	B2	B3	B4	M	R	C	S
HAF [46]	75.9	59.5	45.4	34.4	26.8	–	109.0	–
VIS_SAS [25]	72.5	52.6	38.2	28.1	23.7	55.4	82.1	–
Vis-to-Lang [26]	73.9	56.4	41.7	30.9	27.1	–	–	–
MRRC [47]	75.5	59.8	46.0	35.2	26.5	55.9	108.0	19.7
TAAIC [48]	71.0	–	–	27.7	23.8	51.1	93.2	18.3
NumCap [34]	66.9	49.4	36.5	27.3	24.1	50.7	85.3	17.0
CSA [49]	**77.2**	59.8	46.0	36.2	27.9	56.4	**114.6**	–
r-GRU [50]	**77.2**	**61.3**	46.3	35.6	**30.2**	55.7	109.2	–
VFDICM [51]	76.4	60.4	46.9	36.3	27.7	56.6	113.9	20.6
VARVF (ours)	76.9	60.9	**47.4**	**36.8**	27.8	**56.8**	**114.6**	**20.8**

Table 5: comparative analysis of VARVF and alternative methodologies, utilizing the MS-COCO dataset, during the CIDEr optimization (RL) phase. The highest score is highlighted in bold, while the second highest is marked with an underline

Model	B1	B2	B3	B4	M	R	C	S
UpDown [8]	79.8	–	–	36.3	27.7	56.9	120.1	21.4
RFNet [43]	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
HAF [46]	80.5	62.9	47.7	35.5	27.3	–	116.4	–
RecallNet [44]	75.8	–	–	33.1	24.7	54.9	103.7	–
SCST [45]	–	–	–	34.2	26.7	55.7	114.0	–
Stack-VS [27]	79.4	63.6	49.0	**37.2**	27.9	57.7	**122.6**	**21.6**
TAAIC [48]	78.6	–	–	37.1	27.5	57.2	119.6	21.2
TDA+GLD [52]	78.8	62.6	48.0	36.1	27.8	57.1	121.1	**21.6**
VARVF (ours)	**80.9**	**64.0**	**49.1**	36.8	**28.2**	**57.8**	122.1	**21.6**

The considerable impact of the proposed method in enhancing model performance should be acknowledged through these results. By integrating the VWM and EFAM modules into the suggested network, significant improvements in image-captioning are achieved, as they promote continuous adaptation and synchronization between visual and textual elements. This process of continuous reweighting and refinement enables the model to produce captions that are both semantically and visually coherent, resulting in more accuracy and contextual appropriation of the descriptions, characterized by high consistency and alignment with the actual contents of the image. The VWM adjusts visual features in real-time, based on the changing linguistic context, meanwhile, these features are further refined by the EFAM using an extra attention layer.

4.5 Ablation Studies

In this ablation study, we conduct ablation studies to evaluate the independent effects of the Visual Weighting Mechanism (VWM). We compare two configurations: VARVF-Base, which does not include the VWM module, and VARVF-VWM, which incorporates the VWM module. As demonstrated in Table 6, the results clearly highlight the significant impact of using the VWM module. The inclusion of VWM leads to noticeable improvements in performance, underlining its importance in enhancing the overall model accuracy and efficiency. This comparison provides valuable insights into the individual contributions of each component to the system's performance.

Table 6: Ablation study on the independent effects of the visual weighting mechanism (VWM) using cider optimization and MS-COCO dataset

Model	B1	B4	M	R	C	S
VARVF-Base	80.7	36.8	27.7	56.6	121.5	21.5
VARVF-VWM	80.9	36.8	28.2	57.8	122.1	21.6

5 Qualitative Evaluation

Along with the quantitative score analysis, it is essential to evaluate the quality of captions produced by the VARVF model. Fig. 5 showcases several sample images from the test dataset, each accompanied by their corresponding captions. For each image, three types of descriptions are provided: the first generated by the proposed model, the second by the baseline model, and the third representing the ground truth, which consists of five human-annotated captions. For instance, take the image located in the top-left corner (first row, first column). The caption generated by the proposed model, "a group of elephants standing in a zoo enclosure," offers a more detailed account by explicitly mentioning the "group of elephants" and specifying the location as a "zoo enclosure." This caption bears a close resemblance to the human-annotated actual caption, "Several elephants in a zoo enclosure with onlookers watching," demonstrating the notable capacity of the proposed model to generate captions with human-like quality and precision. The upper right image represents another example. The resulting caption from the proposed model, "a red motorcycle parked on the side of a city street," provides an accurate description of the scene. This caption is aligned with the actual description, "Red motorcycle parked outside of a large building in the city." The high performance of the proposed VARVF model, besides the quality of the produced captions are consistently strong and significantly surpasses the baseline of the standard evaluation metrics, as demonstrated by the scores provided in Table 5.

R :a group of elephants standing in a zoo enclosure

B: a group of elephants standing in a zoo

GT1: Several elephants are in a habitat as heads are in the foreground.

GT2: A small gray elephant standing in an exhibit at a zoo.

GT3: People are watching four elephants in a zoo.

GT4: Several elephants in zoo enclosure with onlookers watching.

GT5: An elephant in a zoo stands in front of the crowd.

R: a red motorcycle parked on the side of a city street

B: a red motorcycle parked on the side of a street

GT1: Red motorcycle parked outside of large building in the city,

GT2: A motorcycle parked on a sidewalk with a man in the background.

GT3: A red motorcylcle parked outside a building on the sidewalk.

GT4: A motorcycle is pictured outside of a building with a man walking away from it.

GT5: A motorcycle on the sidewalk outside of a building.

R: a group of people sitting at a table eating pizza

B: a man and woman sitting at a table with pizza

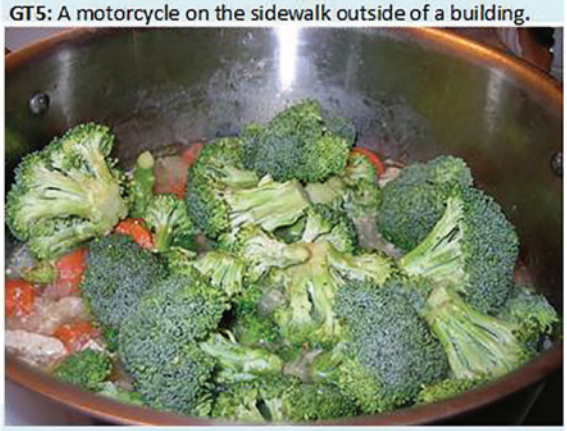
GT1: A group of people sitting at a table holding different pizzas.

GT2: A family is having a pizza dinner at a restaurant

GT3: A group of people sitting around a table with pizza on it.

GT4: Several people that are eating some food together.

GT5: a group of people sitting at a table with different pans of pizza



R: a pan filled with broccoli and vegetables on a stove

B: a pan filled with broccoli and sitting on a table

GT1: A bowl full of broccoli and tomatoes being cooked.

GT2: A pot full of vegetables is sitting on a table.

GT3: A round plate of broccoli and some other vegetables.

GT4: A metal stir fry pan holds broccoli and carrots.

GT5: A bowl filled with broccoli and lots of other vegetables.

Figure 5: Examples of captions generated by the VARVF model. The caption labeled as R is produced by the VARVF model, B is generated by the baseline model, and GT represents the corresponding ground truth captions

Fig. 6 illustrates two sample images generated by our model, including ambiguous cases. For example, consider the image on the left. The caption generated by the proposed model, "a black and white cat laying in front of a door," demonstrates a challenge in accurately describing the dog in the image. The model failed to recognize the dog and instead identified it as a cat. This error is likely due to ambiguity in the image, as the dog appears small and distant within the overall photo. Another example can be seen in the image on the right. The caption generated by the proposed model, "a cow standing in a field next to a building," partially reflects the scene, as the building is indeed visible in the image. However, the caption inaccurately suggests that the cow is positioned next to the building, which is not the case.

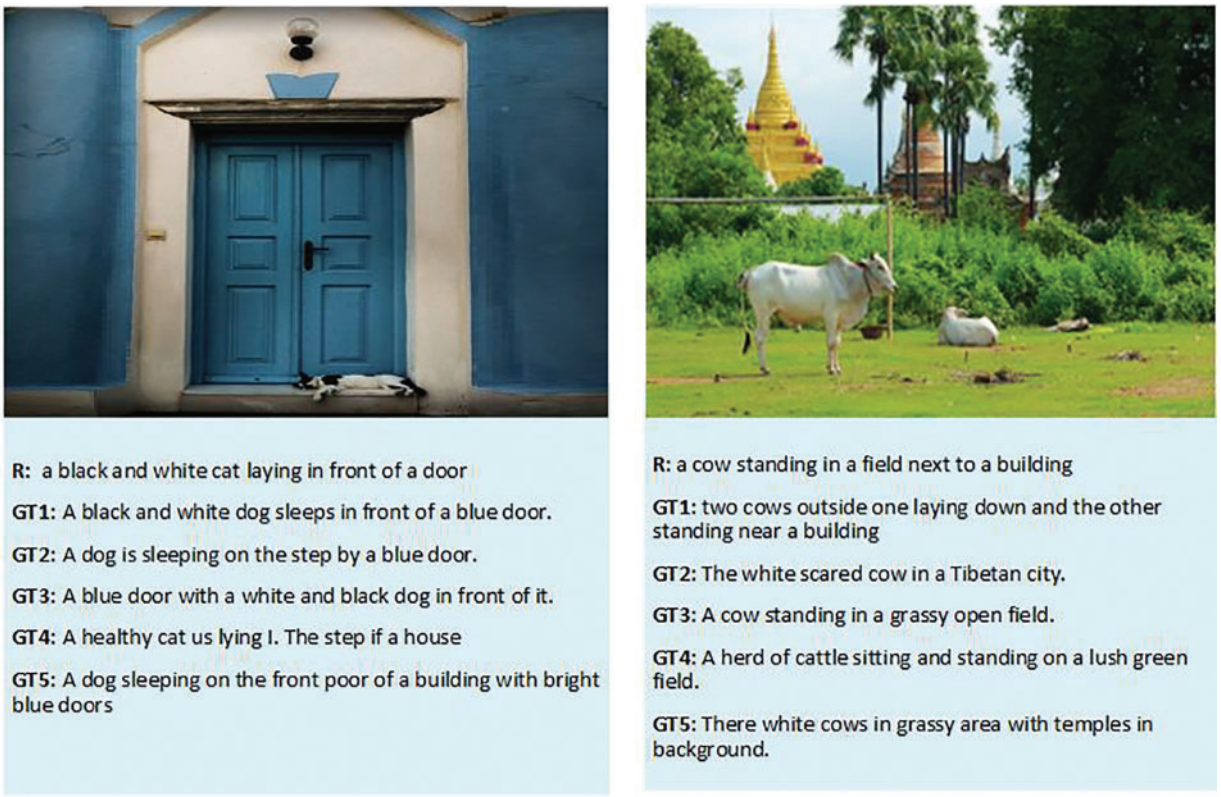


Figure 6: Examples of images generated by VARVF model, including ambiguous cases. The VARVF generates caption R, while the GT represents the ground truth captions

6 Discussion

In this study, we have introduced a novelty framework in images' caption generation that dynamically adjusts visual features based on the evolving linguistic context to generate more coherent captions, with highly contextual relevance. Traditional image-captioning models often rely on static visual features, which limits their ability to form meaningful connections between the visual input and the text output (the produced captions). The proposed model overcomes this limitation by dynamically reweighting the visual features at each time step using a Visual Weighting Module (VWM), followed by further refinement through an Enhanced Features Attention Module (EFAM). This innovative approach enables continuous adaptation of the visual features, ensuring that they remain aligned with the linguistic context throughout the caption generation process. Unlike conventional models, where visual features remain fixed, the proposed

method introduces a novel partial attention to modifying these features in real-time, ensuring stronger visual-linguistic alignment. By incorporating both VWM and EFAM, the model dynamically adjusts and refines the relevant visual elements at each stage, leading to more accurate and semantically coherent captions. This adaptive mechanism results in improving caption quality by making the generated descriptions more closely tied to the visual components of the image. Our approach emphasizes the importance of dynamically adjusting visual features and contributes significantly to advancing image-captioning techniques by enhancing the interaction between visual and linguistic elements. Based on various experiments with the MS-COCO's dataset it has been demonstrated that the proposed framework performs competitively with state-of-the-art methods, achieving substantial improvements in both visual and linguistic coherence in the generated captions.

7 Conclusion

In this paper, we presented a novelty approach to image caption generation that addresses the limitations of traditional models by introducing dynamic interaction between visual features and the evolving linguistic context. The proposed model enhances the relationship between the visual and linguistic elements by reweighting the visual features at each time step using the Visual Weighting Module (VWM) and further refining them through the Enhanced Features Attention Module (EFAM). This continuous adaptation ensures that the visual features remain contextually aligned with the linguistic content throughout the captioning process, resulting in more accurate, coherent captions, with highly contextual relevance. The quality of generated captions significantly improved using the proposed method, this can be interpreted by the application of a methodology that involves making the captions more tied up to the visual input. Evaluation of the MS-COCO dataset demonstrates that the proposed approach has surpassed the competitive performance with recent state-of-the-art methods, highlighting the model's effectiveness in covering the gap between static visual features and dynamic lingual generation. This work adds another contribution to the advancement of image-captioning models and opens new possibilities for further exploration of visual-linguistic alignment. In future work, we plan to explore the potential of advanced model architectures, such as the incorporation of Transformers, to enhance performance. Additionally, we aim to investigate innovative techniques in visual feature extraction, including multi-scale feature extraction and multimodal self-attention mechanisms, to improve the model's ability to capture complex and diverse image features, ultimately refining the accuracy and relevance of generated captions.

Acknowledgment: The authors wish to express their gratitude to Prince Sultan University for their support.

Funding Statement: This work is supported by the National Natural Science Foundation of China (Nos. U22A2034, 62177047), High Caliber Foreign Experts Introduction Plan funded by MOST, and Central South University Research Programme of Advanced Interdisciplinary Studies (No. 2023QYJC020). Also, the authors would like to thank Prince Sultan University for paying the APC of this article.

Author Contributions: The contributions of the authors to this work are as follows: Alaa Thobhani was responsible for investigation, conceptualization, visualization, and software development. Xiaoyan Kui contributed to the review and editing of the manuscript. Beiji Zou provided supervision throughout the project. Amr Abdussalam participated in the investigation, while Muhammad Asim contributed resources. Mohammed ELAffendi handled validation, and Sajid Shah conducted formal analysis. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The widely recognized MS-COCO dataset, which is publicly accessible, was utilized.

Ethics Approval: Not applicable.

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

References

1. Bashir MH, Ahmad M, Rizvi DR, El-Latif AAA. Efficient CNN-based disaster events classification using UAV-aided images for emergency response application. Neural Comput Appl. 2024:1–14. doi:10.1007/s00521-024-09610-4.
2. Ibrahim H, Mohamed AEN, Ammar R, El-Hag NA, Abou-Elazm A, Abd El-Samie FE, et al. Efficient color image enhancement using piecewise linear transformation and gamma correction. J Opt. 2024;53(3):2027–37.
3. Waheed SR, Suaib NM, Rahim MSM, Khan AR, Bahaj SA, Saba T. Synergistic integration of transfer learning and deep learning for enhanced object detection in digital images. IEEE Access. 2024;12:13525–36.
4. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 4651–9.
5. Jiang W, Wang W, Hu H. Bi-directional co-attention network for image captioning. ACM Transact Multimed Comput, Commun Appl (TOMM). 2021;17(4):1–20. doi:10.1145/3460474.
6. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning, ser. Proceedings of Machine Learning Research; 2015; Lille, France: PMLR. Vol. 37, p. 2048–57.
7. Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 375–83.
8. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. 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; 2018. p. 6077–86.
9. Ji J, Xu C, Zhang X, Wang B, Song X. Spatio-temporal memory attention for image captioning. IEEE Trans Image Process. 2020;29:7615–28. doi:10.1109/TIP.2020.3004729.
10. Yu L, Zhang J, Wu Q. Dual attention on pyramid feature maps for image captioning. IEEE Trans Multimedia. 2021;24:1775–86. doi:10.1109/TMM.2021.3072479.
11. Jiang W, Zhu M, Fang Y, Shi G, Zhao X, Liu Y. Visual cluster grounding for image captioning. IEEE Trans Image Process. 2022;31:3920–34. doi:10.1109/TIP.2022.3177318.
12. Yu J, Li J, Yu Z, Huang Q. Multimodal transformer with multi-view visual representation for image captioning. IEEE Trans Circuits Syst Video Technol. 2019;30(12):4467–80. doi:10.1109/TCSVT.2019.2947482.
13. Liu A-A, Zhai Y, Xu N, Nie W, Li W, Zhang Y. Region-aware image captioning via interaction learning. IEEE Trans Circuits Syst Video Technol. 2021;32(6):3685–96. doi:10.1109/TCSVT.2021.3107035.
14. Guo M-H, Lu C-Z, Liu Z-N, Cheng M-M, Hu S-M. Visual attention network. Comput Vis Media. 2023;9(4):733–52. doi:10.1007/s41095-023-0364-2.
15. Wang C, Gu X. Learning joint relationship attention network for image captioning. Expert Syst Appl. 2023;211:118474.
16. Hossen MB, Ye Z, Abdussalam A, Hossain MI. GVA: guided visual attention approach for automatic image caption generation. Multimed Syst. 2024;30(1):50.
17. Zhang M, Yang Y, Zhang H, Ji Y, Shen HT, Chua T-S. More is better: precise and detailed image captioning using online positive recall and missing concepts mining. IEEE Trans Image Process. 2018;28(1):32–44.
18. Huang Y, Chen J, Ouyang W, Wan W, Xue Y. Image captioning with end-to-end attribute detection and subsequent attributes prediction. IEEE Trans Image Process. 2020;29:4013–26.
19. Bae J-W, Lee S-H, Kim W-Y, Seong J-H, Seo D-H. Image captioning model using part-of-speech guidance module for description with diverse vocabulary. IEEE Access. 2022;10:45219–29.
20. Zhang J, Mei K, Zheng Y, Fan J. Integrating part of speech guidance for image captioning. IEEE Trans Multimedia. 2020;23:92–104. doi:10.1109/TMM.2020.2976552.
21. Yu N, Hu X, Song B, Yang J, Zhang J. Topic-oriented image captioning based on order-embedding. IEEE Trans Image Process. 2018;28(6):2743–54. doi:10.1109/TIP.2018.2889922.

22. Wei H, Li Z, Huang F, Zhang C, Ma H, Shi Z. Integrating scene semantic knowledge into image captioning. ACM Transact Multimed Comput Commun Appl. 2021;17(2):1–22. doi:10.1145/3439734.
23. Liu M, Hu H, Li L, Yu Y, Guan W. Chinese image caption generation via visual attention and topic modeling. IEEE Trans Cybern. 2020;52(2):1247–57. doi:10.1109/TCYB.2020.2997034.
24. Al-Qatf M, Wang X, Hawbani A, Abdusallam A, Alsamhi SH. Image captioning with novel topics guidance and retrieval-based topics re-weighting. IEEE Trans Multimedia. 2023;25:5984–99
25. Zhou L, Zhang Y, Jiang Y-G, Zhang T, Fan W. Re-caption: saliency-enhanced image captioning through two-phase learning. IEEE Trans Image Process. 2019;29:694–709.
26. Li X, Yuan A, Lu X. Vision-to-language tasks based on attributes and attention mechanism. IEEE Trans Cybern. 2019;51(2):913–26.
27. Cheng L, Wei W, Mao X, Liu Y, Miao C. Stack-VS: stacked visual-semantic attention for image caption generation. IEEE Access. 2020;8:154953–65.
28. Rotstein N, Bensaid D, Brody S, Ganz R, Kimmel R. FuseCap: leveraging large language models to fuse visual data into enriched image captions. arXiv:2305.17718. 2023.
29. Haque N, Labiba I, Akter S. FaceAtt: enhancing image captioning with facial attributes for portrait images. arXiv:2309.13601. 2023.
30. Hossen MB, Ye Z, Abdussalam A, Hossain MA. ICEAP: an advanced fine-grained image captioning network with enhanced attribute predictor. Displays. 2024;84:102798.
31. Hossen MB, Ye Z, Abdussalam A, Wahab FE. Attribute guided fusion network for obtaining fine-grained image captions. Multimed Tools Appl. 2024:1–35.
32. Dai B, Fidler S, Urtasun R, Lin D. Towards diverse and natural image descriptions via a conditional GAN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2970–9.
33. Mao Y, Zhou C, Wang X, Li R. Show and tell more: topic-oriented multi-sentence image captioning. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 2018. p. 4258–64.
34. Abdussalam A, Ye Z, Hawbani A, Al-Qatf M, Khan R. NumCap: a number-controlled multi-caption image captioning network. ACM Transact Multimed Comput Commun Appl. 2023;19(4):1–24.
35. Daneshfar F, Bartani A, Lotfi P. Image captioning by diffusion models: a survey. Eng Appl Artif Intell. 2024;138:109288.
36. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference; 2014 Sep 6–12; Zurich, Switzerland: Springer. p. 740–55.
37. Karpathy A, Fei Fei L. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 3128–37.
38. Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 4566–75.
39. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002. p. 311–8.
40. Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74–81.
41. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; 2005. p. 65–72.
42. Anderson P, Fernando B, Johnson M, Gould S. SPICE: semantic propositional image caption evaluation. In: Computer Vision–ECCV 2016: 14th European Conference; 2016 Oct 11–14. Amsterdam, The Netherlands: Springer. p. 14–398.
43. Jiang W, Ma L, Jiang Y-G, Liu W, Zhang T. Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 499–515.
44. Wu L, Xu M, Wang J, Perry S. Recall what you see continually using gridlstm in image captioning. IEEE Trans Multimedia. 2019;22(3):808–18.

45. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 7008–24.
46. Wu C, Yuan S, Cao H, Wei Y, Wang L. Hierarchical attention-based fusion for image caption with multi-grained rewards. IEEE Access. 2020;8:57943–51.
47. Sur C. MRRC: multiple role representation crossover interpretation for image captioning with R-CNN feature distribution composition (FDC). Multimed Tools Appl. 2021;80(12):18413–43.
48. Yan C, Hao Y, Li L, Yin J, Liu A, Mao Z, et al. Task-adaptive attention for image captioning. IEEE Trans Circuits Syst Video Technol. 2022;32(1):43–51. doi:10.1109/TCSVT.2021.3067449.
49. Zhao D, Yang R, Wang Z, Qi Z. A cooperative approach based on self-attention with interactive attribute for image caption. Multimed Tools Appl. 2023;82(1):1223–36. doi:10.1007/s11042-022-13279-z.
50. do Carmo Nogueira T, Vinhal CDN, da Cruz Júnior G, Ullmann MRD, Marques TC. A reference-based model using deep learning for image captioning. Multimed Syst. 2023;29(3):1665–81. doi:10.1007/s00530-022-00937-3.
51. Thobhani A, Zou B, Kui X, Abdussalam A, Asim M, Ahmed N, et al. A concise and varied visual features-based image captioning model with visual selection. Comput Mater Contin. 2024;81(2):2873–4. doi:10.32604/cmc.2024.054841.
52. Wu J, Chen T, Wu H, Yang Z, Luo G, Lin L. Fine-grained image captioning with global-local discriminative objective. IEEE Trans Multimedia. 2020;23:2413–27. doi:10.1109/TMM.2020.3011317.