Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065169

ARTICLE

LREGT: Local Relationship Enhanced Gated Transformer for Image Captioning

Yuting He and Zetao Jiang*

Guangxi Key Lab of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin, 541004, China

*Corresponding Author: Zetao Jiang. Email: zetaojiang@guet.edu.cn

Received: 05 March 2025; Accepted: 11 June 2025; Published: 30 July 2025

ABSTRACT: Existing Transformer-based image captioning models typically rely on the self-attention mechanism to capture long-range dependencies, which effectively extracts and leverages the global correlation of image features. However, these models still face challenges in effectively capturing local associations. Moreover, since the encoder extracts global and local association features that focus on different semantic information, semantic noise may occur during the decoding stage. To address these issues, we propose the Local Relationship Enhanced Gated Transformer (LREGT). In the encoder part, we introduce the Local Relationship Enhanced Encoder (LREE), whose core component is the Local Relationship Enhanced Module (LREM). LREM consists of two novel designs: the Local Correlation Perception Module (LCPM) and the Local-Global Fusion Module (LGFM), which are beneficial for generating a comprehensive feature representation that integrates both global and local information. In the decoder part, we propose the Dual-level Multi-branch Gated Decoder (DMGD). It first creates multiple decoding branches to generate multi-perspective contextual feature representations. Subsequently, it employs the Dual-Level Gating Mechanism (DLGM) to model the multi-level relationships of these multi-perspective contextual features, enhancing their fine-grained semantics and intrinsic relationship representations. This ultimately leads to the generation of high-quality and semantically rich image captions. Experiments on the standard MSCOCO dataset demonstrate that LREGT achieves state-of-the-art performance, with a CIDEr score of 140.8 and BLEU-4 score of 41.3, significantly outperforming existing mainstream methods. These results highlight LREGT's superiority in capturing complex visual relationships and resolving semantic noise during decoding.

KEYWORDS: Image captioning; local relation enhancement; local correlation perception; dual-level gating mechanism

1 Introduction

The image captioning task is an important task in the field of artificial intelligence, motivated by the goal of enabling computers to "look at images and generate descriptions". It has applications in areas such as image-text retrieval and assisting people with disabilities in their daily lives. For instance, image captioning technologies are pivotal in developing accessibility tools for the visually impaired. By converting visual content into natural language descriptions, these systems empower users to "see" their surroundings through auditory feedback. Applications like real-time scene narration in smartphones or wearable devices can describe objects (e.g., "a red traffic light ahead"), actions (e.g., "a person waving"), and contextual relationships (e.g., "a dog sitting beside a bicycle"), significantly enhancing independence and safety for visually impaired individuals. Additionally, image captioning plays a vital role in content generation for social media platforms. Automated caption generation streamlines the creation of engaging posts by providing context-aware descriptions for uploaded images. For example, a photo of a sunset over a beach might be tagged with "Golden sunset at Malibu Beach", improving discoverability and user engagement. This not only reduces manual effort but also ensures consistency and relevance in content curation, benefiting both users and platform algorithms. Essentially, the image captioning task is a translation task that involves using natural language to represent the visual semantic information contained in an image. It requires not only learning and understanding the semantic information within the image but also effectively aligning visual and textual information to generate natural language descriptions that accurately convey the content of the image [1]. This is a highly challenging research task.

With the great success of deep learning in the fields of computer vision and natural language processing, image captioning methods based on the encoder-decoder framework have been widely used. Inspired by the idea of "encoding-decoding" in machine translation tasks, the encoder serves as a feature extractor to capture the visual features of the input image, while the decoder acts as a text generator to produce descriptive sentences based on the visual features extracted by the encoder. In early research, encoders based on Convolutional Neural Networks (CNNs) and decoders based on Recurrent Neural Networks (RNNs) were commonly used. Nevertheless, the capacity of these approaches to model long-range dependencies was limited. After the Transformer architecture gained prominence, researchers noticed that its built-in self-attention mechanism, compared to RNNs, has a much stronger ability to model long-range dependencies and correlate contextual information [2,3]. As a result, most current mainstream image captioning methods are designed based on the Transformer.

Although Transformer-based image captioning methods [4–8] have achieved remarkable performance, they still have the following shortcomings: (1) These methods often use the self-attention mechanism to model long-range dependencies, thereby leveraging the global correlation of image features. However, this may cause the model to overlook the potential local correlations between different regional features in the image. As a result, the descriptive statements generated by the model may lack detailed text related to these local correlations between regions. (2) The global and local correlation features extracted by the encoder often focus on different semantic information, which inevitably introduces noise during the decoding stage. Previous methods typically used the encoding results from the last layer of the encoder directly or utilized the encoding outputs from different levels of the encoder for multi-level feature extraction. However, neither approach effectively mitigates the aforementioned noise problems. To perceive and utilize the local correlations between feature sequences, recent works [9,10] have proposed modeling the interactions between multiple feature units in adjacent positions in two-dimensional space using convolution. However, convolutional operations have limitations in integrating high-level semantic signals, which may not be well-suited for the image captioning task. A more effective approach is to use a Multi-Layer Perceptron (MLP) to model the local correlations between sequences. This approach has a simple structure and high computational efficiency, and it can effectively leverage the local semantic relationships in images. Some recent studies [11–13] have also proven this viewpoint.

Therefore, we propose LREGT (Local Relationship Enhanced Gated Transformer), which consists of a Local Relationship Enhanced Encoder (LREE) and a Dual-level Multi-branch Gated Decoder (DMGD). During the encoding stage, a semantic modeling branch based on MLP is introduced to additionally perceive local correlations. In the decoding stage, multiple decoding branches are set up to obtain multi-angle contextual semantics, and a Dual-Level Gating Mechanism (DLGM) is used to refine their intrinsic semantic representations. Specifically, the LREE includes two novel designs: the Local Correlation Perception Module (LCPM) and the Local-Global Fusion Module (LGFM). First, the LCPM extracts local correlation features from the image. Then, the LGFM further integrates and models the non-linear relationships between global and local correlation features, ultimately generating a comprehensive feature representation that includes both global and local information. In the DMGD, multiple decoding branches are first set up to generate multi-perspective feature representations. Subsequently, the DLGM is employed to model the multi-level relationships of these multi-perspective contextual feature representations, enhancing the semantic information and intrinsic relationships of the features. This ultimately leads to the generation of high-quality and semantically rich image captions. Our major contributions are summarized as follows:

- We propose the LREE. First, the LCPM is designed to effectively capture local correlation features between different visual objects. Additionally, the LGFM adaptively fuses global and local correlation features to obtain comprehensive correlation features, ultimately providing higher-quality visual encodings for the decoding process of the decoder.
- We propose the DMGD. It includes multiple decoding branches that generate multi-perspective contextual feature representations. Furthermore, a progressive modeling structure is designed for the DLGM, which uses dual-level relationship weighting to enable the decoder to obtain decoding features with clear semantic relationships and rich semantic information. This approach also effectively addresses potential noise issues arising from the different correlations of visual features during the decoding stage through adaptive fusion.
- By integrating the above two modules, we construct the novel LREGT model. This model can fully perceive the local correlation information between different visual regions in an image and resolve semantic noise arising from the interaction between global and local correlation features during decoding. Extensive experiments on the standard MSCOCO dataset demonstrate that our approach achieves superior performance compared to existing state-of-the-art image captioning methods.

The remainder of this paper is organized as follows: Section 2 reviews related work on image captioning methods and the role of Multi-Layer Perceptrons (MLPs) in visual modeling. Section 3 introduces the proposed LREGT framework, detailing the Local Relationship Enhanced Encoder (LREE) and the Dual-level Multi-branch Gated Decoder (DMGD). Section 4 presents extensive experiments on the MSCOCO dataset, including ablation studies, comparisons with state-of-the-art methods, and qualitative visualizations. Finally, Section 5 concludes the paper and discusses potential future directions.

2 Related Work

2.1 Image Captioning

Existing image captioning methods are broadly classified as traditional and deep learning-based methods. The traditional methods encompass retrieval-based and template-based methods. Retrieval-based methods involve selecting the most similar image from a given database and using its corresponding caption as the final generated description. Template-based methods, on the other hand, generate new image captions using a set of predefined grammar and semantic rules that are hard-coded. However, both types of methods rely heavily on complex image feature extraction and devote insufficient attention to optimizing the language model for sentence generation. As a result, they struggle to produce high-quality captions that are both accurate in description and diverse in expression [5,8].

With the development of deep learning, which provides effective solutions for vision and language modeling, deep learning-based methods have been widely adopted in image captioning. Most existing image captioning methods are based on the encoder-decoder framework, using a Convolutional Neural Network (CNN) as the encoder to convert input images into vector representations and either a Recurrent Neural Network (RNN) or another CNN as the decoder to generate descriptive sentences. Kiros et al. [12] first proposed an image captioning method based on the combination of CNN and RNN. However, since the encoder typically represents the input image as a fixed-dimensional vector, it treats all objects in the image as identical and does not dynamically focus on different salient regions when generating words.

To address this issue, the attention mechanism has been introduced into the encoder-decoder framework. This allows different weights to be assigned to input image regions or sequence words, focusing only on the key information needed to generate the next word. This approach enables the generation of more accurate descriptions.

In 2017, Vaswani et al. [13] proposed the Transformer, which departed from traditional CNN and RNN frameworks. The Transformer is entirely based on the attention mechanism, eliminating the need for recurrence and convolution. It uses the self-attention mechanism to model global dependencies between different inputs, while multi-head attention replaces the recurrent layer commonly used in the encoder-decoder framework. Inspired by the great success of the Transformer in natural language processing, Transformer-based methods have been widely studied due to their superior capacity for parallel training and excellent performance. For example, Li et al. [14] proposed a Transformer-based model that includes only attention and feed-forward layers. The model employs a bilateral gating mechanism to control the propagation of visual and semantic information. To capture higher-order interactions, Pan et al. [15] designed an X-Linear Attention Block based on bilinear pooling. This block uses spatial and channel-wise bilinear attention features to capture higher-order interactions between input features. Cornia et al. [16] introduced a Meshed-Memory Transformer, which models relationships between regions and incorporates prior knowledge. Additionally, they designed mesh-like connectivity for decoding. Dubey et al. [17] proposed a label-attention transformer with geometrically coherent objects. They use a deep neural network (DNN) to acquire proposals of geometrically coherent objects and generate captions by investigating their relationships using the Label Attention Mechanism (LAM). Fang et al. [18] proposed a pure vision transformer-based model, which introduces a novel Concept Token Network (CTN) to predict the semantic concepts and then incorporate them into the end-to-end captioning. Zeng et al. [19] proposed a novel Progressive Tree-Structured Prototype Network (PTSN), which is the first attempt to narrow down the scope of predicted words by modeling hierarchical textual semantics. Ge et al. [20] proposed a double decoding transformer to correct wrong words and thereby improve caption quality. Yang et al. [21] proposed an innovative variational transformer framework to boost diversity while maintaining high accuracy. Zhang et al. [22] proposed an end-to-end adaptive semantic-enhanced transformer (AS-Transformer) to produce more accurate semantic guiding information and further optimize the decoding process. The aforementioned studies have demonstrated the superior performance of Transformer-based methods in the image captioning task. Therefore, our method is also based on the Transformer model. However, existing Transformer methods still face challenges in effectively capturing local correlations and semantic noise during the decoding stage. To address these issues, we have designed a novel encoder, LREE, and a novel decoder, DMGD.

2.2 Multi-Layer Perceptron

Benefiting from inductive bias, CNNs have achieved good performance in computer vision tasks. Although CNNs are easy to train and have fewer parameters, their learning ability on large-scale training sets is relatively weak. Compared with CNNs, MLPs demonstrate stronger feature extraction capabilities without relying on inductive bias. Due to their simple structure and high computational efficiency, MLPs have become a hot topic in the field of computer vision.

Existing visual MLP models demonstrate strong capabilities in extracting local features from images, using fully connected layers with fixed weights to aggregate different regions of the input image. Tolstikhin et al. [23] proposed the MLP-Mixer network, a visual backbone network entirely based on a multi-layer perceptron architecture. Research results show that this network can achieve excellent performance using only simple operations. Liu et al. [24] proposed a new full MLP model, which replaces the self-attention mechanism in some visual task models with an external attention mechanism, solving the problem of ignoring the external correlations between multiple samples. For the first time, Li et al. [25] combined a spiking neural network module with an MLP model to optimize the local feature communication mechanism of the MLP model, significantly enhancing its local feature communication ability. The widespread application of the MLP model provides a new feasible solution for optimizing image feature encoding. While MLPs lack the spatial inductive bias inherent to CNNs, their flexibility in modeling irregular semantic relationships and computational efficiency make them better suited for captioning tasks. In our framework, the LGFM bridges this gap by integrating global spatial context from self-attention, ensuring robust handling of both local semantics and global geometry. Therefore, we use MLP to construct LCPM to capture the local correlation information between visual features in different regions, enabling the model to effectively leverage the local semantic relationships within the image.

3 Proposed Method

During the image captioning process, it is crucial not only to focus on the global relationships of the entire image but also to consider the local relationships between different visual regions as an important reference. By deeply processing these relationships, the model can effectively integrate global and local correlation information, thereby generating more accurate and richer language descriptions. To this end, we propose the LREGT model, which is based on an encoder-decoder framework and features a specialized encoder, LREE, and a decoder, DMGD. This model is designed to capture both global and local correlations, ensuring that the generated captions reflect a comprehensive understanding of the image content. The specific approach of this paper is as follows: In the encoding stage, we propose the LREM. This module captures local correlation features through the LCPM and global correlation features through the self-attention mechanism. It then employs the LGFM to model the non-linear relationships between global and local correlation features, thereby obtaining comprehensive correlation features. In the decoding stage, we propose the DMGD, which adopts a progressive model structure to design the DLGM. This mechanism uses dual-level weighting of relationship features to enable the decoder to obtain decoding features with clear semantic relationships and rich semantic information. It also effectively addresses semantic noise issues between cross-domain features through adaptive fusion. The overall structure of the LREGT model is shown in Fig. 1.

3.1 Local Relationship Enhanced Encoder

In the image captioning task, the role of the encoder is to refine the visual representation of the image. Therefore, its ability to capture visual semantic information is crucial for the task. To enable the encoder to capture richer semantic information from the input image and address the issue of previous methods ignoring the potential correlations between different visual objects, we designed a novel encoder, LREE. Specifically, we introduced LREM. In this module, we retained the strategy of using self-attention to extract global correlations, which is commonly used in previous methods. Additionally, we incorporated the LCPM to perceive local correlations. Finally, we integrated these features using the LGFM. This design allows the model to comprehensively capture both global and local semantic correlations in the image and model their non-linear relationships. As a result, the encoder provides higher-quality visual encodings for the decoding process, enabling more accurate and richer language descriptions.

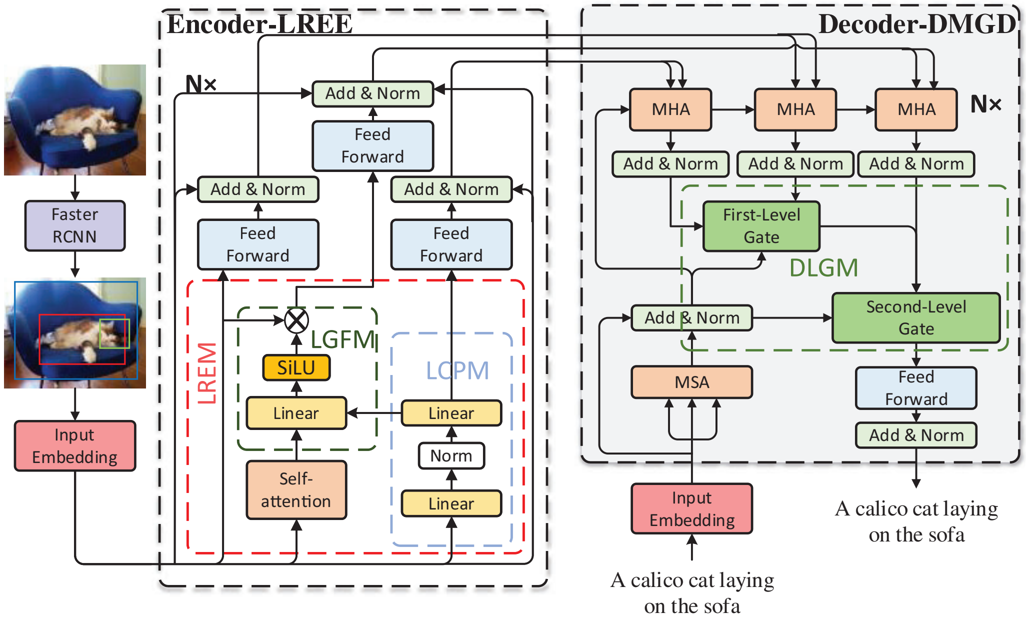


Figure 1: The overall architecture diagram of the LREGT model

3.1.1 Local Relationship Enhanced Module

To ensure that the image captioning model not only focuses on the global correlation information of the whole image but also captures the local correlation information between different visual regions within the image. We propose the LREM, which is mainly composed of the LCPM and the LGFM. The LCPM adopts a fully MLP structure, which is constructed by stacking two linear layers. Because the parameter weights of the MLP can be shared through the local feature-sharing mechanism [25], the weights of the MLP module can be globally shared during training. This significantly enhances the ability of the model to capture the potential correlations between different visual objects in the image. As a result, the LCPM module is capable of capturing local correlation information between visual features from different regions, enabling the model to effectively leverage the local semantic relationships within the image. The LCPM module is not only structurally simple but also effectively realizes local semantic correlations. The specific steps are as follows:

First, the Local Correlation Perception Module (LCPM) is employed to extract local correlation information from the set of visual region features extracted by Faster-RCNN, denoted as $VF = \{v_1, v_2, \ldots, v_N\}$. The operation formula of LCPM is defined as follows:

$$FC\left(v_j\right) = W_i v_j + b_i \tag{1}$$

$$LCPM\left(v_j\right) = FC_1\left(Norm\left(FC_2\left(v_j\right)\right)\right) \tag{2}$$

where v_j represents the j-th regional feature extracted from the input image, FC_i denotes the linear mapping operation of the i-th linear layer, W_i is the parameter weight of the i-th linear layer, b_i is the learnable bias coefficient of the i-th linear layer, and $Norm\left(\right)$ is the L1 regularization operation. The first linear layer projects regional features into a latent space, while the second reconstructs them with shared weights. This forces the MLP to learn translation-invariant correlations, akin to human perception of object relationships regardless of spatial positions.

Subsequently, the self-attention mechanism is employed as the global correlation extraction module. The self-attention mechanism models the relationships between global and local information through attention-based dot-product operations. This process endows the local information of the image with global relevance, enabling the model to capture global relationships within the image. This is beneficial for exploring the global semantic relationships of the image. The operation formula for the self-attention mechanism is defined as follows:

$$Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q is the query vector matrix with dimension n_q, and both K and V are the key and value vector matrices with dimension n_k, respectively. $\frac{1}{\sqrt{d_k}}$ is a scaling factor, where d_k is the dimension of the key vectors. Its purpose is to stabilize the distribution of attention scores and mitigate gradient instability during training.

Finally, to better realize the semantic relationship between global and local correlation information, we design LGFM to encapsulate the function of LCPM and the self-attention mechanism. The design principle is as follows: First, we aggregate the extracted global and local correlation features to obtain the full correlated features. Then we send the fully correlated features into the linear layer with the SiLU activation function for linear mapping, projecting the features into an abstract space. Finally, a gating operation dynamically weights the features in this abstract space, enabling the model to learn the non-linear relationships between different correlation features. The specific representation is as follows:

$$LGFM = att \otimes SiLU\left(W\left[att, vrem\right] + b\right) \tag{4}$$

$$att = Attention\left(v_j, v_j, v_j\right) \tag{5}$$

$$vrem = LCPM\left(v_j\right) \tag{6}$$

where att represents the global correlation feature extracted through the self-attention mechanism, $vrem$ represents the local correlation feature extracted through the LCPM, $SiLU$ denotes the sigmoid-weighted linear unit activation function, [,] represents the feature connection operation, ⊗ is matrix multiplication.

3.1.2 Encoder Layer

Unlike CNNs, which have a relatively small parameter scale, gated attention units typically have a larger parameter scale. As a result, gated attention units can only perform shallow stacking and cannot achieve the deep stacking of dozens or even hundreds of layers like CNNs. However, shallow network layers often lead to weaker model-fitting capabilities. The larger parameter scale of gated attention units arises from auxiliary gating operations, which compound with network depth. This increases memory consumption and complicates gradient flow, empirically limiting practical stacking depth compared to lightweight modules like CNNs. Gated attention mechanisms, unlike standard self-attention, require additional parameters for dynamic feature weighting. This increased parameter scale per layer limits deep stacking due to memory constraints and optimization difficulties, as discussed in studies on parameter-efficient attention designs. To address this issue, we introduce a linear layer structure and regularization operations [26]. The steps are as follows: (1) First, we introduce a linear layer structure to increase network depth and provide a rich set of trainable parameters to enhance the model's fitting capabilities. (2) Then, we apply regularization operations to reduce the model's sensitivity to parameters [27], which effectively mitigates the overfitting phenomenon. Each encoder layer of the LREGT consists of two components: the Local Relationship Enhanced Module (LREM) and the Feed Forward Network (FFN). Based on the above theory, LREM is embedded into a Transformer-based encoder layer structure. This structure uses the feature set output by LREM as the input to the linear layer structure of the FFN, which includes regularization operations. The FFN enhances the model's fitting capabilities by applying non-linear affine transformations to each element of the feature set. The operation formula for FFN is defined as follows:

$$FFN\left(x\right) = W_{F_1}relu\left(W_{F_2}x + c_1\right) + c_2 \tag{7}$$

where $relu$ represents the ReLU activation function operation, W_{F_i} is the learnable weight parameter of the i-th linear layer in FFN, c_1 and c_2 are is the bias term coefficients.

Overall, the encoder layer structurally encapsulates the LREM and FFN layers through residual connections and L1 regularization. The operation formula for the encoder layer is defined as follows:

$$f = Norm\left(v + LREM\left(v\right)\right) \tag{8}$$

$$E = Norm\left(f + FFN\left(f\right)\right) \tag{9}$$

where $Norm\left(\right)$ is an L1 regularization operation.

3.2 Dual-Level Multi-Branch Gated Decoder

The decoder typically employs the output features from the encoder to achieve semantic alignment between visual and textual features by exploring cross-modal semantic relationships. This enables the model to generate natural language descriptions based on the input image. However, since the global and local relational features extracted by the encoder focus on different semantic information, this inevitably leads to semantic noise between them. Directly using the decoder structure to fuse these global and local relational features would further amplify the semantic noise, thereby affecting the accuracy of the generated text descriptions. To address this issue, we propose DMGD. Firstly, it generates multi-perspective contextual feature representations by setting up multiple decoding branches. Secondly, we design DLGM with a hierarchical structure to perform multi-level relationship modeling on multi-correlation features. This allows the decoder to obtain decoding features with clear semantic relationships and rich semantic information, effectively solving the noise problem that may arise between visually relational features with different correlations during the decoding stage. The specific detailed steps are as follows:

Firstly, multiple decoding branches with Multi-Head Self-Attention (MHA) are set up to model the semantic relationships of the visual encoding features extracted by the encoder, which encompass various semantic characteristics, thereby generating multi-perspective contextual feature representations. Meanwhile, Masked Self-Attention (MSA) based on the Transformer decoder is employed to model the semantic relationships of the text features, to obtain multi-referential decoding features for guiding text generation. The specific representation is as follows:

$$MHA\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

$$MSA\left(Q, K, V\right) = \text{softmax}\left(mask\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)V \tag{11}$$

$$l' = MSA\left(l, l, l\right) \tag{12}$$

$$v' = MHA\left(l', v, v\right) \tag{13}$$

where $mask\left(\right)$ denotes the masking operation, l represents the input text features, Q is the query vector matrix with dimension n_q, and both K and V are the key and value vector matrices with dimension n_k, respectively. d_k is a scaling factor.

Secondly, due to the semantic noise between global and local correlation features, directly using a multi-level decoder structure to fuse these features would further amplify the semantic noise, thereby affecting the accuracy of the generated text descriptions. To address this issue, we propose DLGM, which enables the decoder to obtain decoding features with clear semantic relationships and rich semantic information through dual-level relationship weighting. This approach effectively solves the problem of semantic noise between cross-domain features through weight calculation. The DLGM operates in two stages to progressively refine multi-perspective features while suppressing semantic noise: (1) First-Level Gating: This stage weights the global and local correlation features independently to generate intermediate representations. It addresses coarse-grained noise by emphasizing features with strong semantic coherence. (2) Second-Level Gating: The first-level outputs are combined with the fused global-local features to model fine-grained relationships. This step resolves residual noise by adaptively balancing hierarchical semantics, ensuring precise alignment between visual and textual modalities. Specifically, the first-level gating mechanism is used to jointly weigh the global and local correlation features after the decoding operation. The resulting weighted values serve as the first-level relationship weights to obtain the first-level weighted features. Subsequently, the second-level gating mechanism is employed to weigh the first-level weighted features and the full-correlation features that integrate both local and global information. This process yields decoding features with dual-level relationship weights, thereby achieving higher-level semantic relationship modeling. The formulas for the first-level and second-level gating units are defined as follows:

$$gate1_i\left(v',l'\right) = gelu\left(FC\left(\left[v',l'\right]\right)\right) \otimes l' \tag{14}$$

$$gate2\left(g_i,g_j,l'\right) = \left(g_i + g_j\right) \otimes l' \tag{15}$$

$$g_x = gate1_x\left(v',l'\right) \tag{16}$$

where GELU (Gaussian Error Linear Units) is the activation function, $\otimes$ is matrix multiplication, g_x represents the first-level gating unit of the x-way. The first-level gating computes per-branch weights to filter noisy features, while the second-level dynamically adjusts the contribution of fused features based on cross-modal consistency. This hierarchical approach mimics human cognition, where coarse object recognition precedes fine-grained relationship reasoning.

Finally, the output of the DLJGM is fed into the position feed-forward layer to refine the feature decoding, to better guide the model to generate text description.

3.3 Training Strategy

The LREGT is trained in the same way as mainstream image captioning methods [28]. The training process is mainly divided into two stages: (1) Pre-training using the traditional cross-entropy loss function; and (2) Fine-tuning using a reinforcement learning method based on the reward and punishment mechanism.

During the first stage of pre-training, the model trained using cross-entropy loss predicts the next state based on the current state. The operational formula of the cross-entropy loss function is defined as follows:

$$L_{XE}\left(\theta\right) = -\sum_{t=1}^{T}\log\left(p\left(y_t^* \mid y_{1:t-1}^*\right)\right) \tag{17}$$

where y_t represents the t-th word in the sentence sequence.

In the second stage of fine-tuning, we adopt the reinforcement learning-based SCST method [29]. When calculating the reward score, beam search [30] is used to sample the top-k candidates from the probability distribution at each time step of the decoder. We use the CIDEr-D score as the reward mechanism, and the average value of the reward score is used to define the benchmark reward. The operational formula is as follows:

$$b = \frac{1}{k}\left(\sum_{j}^{k} r\left(y_{1:T}^{j}\right)\right) \tag{18}$$

$$\nabla_\theta L_R\left(\theta\right) = -\frac{1}{k}\sum_{j=1}^{k}\left(r\left(y_{1:T}^{j}\right) - b\right)\nabla_\theta \log p\left(y_{1:T}^{j}\right) \tag{19}$$

where $y_{1:T}^{j}$ is the j-th sampling title, k is the beam size of the beam search process, r defines the baseline score (the reward is calculated using the mean of all rewards), and b is the reward bias coefficient. When predicting, we use beam search again for decoding and maintaining the sequence with the highest prediction probability in the last cluster.

4 Experiments

4.1 Dataset and Evaluation Metrics

Dataset: To validate the effectiveness of our proposed LREGT, we conduct extensive experiments on the MS-COCO dataset, which is the most widely used and the largest benchmark dataset in image captioning. The dataset contains 82,783 training images, 40,504 validation images, and 40,775 testing images. Each image is annotated with at least five captions. For a fair comparison, we divide all those images and their corresponding captions into three pairs of sets, 113,287 for training, 5000 for validation, and 5000 for testing.

Evaluation metrics: To effectively evaluate the performance of the model, we employ a comprehensive set of text description evaluation metrics, including ROUGE [31], BLEU [32], METEOR [33], CIDEr [34], and SPICE [35]. These metrics are used to assess the quality of the image descriptions generated by the model, in accordance with the standard evaluation protocol.

4.2 Implementation Details

In the model training process, we use a pre-trained Faster R-CNN object detection model to extract features from the input image. The extracted regional visual features are then linearly mapped using a fully connected layer. The input and output dimensions d_{model} of each layer in the model are set to 512, and the number of heads in the multi-head attention mechanism is set to 8. During training, we use the Adam optimizer with a batch size of 50. The beam size for the cluster search algorithm is set to 5. In the first stage of cross-entropy pre-training, a standard learning rate is set to across 19 epochs. In the second stage of CIDEr-D optimization, we adjust the learning rate in three stages based on the number of training epochs: (1) Before epoch 29, we use a fixed baseline learning rate of 5×10^{-6}. (2) Between epochs 29 and 40, we adjust the baseline learning rate to 5×10^{-7}. (3) After epoch 40, we apply a composite exponential decay strategy to adjust the learning rate dynamically. The formula for the reinforcement learning rate is as follows:

$$lr = \begin{cases} base\_lr & e <= 19 \\ base\_lr * 0.1 & 19 < e <= 40 \\ d_{model}^{-0.5} * \min\left(e^{-0.5}, e * w^{-1.5}\right) & e > 40 \end{cases} \tag{20}$$

where $base\_lr$ is the baseline learning rate, w is the learning rate scheduling strategy period of the warming cycle, set to 20,000, e is the number of rounds of the current training, $\min()$ is the minimum value calculation function, and d_{model} is the input and output dimension of each layer of the model.

4.3 Quantitative Analysis

4.3.1 Analysis of Module Ablation Experiments

To verify the effectiveness of the proposed LREM and DMGD modules, we conducted a series of experiments using the Transformer model as the baseline. Specifically, we followed a step-by-step approach: (1) Baseline Model: We used the standard Transformer model as the baseline. (2) LREM Integration: The LREM module was used to replace the self-attention mechanism in the encoder of the Transformer model. (3) DMGD Integration: The decoder portion of the baseline Transformer was completely replaced by the DMGD module. (4) Combined Model: Finally, the LREM and DMGD modules were combined to construct the LREE model. These experiments were conducted using both grid and regional features extracted from the MS-COCO dataset, as well as cross-modal features retrieved using CLIP. The results are presented in Tables 1–3.

Table 1: Ablation experiments on grid features of the MS-COCO dataset

LREM	DMGD	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
×	×	80.9	38.9	29.0	58.5	131.2	22.7
×	√	81.0	39.2	28.9	58.8	132.0	22.9
√	×	81.2	39.5	29.1	59.1	132.4	23.0
√	√	81.6	39.9	29.5	59.2	133.4	23.1

Table 2: Ablation experiments on regional features of the MS-COCO dataset

LREM	DMGD	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
×	×	79.1	36.2	27.7	56.9	121.8	20.9
×	√	80.9	38.6	28.5	58.6	127.1	22.7
√	×	81.1	39.1	28.8	58.8	130.5	22.7
√	√	81.5	39.6	29.3	59.0	132.3	22.9

Table 3: Ablation experiments on cross-modal features of the MS-COCO dataset

LREM	DMGD	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
×	×	82.2	40.0	29.9	59.4	137.7	23.3
×	√	82.2	40.4	30.0	59.9	138.4	23.6
√	×	82.5	40.8	29.8	59.7	139.1	23.9
√	√	83.5	41.3	30.6	60.4	140.8	24.1

From the results of the ablation experiments in Tables 1–3, it can be seen that adding the LREM and DMGD modules to the Transformer baseline model significantly improves the model's performance metrics. Specifically: (1) Using only DMGD: The BLEU-4 scores increased by 0.3, 2.4, and 0.4, respectively. The ROUGE-L scores increased by 0.3, 1.7, and 0.5, respectively. The CIDEr scores increased by 0.8, 5.3, and 0.7, respectively. These improvements are primarily due to the DMGD module's ability to utilize multiple visual features, providing richer reference information to guide text generation. (2) Using only LREM: The BLEU-4 scores increased by 0.6, 2.9, and 0.8, respectively. The ROUGE-L scores increased by 0.6, 1.9, and 0.3, respectively. The CIDEr scores increased by 1.2, 8.7, and 1.4, respectively. These enhancements are mainly attributed to the encoder's ability to extract both global and local semantic relationships from the image, providing more comprehensive semantic visual encoding for the decoder. This plays a key role in improving model performance. (3) Integrating LREM and DMGD: The combined model achieves excellent performance, outperforming the individual modules.

4.3.2 Ablation Experiments on LREM and DMGD Components

To provide a more comprehensive understanding of each component's contribution to the LREGT model, we have expanded the ablation experiments. In addition to the original experiments that validated the effectiveness of the LREM and DMGD modules, we now present an in-depth analysis of their internal components. For the LREM, we have evaluated the impact of the Local Correlation Perception Module (LCPM) and the Local-Global Fusion Module (LGFM) individually. The LCPM demonstrates a significant improvement in capturing local correlations within images, enhancing the model's ability to describe detailed interactions between different visual elements. The LGFM effectively fuses local and global features, leading to a more comprehensive feature representation and a notable increase in the accuracy and richness of the generated captions. Regarding the DMGD, we have assessed the contribution of the Dual-Level Gating Mechanism (DLGM) through a two-stage approach. The first-level gating mechanism significantly reduces semantic noise by adaptively weighting global and local features, improving the clarity of the generated text descriptions. The second-level gating further refines the feature representations by incorporating cross-modal information, resulting in more accurate and diverse captions. The results of these additional ablation experiments are shown in Table 4:

Table 4: Ablation experiments on LREM and DMGD components

Component	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Transformer [13]	80.9	38.9	29.0	58.5	131.2	22.7
+LCPM	81.2	39.5	29.1	59.1	132.4	23.0
+LGFM	81.5	39.9	29.5	59.2	133.4	23.1
+First-level DLGM	81.8	40.2	29.8	59.5	134.5	23.4
+Second-level DLGM	82.1	40.6	30.1	59.8	135.6	23.7
LREGT (Our)	83.5	41.3	30.6	60.4	140.8	24.1

In Table 4, the ablation experiments demonstrate the incremental improvements contributed by each component of the LREGT model. The LCPM and LGFM in the LREM enhance the model's ability to capture and integrate local and global visual information. In the DMGD, the two-stage DLGM progressively refines the decoding features, effectively reducing semantic noise and improving the accuracy and diversity of the generated captions. These results confirm the effectiveness of our proposed components and their synergistic contribution to the model's overall performance.

4.3.3 Comparison and Analysis with Advanced Baseline Models

To demonstrate that the high performance of the LREGT model is not dependent on the grid and regional features extracted by the Faster R-CNN object detector or the cross-modal features extracted by CLIP, we conducted comparative experiments with several state-of-the-art baseline models. These experiments were performed on both the grid and regional visual features, as well as the cross-modal features of the MS-COCO dataset. To ensure the fairness of the experimental settings, we standardized the input and output dimensions d_{model} of all participating models to 512 and set the number of training epochs to 50. The experimental results are presented in Tables 5–7. These results show that, under the same feature and parameter configurations, the LREGT model achieves better performance compared to other advanced baseline methods.

Table 5: Comparison experiments with advanced baseline models on grid features of the MS-COCO dataset

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Up-Down [30]	75.0	37.3	28.1	57.9	123.8	21.6
Transformer [13]	80.9	38.9	29.0	58.5	131.2	22.7
X-Transformer [15]	81.0	39.5	29.1	59.0	130.2	22.8
M2-Transformer [16]	80.9	38.9	29.1	58.5	131.8	22.7
LREGT (Our)	81.6	39.9	29.5	59.2	133.4	23.1

Table 6: Comparison experiments with advanced baseline models on regional features of the MS-COCO dataset

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Up-Down [30]	79.8	36.3	27.7	56.9	120.1	21.4
Transformer [13]	79.1	36.2	27.7	56.9	121.8	20.9
AOA-Transformer [36]	80.2	38.9	29.1	58.8	129.8	22.4
M2-Transformer [16]	80.8	39.1	29.1	58.6	131.2	22.6
LREGT (Our)	81.5	39.6	29.3	59.0	132.3	22.9

Table 7: Comparison experiments with advanced baseline models on cross-modal features of the MS-COCO dataset

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Transformer [13]	79.8	36.3	27.7	56.9	120.1	21.4
CTX + M^2 [37]	81.5	39.7	30.0	59.5	135.9	23.7
LREGT (Our)	83.5	41.3	30.6	60.4	140.8	24.1

4.3.4 Comparative Analysis with Advanced Models on the MSCOCO Dataset

We used the MS-COCO dataset divided according to the Karpathy split rules as the benchmark and compared the performance of the proposed LREGT model with existing mainstream models. The experimental results are shown in Table 8. The mainstream methods involved in the comparison include Google NIC [38], Soft-Attention [39], SCST [28], RFNet [40], Up-Down [30], GCN-LSTM [41], SGAE [42], ORT [43], AoANet [36], M^2 Transformer [16], X-Transformer [15], RSTNet [44], DGET [45], GAT [46], ViTCAP [18], CATNet [47], MAENet [48], D2Transformer [20], VaT [21], AS-Tranformer [22], LATGeO [17] and CTX + M^2 [37]. It can be observed that our LREGT achieves significant performance improvements across various metrics compared to current state-of-the-art models. Notably, the CIDEr score of LREGT reaches 140.8, which proves the superiority of LREGT. The CIDEr score is widely recognized as the primary measure for assessing the quality of generated captions in image captioning tasks. This indicates that our LREGT generates more accurate captions when describing image content than previous approaches.

Table 8: Performance comparison with state-of-the-art models on the MSCOCO dataset. All values are reported as percentages (%)

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Google NIC [38]	66.6	24.6	–	–	–	–
Soft-Attention [39]	70.7	24.3	23.9	–	–	–
SCST [28]	–	34.2	26.7	55.7	114.0	–
RFNet [40]	79.1	36.5	27.7	57.3	121.9	21.2
Up-Down [30]	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [41]	80.5	38.2	28.5	58.3	127.6	22.0
SGAE [42]	80.8	38.4	28.4	58.6	127.8	22.1
ORT [43]	80.5	38.6	28.7	58.4	128.3	22.6
AoANet [36]	80.2	38.9	29.1	58.8	129.8	22.4
M^2-Transformer [16]	80.8	39.1	29.1	58.6	131.2	22.6
X-Transformer [15]	81.0	39.1	29.1	58.8	130.2	22.8
RSTNet [44]	81.1	39.3	29.3	58.8	133.3	23.0
DGET [45]	81.3	40.3	29.2	59.4	132.4	23.3
GAT [46]	80.8	39.7	29.1	59.0	130.5	22.9
ViTCAP [18]	–	40.1	29.4	59.4	133.1	23.0
CATNet [47]	81.1	39.7	29.5	59.3	133.4	23.5
MAENet [48]	81.3	39.8	29.6	59.1	133.2	23.5
D2 Transformer [20]	80.8	38.9	29.1	58.5	131.8	22.7
VaT [21]	80.9	39.8	29.2	59.0	131.2	23.1
AS-Transformer [22]	80.6	39.3	29.2	58.9	131.0	23.1
LATGeO [17]	81.0	38.8	29.2	58.7	131.7	22.9
CTX+ M^2 [37]	81.5	39.7	30.0	59.5	135.9	23.7
LREGT (Our)	83.5	41.3	30.6	60.4	140.8	24.1

4.3.5 Comparative Analysis with Advanced Models on the Flickr30k Dataset

To further evaluate the generalization capability of GSTNet on a different, we have conducted experiments on the Flickr30k dataset. This dataset, while smaller in scale compared to MSCOCO, offers a distinct testing environment that helps validate the model's performance across diverse scenarios. We have trained and evaluated the LREGT model on the Flickr30k dataset and compared its performance with state-of-the-art methods. The experimental results are shown in Table 9.

In Table 9, the LREGT model outperforms other methods in multiple evaluation metrics, especially achieving a CIDEr score of 68.7%, which is a 5.1% improvement over the previously best-performing model JRAN. This significant enhancement is crucial for the task of image captioning as it directly relates to the quality and relevance of the generated descriptions. Additionally, LREGT demonstrated superior performance in other key metrics such as BLEU-1, BLEU-4, METEOR, and ROUGE-L, further confirming its adaptability and generalization capability across different datasets.

Table 9: Comparison with state-of-the-art on the Flickr30k dataset

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
ALT-ALTM [49]	68.5	27.0	21.2	48.0	56.2
ARL [50]	69.8	27.7	21.5	48.5	57.4
Trans + KG [51]	68.3	26.5	21.7	–	57.7
LSTNet [10]	67.4	24.3	21.5	44.8	63.6
JRAN [52]	71.3	28.3	25.3	53.5	58.2
LREGT (Our)	73.6	31.0	23.2	51.4	68.7

4.4 Qualitative Analysis

4.4.1 Visualization of Visual Attention

To better qualitatively evaluate the visual representation effect of the LREM module, we visualize the contribution of each visual feature in the image to the model's output, as shown in Fig. 2. Fig. 2 displays the visualization effect and attention heatmap of the LREGT model. The image annotations include the words generated step-by-step and their corresponding probability coefficients. For each generated word, we highlight the top 3 attention regions (colored red, blue, and green, respectively), which have the highest attention weights in the headline.

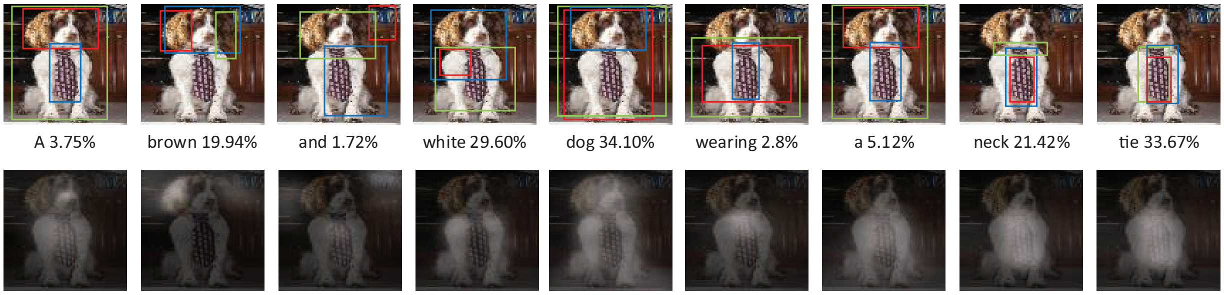
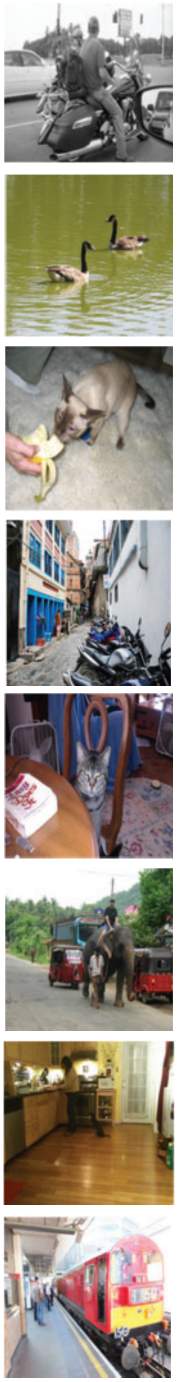


Figure 2: Visualization of the effect and attention heatmaps

4.4.2 Example of Generated Descriptions

To visually demonstrate the accuracy and vividness of the text descriptions generated by our proposed model, Fig. 3 presents multiple sets of comparisons between the descriptions generated by the LREGT model and other state-of-the-art models for the same images. It can be intuitively seen that the text descriptions generated by the LREGT model are not only highly accurate but also more vivid and diverse compared to those generated by the other state-of-the-art methods. From the comparison of the third image in Fig. 3, the LREGT model effectively captures both the global and local relationships within the image. It not only focuses on the main relationship and behavior of the primary object ("cat eating") but also considers the influence of other objects ("person") on the "eating" behavior during text generation. By leveraging local correlations, our model describes the background of the "cat" and its relationship with the "person", leading to the generation of the descriptive phrase "a house Cat eating a banana with a person feeding it on the floor". In contrast, other models fail to capture such detailed and context-rich descriptions. From the comparison of the fourth image in Fig. 3, LREGT captures the arrangement of motorcycles as a "row" and their precise location on the "side of a street", offering a clearer and more structured caption. Other models either inaccurately place the motorcycle or fail to capture the precise arrangement and location. These comparisons clearly illustrate that the LREGT model has a significant advantage in capturing both global and local relationships in images, making the generated descriptions not only accurate but also more vivid and diverse.



Transformer: photo of two men on motorcycles.
Up-Down: two people riding motorcycles on a street.
AOANet: two men on motorcycles at an intersection.
M^2-Transformer: two men riding motorcycles at a traffic light.
LREGT: black and white photo of two men on motorcycles at a stop light.

Transformer: two ducks are swimming in the water.
Up-Down: two ducks in a body of water.
AOANet: ducks swimming on a pond.
M^2-Transformer: two ducks floating on a green pond.
LREGT: two ducks floating together on the green-colored pond.

Transformer: a cat that is eating a banana.
Up-Down: a cat eating something from a person's hand.
AOANet: a cat being fed by a person.
M^2-Transformer: a cat eating a banana held by a person.
LREGT: a house Cat eating a banana with a person feeding it on the floor.

Transformer: a motorcycle parked in front of a street.
Up-Down: motorcycles parked along a street.
AOANet: several motorcycles near a building.
M^2-Transformer: motorcycles lined up on a narrow street.
LREGT: a row of motorcycles parked on the side of a street.

Transformer: a cat sitting on a wooden table.
Up-Down: a cat sitting on furniture near a table.
AOANet: a cat on a chair beside a table.
M^2-Transformer: a gray cat sitting on a chair next to a table.
LREGT: a gray tiger cat sitting on a chair next to a wooden table.

Transformer: a man walking next to an elephant on a street.
Up-Down: an elephant with people on its back on a road.
AOANet: a man riding an elephant with another person alongside.
M^2-Transformer: two men atop an elephant walking on a street.
LREGT: two men riding on the back of an elephant down a street.

Transformer: a dog standing on the floor.
Up-Down: a woman in a kitchen with a pet.
AOANet: a person standing in a kitchen near a dog.
M^2-Transformer: a woman in a kitchen with a dog on the floor.
LREGT: a woman standing in a kitchen with a little dog.

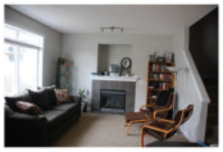
Transformer: a train stop on a station.
Up-Down: a red train on a platform with passengers.
AOANet: a train arriving at a station with people around.
M^2-Transformer: a red train stationed at a platform with passengers waiting.
LREGT: a red train pulling into a train station with people.

Figure 3: Comparison of description generation examples between LREGT and other state-of-the-art models

4.4.3 Failure Cases and Limitations

While LREGT demonstrates strong performance, we analyze scenarios where it underperforms to provide a holistic perspective. Below are representative failure cases and their potential causes:

Failure Case 1: Indoor Living Room Scene

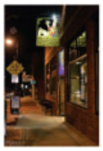


Ground Truth: A modern living room with a black leather sofa, marble coffee table, vertically aligned books on a shelf, and a tile fireplace.

LREGT Output: A living room with furniture and a fireplace.

In this image of a living room with furniture such as a sofa, coffee table, bookshelf, and a fireplace, our model struggled to accurately describe the finer details and relative positions of the objects. For instance, it failed to capture the specific style of the coffee table or the exact arrangement of the books on the shelf. We believe this is due to the complex layout and the numerous objects present in the scene, which may have overwhelmed the model's ability to discern and describe all elements precisely. The model's generated caption might have been too generic, such as "a living room with furniture and a fireplace", without mentioning the specific types or the aesthetic details of the furniture.

Failure Case 2: Nighttime Street Scene



Ground Truth: A cobblestone sidewalk lit by vintage streetlights leads to "Fisher King Winery", where neon signs glow above outdoor tables.

LREGT Output: A street at night with shops and lights.

In this nighttime street scene with shops, signs, and streetlights, the model had difficulty generating a comprehensive and accurate description. It might have missed some of the smaller details like the exact content of the shop signs or the specific types of streetlights. The low-light conditions and the resulting lower image clarity could have contributed to the model's inability to fully capture the scene's details. The generated caption might have been something like "a street at night with shops and lights", lacking specifics about the shops' appearances or the street's atmosphere.

4.5 Computational Complexity and Training Efficiency

To evaluate the practical applicability of LREGT, we compare its computational complexity and training efficiency with key baseline models, including the standard Transformer [13], M^2-Transformer [16], and X-Transformer [15]. All experiments were conducted on an NVIDIA A100 GPU with consistent hyperparameter settings. The results are summarized in Table 10.

(1) Parameter Count: LREGT achieves competitive parameter efficiency compared to M^2-Transformer and X-Transformer. This is attributed to the lightweight MLP-based LCPM module and shared-weight gating mechanisms in DLGM. (2) Inference Time: LREGT achieves the fastest inference time, outperforming M^2-Transformer and X-Transformer. The MLP's parallel computation and reduced dependency on multi-head

attention operations contribute to this improvement. (3) Training Time: LREGT requires only 32.5 h per 50 epochs, demonstrating faster convergence due to the simplified feature fusion in LGFM and dual-level gating. (4) FLOPs Reduction: With 197 GFLOPs, LREGT reduces computational overhead compared to M^2-Transformer and X-Transformer.

Table 10: Computational efficiency comparison on MSCOCO dataset

Model	Params (M)	Inference time (ms)	Training time (50 epochs)	FLOPs (G)
Transformer [13]	166	254	30.2 h	189
M^2-Transformer [16]	191	500	33.1 h	214
X-Transformer [15]	210	276	35.8 h	265
LREGT (Ours)	185	267	32.5 h	197

5 Limitations and Future Work

While the LREGT model demonstrates superior performance on the MSCOCO benchmark, we acknowledge several limitations and outline future directions for improvement that warrant further investigation:

1. Computational Complexity

The integration of the LREM and DMGD modules introduces additional parameters and computational overhead compared to standard Transformer architectures. Specifically: The LREM's MLP-based Local Correlation Perception Module (LCPM) and multi-branch decoding in DMGD increase memory usage during training. The dual-level gating mechanism (DLGM) requires iterative feature weighting, which marginally prolongs inference time. However, we mitigated these costs through parameter sharing in the LCPM and efficient parallelization of the DMGD branches. Despite this, deploying LREGT on resource-constrained devices (e.g., mobile platforms) remains challenging, necessitating future work on lightweight variants or pruning techniques.

2. Adaptability to Diverse Datasets

Our experiments are currently limited to the MSCOCO dataset, which is large-scale and well-annotated. The model's performance on smaller datasets (e.g., Flickr30k) or domain-specific datasets (e.g., medical imaging) remains untested. Variations in data distribution, annotation quality, or object scales may affect robustness. Future studies will explore cross-dataset generalization and few-shot learning strategies to enhance adaptability.

3. Training Efficiency

The two-stage training strategy (cross-entropy pre-training + CIDEr-D optimization) requires significant computational resources and time. While this is common in image captioning, optimizing the training pipeline (e.g., via mixed-precision training or early stopping) could reduce costs.

4. Interpretability and Robustness

Although attention heatmaps (Section 4.4.1) provide partial insights into the model's focus, the interplay between global and local features in the LGFM and DLGM remains complex. Additionally, the model's robustness to adversarial perturbations or noisy inputs (e.g., blurred images) is unexplored. Future work will incorporate explainability tools (e.g., gradient-based attribution) and adversarial training to address these gaps.

5. Handling Rare or Complex Scenes

The model may struggle with images containing rare object interactions or abstract concepts (e.g., metaphors in captions). Enriching the training data with synthetic scenes or leveraging external knowledge bases could improve performance in such scenarios.

6 Conclusion

Given that the thinking model of human beings in visual description is often not limited to the global relationship information of the image, but fully utilizes the integration of the local and global relationships between different visual objects in the image, we propose LREGT, a novel Local Relationship Enhanced Gated Transformer for image captioning. LREGT can capture both global semantic relationships and local potential relationships, and aggregate these relationships through a dual-level gating mechanism. Specifically, in the encoding phase, it introduces a semantic modeling branch based on the MLP to perform additional perception of local associations. In the decoding phase, it sets up multi-branch decoding to obtain multi-angle contextual feature representations and refines its intrinsic semantic representation through a dual-level gating mechanism. LREGT ultimately generates high-quality and semantically rich image captions. The experimental results show that our proposed method achieves excellent performance on various evaluation metrics for image captioning.

1. Societal Impact and Practical Applications

The LREGT model has significant potential to enhance accessibility and human-computer interaction. For instance: (1) Assistive Technologies: LREGT can empower visually impaired individuals by providing real-time, context-aware image descriptions, enabling independent navigation of visual content in scenarios such as online education, social media, or public spaces. (2) Healthcare: In medical imaging, LREGT could assist clinicians by generating preliminary diagnostic reports from radiology scans, reducing interpretation time while maintaining accuracy. (3) Education: The model could be integrated into e-learning platforms to automatically generate descriptive captions for educational materials, aiding students in understanding complex diagrams or illustrations.

2. Ethical Considerations

While LREGT advances image captioning capabilities, its deployment requires careful ethical scrutiny: (1) Bias Mitigation: Like many AI systems, LREGT may inherit biases from training data (e.g., MSCOCO's geographic and cultural skew), potentially leading to stereotypical or inaccurate captions for underrepresented groups. Future work should incorporate fairness-aware training strategies and diversify training datasets to mitigate such biases. (2) Privacy and Consent: When applied to user-generated content, the model must respect privacy norms. For example, captions describing sensitive scenes (e.g., personal photos) should adhere to strict consent protocols to avoid misuse.

Acknowledgement: The authors sincerely acknowledge the support from Guangxi Key Lab of Image and Graphic Intelligent Processing.

Funding Statement: This work is supported by the Natural Science Foundation of China (62473105, 62172118), Nature Science Key Foundation of Guangxi (2021GXNSFDA196002), in part by the Guangxi Key Laboratory of Image and Graphic Intelligent Processing under Grants (GIIP2302, GIIP2303, GIIP2304), Innovation Project of Guang Xi Graduate Education (2024YCXB09, 2024YCXS039).

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design, Yuting He; Data collection, Yuting He and Zetao Jiang; Analysis and interpretation of results, Yuting He; Draft manuscript preparation, Yuting He and Zetao Jiang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated during and/or analysed during the current study are available in the MSCOCO repository, http://images.cocodataset.org/zips/train2014.zip (accessed on 10 June 2024).

Ethics Approval: Not applicable.

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

References

1. Sharma D, Dhiman C, Kumar D. Evolution of visual data captioning methods, datasets, and evaluation metrics: a comprehensive survey. Expert Syst Appl. 2023;221(7):119773. doi:10.1016/j.eswa.2023.119773.
2. Abdar M, Kollati M, Kuraparthi S, Pourpanah F, McDuff D, Ghavamzadeh M, et al. A review of deep learning for video captioning. IEEE Trans Pattern Anal Mach Intell. 2024;2024:1–20. doi:10.1109/tpami.2024.3522295.
3. Zohourianshahzadi Z, Kalita JK. Neural attention for image captioning: review of outstanding methods. Artif Intell Rev. 2022;55(5):3833–62. doi:10.1007/s10462-021-10092-2.
4. Reale-Nosei G, Amador-Domínguez E, Serrano E. From vision to text: a comprehensive review of natural image captioning in medical diagnosis and radiology report generation. Med Image Anal. 2024;97(2):103264. doi:10.1016/j.media.2024.103264.
5. Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G, Cucchiara R. From show to tell: a survey on deep learning-based image captioning. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):539–59. doi:10.1109/tpami.2022.3148210.
6. Nivedita M, Asnath Victy PY. A survey on different deep learning architectures for image captioning. WSEAS Trans Syst Control. 2020;15:635–46. doi:10.37394/23203.2020.15.63.
7. Chen F, Li X, Tang J, Li S, Wang T. A survey on recent advances in image captioning. J Phys Conf Ser. 2021;1914(1):012053. doi:10.1088/1742-6596/1914/1/012053.
8. Sharma H. A survey on image encoders and language models for image captioning. IOP Conf Ser Mater Sci Eng. 2021;1116(1):012118. doi:10.1088/1757-899x/1116/1/012118.
9. Ji J, Huang X, Sun X, Zhou Y, Luo G, Cao L, et al. Multi-branch distance-sensitive self-attention network for image captioning. IEEE Trans Multimed. 2022;25:3962–74. doi:10.1109/TMM.2022.3169061.
10. Ma Y, Ji J, Sun X, Zhou Y, Ji R. Towards local visual modeling for image captioning. Pattern Recognit. 2023;138(6):109420. doi:10.1016/j.patcog.2023.109420.
11. Pham MT, Pham QH, Tran QD, Ho HT, Nguyen LV, Huy DNM, et al. A review on vision-language-based approaches: challenges and applications. Comput Mater Contin. 2025;82(2):1733–56. doi:10.32604/cmc.2025.060363.
12. Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. Proc Mach Learn Res. 2014;32(2):595–603.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
14. Li J, Yao P, Guo L, Zhang W. Boosted transformer for image captioning. Appl Sci. 2019;9(16):3260. doi:10.3390/app9163260.
15. Pan Y, Yao T, Li Y, Mei T. X-linear attention networks for image captioning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. doi:10.1109/cvpr42600.2020.01098.
16. Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. doi:10.1109/cvpr42600.2020.01059.

17. Dubey S, Olimov F, Rafique MA, Kim J, Jeon M. Label-attention transformer with geometrically coherent objects for image captioning. Inf Sci. 2023;623:812–31. doi:10.1016/j.ins.2022.12.018.
18. Fang Z, Wang J, Hu X, Liang L, Gan Z, Wang L, et al. Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.01748.
19. Zeng P, Zhu J, Song J, Gao L. Progressive tree-structured prototype network for end-to-end image captioning. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisboa, Portugal. doi:10.1145/3503161.3548024.
20. Ge G, Han Y, Hao L, Hao K, Wei B, Tang XS. Show, tell and rectify: boost image caption generation via an output rectifier. Neurocomputing. 2024;585(1):127651. doi:10.1016/j.neucom.2024.127651.
21. Yang L, He L, Hu D, Liu Y, Peng Y, Chen H, et al. Variational transformer: a framework beyond the tradeoff between accuracy and diversity for image captioning. IEEE Trans Neural Netw Learn Syst. 2025;36(5):9500–11. doi:10.1109/tnnls.2024.3440872.
22. Zhang J, Fang Z, Sun H, Wang Z. Adaptive semantic-enhanced transformer for image captioning. IEEE Trans Neural Netw Learn Syst. 2024;35(2):1785–96. doi:10.1109/TNNLS.2022.3185320.
23. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, et al. MLP-Mixer: an all-MLP architecture for vision. Adv Neural Inf Process Syst. 2021;34:24261–72.
24. Liu H, Dai Z, So D, Le QV. Pay attention to MLPs. Adv Neural Inf Process Syst. 2021;34:9204–15.
25. Li W, Chen H, Guo J, Zhang Z, Wang Y. Brain-inspired multilayer perceptron with spiking neurons. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.00086.
26. Yang X, Yang Y, Ma S, Li Z, Dong W, Woźniak M. SAMT-generator: a second-attention for image captioning based on multi-stage transformer network. Neurocomputing. 2024;593(4):127823. doi:10.1016/j.neucom.2024.127823.
27. Socher R, Karpathy A, Le QV, Manning CD, Ng AY. Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comput Linguist. 2014;2(8):207–18. doi:10.1162/tacl_a_00177.
28. Daneshfar F, Bartani A, Lotfi P. Image captioning by diffusion models: a survey. Eng Appl Artif Intell. 2024;138(1):109288. doi:10.1016/j.engappai.2024.109288.
29. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. doi:10.1109/CVPR.2017.131.
30. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. doi:10.1109/CVPR.2018.00636.
31. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out; 2004 Jul 25–26; Barcelona, Spain.
32. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; 2002 Jul 7–12; Stroudsburg, PA, USA. doi:10.3115/1073083.1073135.
33. 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 Jun 29; Beijing, China.
34. Vedantam R, Zitnick CL, Parikh D. CIDEr: consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. doi:10.1109/CVPR.2015.7299087.
35. Anderson P, Fernando B, Johnson M, Gould S. SPICE: semantic propositional image caption evaluation. In: Proceedings of the Computer Vision—ECCV 2016; 2016 Oct 11–14; Amsterdam, The Netherlands. doi:10.1007/978-3-319-46454-1_24.

36. Huang L, Wang W, Chen J, Wei XY. Attention on attention for image captioning. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. doi:10.1109/iccv.2019.00473.
37. Kuo CW, Kira Z. Beyond a pre-trained object detector: cross-modal textual and visual context for image captioning. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/CVPR52688.2022.01744.
38. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. doi:10.1109/CVPR.2015.7298935.
39. 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. Proc Mach Learn Res. 2015;37(1):2048–57. doi:10.1088/1742-6596/2589/1/012012.
40. Jiang W, Ma L, Jiang YG, Liu W, Zhang T. Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany.
41. Yao T, Pan Y, Li Y, Mei T. Exploring visual relationship for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. doi:10.1007/978-3-030-01264-9_42.
42. Yang X, Tang K, Zhang H, Cai J. Auto-encoding scene graphs for image captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. doi:10.1109/cvpr.2019.01094.
43. Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: transforming objects into words. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS); 2019 Dec 8–14; Vancouver, BC, Canada.
44. Zhang X, Sun X, Luo Y, Ji J, Zhou Y, Wu Y, et al. RSTNet: captioning with adaptive attention on visual and non-visual words. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. doi:10.1109/CVPR46437.2021.01521.
45. Xian T, Li Z, Zhang C, Ma H. Dual global enhanced transformer for image captioning. Neural Netw. 2022;148(12):129–41. doi:10.1016/j.neunet.2022.01.011.
46. Wang C, Shen Y, Ji L. Geometry Attention Transformer with position-aware LSTMs for image captioning. Expert Syst Appl. 2022;201(4):117174. doi:10.1016/j.eswa.2022.117174.
47. Gao Y, Wang N, Suo W, Sun M, Wang P. Improving image captioning via enhancing dual-side context awareness. In: Proceedings of the 2022 International Conference on Multimedia Retrieval; 2022 Jun 27–30; Newark, NJ, USA. doi:10.1145/3512527.3531379.
48. Hu N, Fan C, Ming Y, Feng F. MAENet: a novel multi-head association attention enhancement network for completing intra-modal interaction in image captioning. Neurocomputing. 2023;519(1):69–81. doi:10.1016/j.neucom.2022.11.045.
49. Ye S, Liu N, Han J. Attentive linear transformation for image captioning. IEEE Trans Image Process. 2018;27(11):5514–24. doi:10.1109/TIP.2018.2855406.
50. Wang J, Wang W, Wang L, Wang Z, Feng DD, Tan T. Learning visual relationship and context-aware attention for image captioning. Pattern Recognit. 2020;98(4):107075. doi:10.1016/j.patcog.2019.107075.
51. Zhang Y, Shi X, Mi S, Yang X. Image captioning with transformer and knowledge graph. Pattern Recognit Lett. 2021;143(6):43–9. doi:10.1016/j.patrec.2020.12.020.
52. Wang C, Gu X. Learning joint relationship attention network for image captioning. Expert Syst Appl. 2023;211(20):118474. doi:10.1016/j.eswa.2022.118474.