SG-DLCT: Saliency-Guided Dual-Level Collaborative Transformer for Image Captioning

Xinchao Zhu[1], Jingzheng Deng[1], Kanjian Zhang[2,1], Jinxia Zhang[2,1]

1. The Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China

2. Southeast University Shenzhen Research Institute, Shenzhen 518057, China

Abstract: Improving the generation of feature representations for images has been a central focus in the field of image captioning. However, existing approaches, whether based on grid features or region features, only consider the features within individual regions without capturing the interactions between objects. In this paper, we propose a novel **S**aliency-**G**uided **D**ual-**L**evel **C**ollaborative **T**ransformer (SG-DLCT), which leverages saliency information to guide the network in capturing the importance of different objects and their interrelationships. Concretely, we utilize a pre-trained RGBD salient object ranking model to obtain saliency information. Building upon this, we propose the Saliency Feature Enhanced Module (SFEM), which optimizes region features using saliency values as weights, thereby guiding the model to focus on image regions that are more meaningful for generating image captions. Additionally, we propose the Saliency-Guided Cross Attention (SGCA), which leverages saliency information to model the interaction between region features and grid features, allowing the final model's features to benefit from the advantages of both types of features. Experimental results validate the significance of saliency information in image captioning tasks and demonstrate the effectiveness of our proposed method.

Key Words: Image Captioning, Salient Object Ranking, Transformer

1 Introduction

Image Captioning aims to utilize natural language to describe the content of an image, serving as an intersection between computer vision and natural language processing domains. During describing an image, human observers first perceive the visual content to comprehend the objects, scenes, and contexts. Subsequently, they establish relationships between objects and select appropriate adjectives and nouns to form coherent and fluent sentences[1]. Understanding image information, encoding textual information, and capturing relationships between objects are crucial elements for image captioning. Drawing inspiration from the sequence-to-sequence models used in machine translation, most existing models[1–7] employ an encoder-decoder framework to model such process. Initially, the encoder extracts image features. Then, the decoder leverages these features to model the relationship with a textual vector and generate a descriptive sentence corresponding to the image.

Vinyals et al.[2] employed the last layer output of convolutional neural networks (CNNs) as the image features. However, such global image features excessively compress the image information, leading to the loss of certain details and fine-grained content. To address this issue, Xu et al.[3] utilized fixed-sized feature maps generated by CNNs as grid features to replace global image features. However, this uniform grid partitioning may introduce the issue of splitting the same object across multiple grids, thereby compromising the integrity of the object information. To better represent object information, Anderson et al.[5] highlighted that region features detected from images can provide a more refined representation of each object. These methods mainly focus on how to better represent features within regions from different perspectives but overlook the relationships between objects and the saliency ranking of objects[8].

When generating image captions, salient object ranking provides crucial prior information, including the emphasis on specific objects and the importance relationship among various objects. This information is vital for the model to generate more accurate, detailed, and human-like descriptions in line with visual perception. As depicted in Fig. 1(a), both the dog and the stop sign are the most salient objects in the image, and they are highlighted in their respective captions. Furthermore, recognizing the saliency ranking of objects and identifying salient regions contributes to the network's modeling of the depicted actions or scenes within the entire image. By integrating the saliency ranking and semantic information of various objects in the scene, key vocabulary to be included in the image captions can be accurately selected. As shown in Fig. 1(b), salient objects such as "two young boys" and "carrots" are missing in the image captions generated by the vanilla Transformer, resulting in a misjudgment of the depicted action as "brush teeth", which greatly deviates from the actual content of the image caption.

Fig. 1: (a) Examples to illustrate the impact of saliency on image captioning. White regions on the right side indicate the salient regions, underlined words indicate the corresponding word. (b) Limitations of capturing salient objects with the vanilla Transformer.

Based on the above analysis, Saliency-Guided Dual-Level

Collaborative Transformer (SG-DLCT) is proposed for image captioning. Specifically, we propose the Saliency Feature Enhanced Module (SFEM), which is based on salient object ranking. SFEM assigns weights to region features according to the saliency levels of objects detected within bounding boxes. SFEM guides the model to prioritize image regions that hold greater significance for generating image captions across different regions of the image.

Furthermore, we redesign the cross attention mechanism and propose the Saliency-Guided Cross Attention (SGCA). SGCA utilizes saliency information to guide the model in capturing the intricate interactions between region features and grid features. By incorporating saliency guidance, the module facilitates the better fusion of these two features.

To sum up, the contributions of this work are as follows:

- We propose the Saliency-Guided Dual-Level Collaborative Transformer (SG-DLCT) to incorporate saliency information into both grid features and region features, enabling the model to generate more accurate and vivid image captions.
- We propose the Saliency Feature Enhanced Module (SFEM) based on salient object ranking. SFEM applies weights to regions based on the saliency information of objects within those regions, directing the model's attention towards salient regions.
- We redesign the cross attention mechanism and propose the Saliency-Guided Cross Attention (SGCA), which leverages saliency guidance to facilitate better fusion of region features and grid features.
- Experimental results demonstrate the effectiveness of saliency in image captioning tasks and validate the competitiveness of our proposed approach.

2 Related Work

Inspired by the sequence-to-sequence models in machine translation, most methods[1–5, 9] adopt an encoder-decoder architecture. In this architecture, the encoder takes the image as input and produces different image features, which are subsequently utilized as inputs for the decoder.

Previous studies[2, 4, 9] employed fixed-size grid features as inputs to the encoder. However, these uniformly divided grids may forcefully separate objects in the image, resulting in incomplete object information. Subsequently, Anderson et al.[5] proposed using region features obtained from object detectors as inputs to the encoder, which effectively improved the performance of image captioning[6, 10–12]. Jiang et al.[13] pointed out that the superior performance of region features is primarily attributed to pre-training on the Visual Genome dataset rather than the features themselves. Luo et al.[7] highlighted the pros and cons of grid features and region features and tried to integrate these two features using the cross attention mechanism. These studies primarily focus on enriching the representation of image features but overlook the core of image captioning task, which is to describe the interactions between objects. However, salient object ranking reflects the importance differences between different objects in an image and serves as an effective tool for facilitating the interaction of features between objects.

3 Methodology

3.1 Model Overview

The overall of our model is illustrated in Fig. 2. Our model comprises three parts: feature extraction, saliency-guided feature interaction, and feature encoding.

In the feature extraction part, given an input RGB image, we first employ a pretrained object detection model to extract region features and grid features. Additionally, we incorporate depth map information and utilize a pretrained RGBD salient object ranking model to predict the saliency levels of individual objects. By calculating the average saliency level within each bounding box, we obtain saliency values for different detection boxes.

In the saliency-guided feature interaction part, we leverage the Saliency Feature Enhanced Module (SFEM) to perform saliency-weighted optimization on different regions based on their corresponding saliency values. Subsequently, we employ the Saliency-Guided Cross Attention (SGCA) to guide the model in modeling the complex interaction between region features and grid features.

Finally, in the feature encoding part, the saliency-guided image feature and text embedding are input into a Transformer decoder to generate the image caption.

3.2 Saliency Feature Enhanced Module (SFEM)

As shown in Fig. 1(a), salient objects are often emphasized in the image captions, indicating that different regions correspond to varying degrees of importance. By assigning higher weights to the features of salient regions, the generated image captions can prioritize regions with higher saliency levels, leading to more accurate outcomes.

Thus, to allocate attention weights to region features based on saliency levels, we propose the Saliency Feature Enhanced Module (SFEM). As illustrated in Fig. 3, given an image I, we initially feed it into a pretrained Region Feature Extraction Network, yielding region features $F = \{f_1, \cdots, f_k\}$ and region bounding boxes $B = \{b_1, \cdots, b_k\}$, where k denotes the number of regions, $f_i \in \mathcal{R}^{d_f}$ represents the feature vector for the i-th region in I, with d_f denoting the feature dimension of f_i. Additionally, $b_i = (x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max})$ denotes the bounding box coordinates of the i-th region, specifying the top-left and bottom-right corners of the bounding box.

Subsequently, to construct the input for the RGBD Salient Object Ranking Network, we utilize ZoeDepth[14] to estimate the depth map D of I. We combine the RGB image I and the depth map D, forming the input $X = [I, D]$, which serves as the input to SOR-PPA[15], yielding the saliency prediction map I'.

Assuming that the coordinates of the top-left and bottom-right corners within the i-th region are (x_0, y_0) and (x_1, y_1), respectively, the saliency level s_i is given by:

$$s_i = \frac{\sum_{p=x_0}^{x_1} \sum_{q=y_0}^{y_1} I'_{p,q}}{(x_1 - x_0)(y_1 - y_0)}, \tag{1}$$

where $I'_{p,q}$ denotes the saliency value of the pixel (p, q) within the i-th region. Next, we normalize the saliency levels

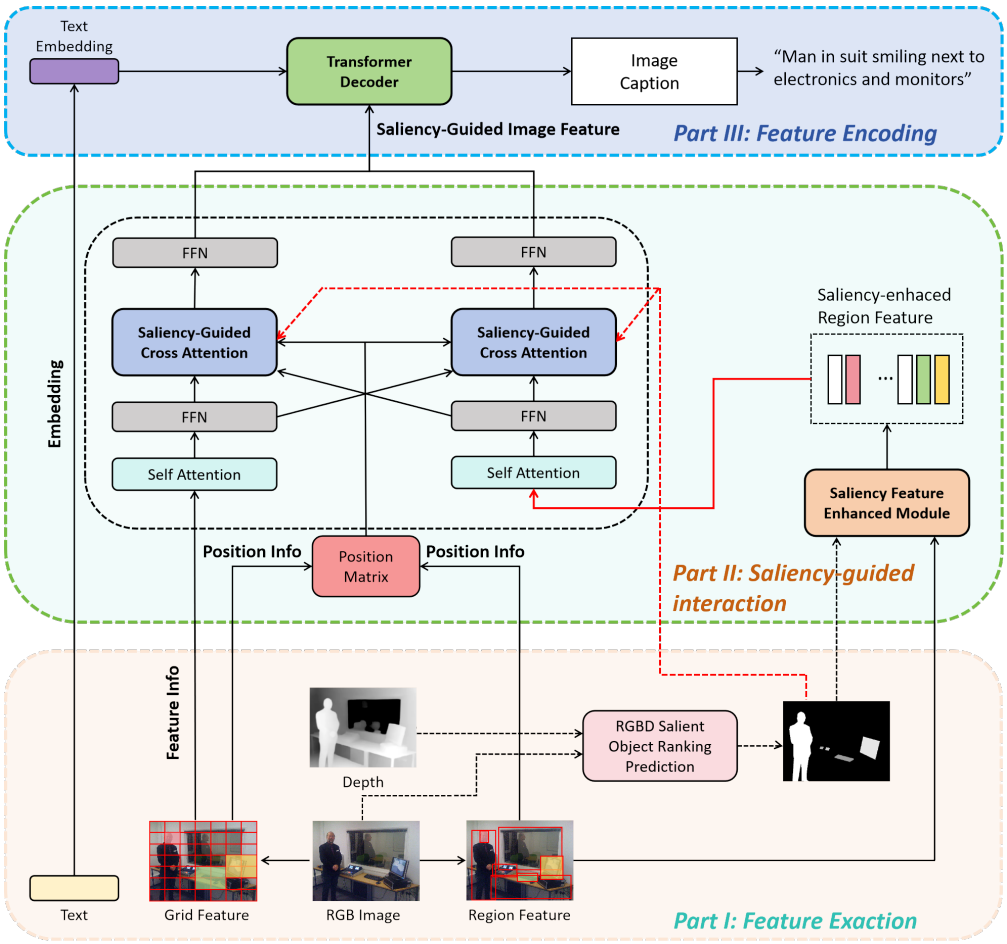

Fig. 2: Overview of our proposed Saliency-Guided Dual-Level Collaborative Transformer architecture for image captioning. The Saliency Feature Enhanced Module (SFEM) is applied to weight the region features based on the saliency features, followed by the Saliency-Guided Cross Attention (SGCA) which facilitates the interaction of grid features and saliency-enhanced region features under the guidance of saliency. Note that the Transformer decoder is the same as the vanilla Transformer's.

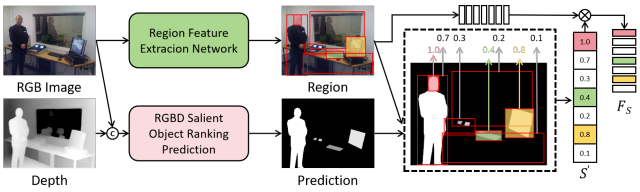

Fig. 3: The architecture of Saliency Feature Enhanced Module.

across all regions, using the formula (2):

$$s'_i = \frac{s_i - min(s)}{max(s) - min(s)} \in [0, 1], \tag{2}$$

where $min(\cdot)$ and $max(\cdot)$ respectively represent the minimum and maximum values among all saliency values from k regions. Then the saliency ranking relationship among the region features can be represented as $S' = \{s'_1, \cdots, s'_k\}$.

Finally, we perform feature enhancement on the region features F based on the saliency ranking relationship S', resulting in saliency-enhanced region features F_S. The process is represented as formula (3):

$$F_S = \{s'_1 f_1, \cdots, s'_k f_k\}. \tag{3}$$

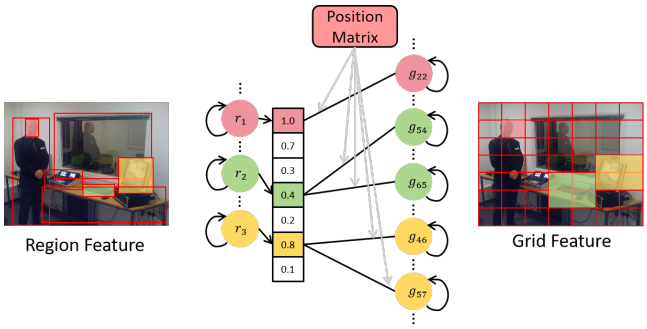

Fig. 4: Illustration of Saliency-Guided Cross Attention.

3.3 Saliency-Guided Cross Attention (SGCA)

We propose the Saliency-Guided Cross Attention (SGCA) to model the complex interaction between region features and grid features, while amplifying the weights of salient regions. To effectively distinguish the importance of different region features, we employ a histogram-based approach to partition the saliency levels of all region features. Specifically, we generate a histogram that encompasses all saliency values of the regions. The x-axis represents the saliency levels, while the y-axis represents the proportion of regions at each saliency value relative to the total number of regions.

Subsequently, we divide the saliency values into 10 intervals proportionally, as indicated by the following formula (4):

$$p_i = \frac{1}{k}\sum_{j=1}^{n}[s_j < s_i], i = 1, 2, \cdots, k, \tag{4}$$

where $[s_j < s_i]$ indicates that it takes the value 1 when $s_j < s_i$, and 0 otherwise. Thus, p_i represents the proportion of regions with saliency values smaller than s_i. Based on the proportions obtained from formula (4) and the histogram, we divide the saliency values into 10 intervals, corresponding to threshold values $t_0, t_1, \cdots, t_9$, where $t_0 = 0$, $t_9 = 1$, and $t_j < t_{j+1}$. Therefore, the m-th interval is defined within the range $[t_{m-1}, t_m)$. Subsequently, we calculate the attention weight for the i-th region based on the aforementioned intervals, as computed as:

$$s_i' = j, where \quad t_{j-1} < s_i < t_j. \tag{5}$$

Additionally, a geometric alignment graph $G = (V, E)$ is created. All region features and grid features are represented as independent nodes, forming a set of feature nodes V. They are connected together to form an undirected graph G only when the bounding boxes of the grid nodes and region nodes have an intersection. Following the above rules, an undirected graph can be conducted, as illustrated in Fig. 4. Based on the geometric alignment graph, we apply SGCA to model the interaction between region features and grid features. Firstly, the positional relationship $L(r, g)$ (corresponding to the position matrix in Fig. 4) between region nodes and grid nodes is calculated based on a Gaussian function:

$$L(r, g) = exp(-\frac{(x_r - x_g)^2 + (y_r - y_g)^2}{2\sigma^2}), \tag{6}$$

where $L(r, g)$ represents the positional relationship between region feature r and grid feature g, (x_r, y_r) and (x_g, y_g) denote the center coordinates of region feature r and grid feature g, respectively. σ is the standard deviation of the Gaussian function and is set to 5 in all experiments.

By considering the saliency weights and the positional relationship, we can obtain the relationship $R(r, g)$ between region feature r and grid feature g:

$$R(r, g) = s_r' \cdot L(r, g). \tag{7}$$

Finally, we utilize the relation matrix R obtained from formula (7) as weights in the conventional cross attention mechanism, our Multi-Head SGCA (MHSGCA) can be formulated as:

$$MHSGCA(Q, K, V) = Concat(head_1, \cdots, head_h)W^O, \tag{8}$$

$$head_i = SGCA(QW_i^Q, KW_i^K, VW_i^V, G, R), \tag{9}$$

$$SGCA(Q, K, V, G, R) = \underset{G}{graph - softmax}(\frac{QK^T \odot R}{\sqrt{d}})V, \tag{10}$$

where $Q, K, V \in \mathcal{R}^{N \times d}$ represent the input query, key, and value, respectively. N is the length of input, d is the hidden dimension in each head. $W^O \in \mathcal{R}^{hd \times d}$ is a learnable matrix for the output of all heads. $\underset{G}{graph - softmax}$ means applying the softmax operation only among adjacent nodes, while assigning a weight of 0 to non-adjacent nodes.

4 Experiments

4.1 Datasets and Metrics

Our experiments are conducted on the benchmark image captioning dataset COCO[16]. The dataset contains 123287 images, each annotated with 5 different captions. It is worth noting that the original COCO only provides corresponding reference captions in the training set and validation set while the test set only contains images. To establish a unified testing standard, we follow the widely adopted Karpathy split[17], where 113287, 5,000 and 5,000 images are used for training, validation, and testing respectively.

Following the standard evaluation criterion, we utilize BLEU[18], METEOR[19], ROUGE[20] and CIDEr[21] to evaluate our model.

4.2 Implementation Details

The three types of features utilized in all our experiments are grid features, region features, and saliency features. The extraction of grid features and region features follows the same approach as described in [7]. To extract saliency features, we initially employ ZoeDepth[14] to obtain depth estimation maps for the images in the COCO dataset. Subsequently, we fine-tune the SOR-PPA[15] on the COCO dataset using both RGB and RGBD input formats respectively. During the feature extraction stage, we utilize ZoeDepth again to acquire the depth estimation map for the input image and concatenate it with the RGB image. This combined input is then fed into SOR-PPA to obtain the predicted saliency map, which serves as saliency features.

In practice, our encoder and decoder both have 3 layers, where each layer uses 8 self-attention heads and the inner dimension is 512. The remaining hyperparameters and training strategies are kept consistent with those described in [7]. The training is conducted on one Nvidia RTX 3090, with a batch size of 50.

4.3 Main Results

4.3.1 Quantitative Comparisons

We compare our method with various state-of-the-art approaches, including MT[24], AOA[23], ORT[10], M2-Transformer[11], X-transformer[22], RSTNet[6] and DLCT[7]. Tab.1 shows the quantitative results.

SG-DLCT outperforms all other models on the BLEU-1 and BLEU-4 metrics, achieving scores of 81.8 and 40.0, respectively. These results indicate the superiority of our method in generating captions that are more similar to the reference captions. In terms of the ROUGE metric, SG-DLCT also exhibits excellent performance, reaching a score of 59.2. The ROUGE metric measures the overlap between the generated and reference captions, and higher ROUGE scores indicate better capturing of key information in the images. This further confirms the significant role of saliency information in improving model performance. Moreover, in the CIDEr metric, SG-DLCT surpasses the highest-scoring DLCT (133.8) by 0.7 points. CIDEr primarily evaluates the

semantic similarity between the generated captions and the reference captions, and higher scores suggest that SG-DLCT better understands the semantic information of the images. It is worth noting that SG-DLCT achieves parity with the highest-scoring DLCT (29.5) on the METEOR metric. This indicates that our method remains competitive in generating captions with reasonable grammatical structures and accurate word choices. The results presented comprehensively demonstrate the competitiveness of our approach in the image captioning task.

Table 1: Quantitative comparisons with state-of-the-art methods. The data presented are all sourced from the original papers. The bold number is the top score.

Model	B-1	B-4	M	R	C
MT[24]	80.8	38.9	28,8	58.7	129.6
AOA[23]	80.7	39.0	28.9	58.7	129.5
ORT[10]	80.5	38.6	28.7	58.4	128.3
M2-Transformer[11]	80.6	38.8	29.0	58.4	130.8
X-transformer[22]	81.0	39.7	29.4	58.9	132.5
RSTNet[6]	81.1	39.3	29.4	58.8	133.3
DLCT[7]	81.4	39.8	**29.5**	59.1	133.8
Ours	**81.8**	**40.0**	**29.5**	**59.2**	**134.5**

4.3.2 Qualitative Comparisons

We show the images and the corresponding generated image captions in Fig. 5. It can be observed that our method, by incorporating saliency factors, is able to capture crucial objects within the images, resulting in more accurate overall captions of the image content. For instance, in the first and third rows, the vanilla Transformer fails to recognize "bus" and "woman", leading to deviations in the generated image captions, while our method generates the correct captions. Furthermore, our model achieves higher precision in capturing image details due to the incorporation of saliency-guided interactions between region features and grid features. For example, in the second row, our method more accurately describes "in a bakery" compared to the vanilla Transformer. The results show that our model can generate more accurate and diverse captions compared to the vanilla Transformer.

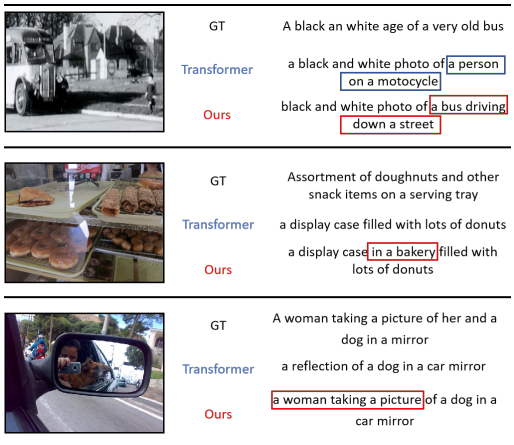

Fig. 5: Examples of captions generated by our approach and the vanilla Transformer.

4.4 Ablation Studies

4.4.1 Effectiveness of Each Component of Our Method

To validate the effectiveness of each component in our method, we conduct ablation studies on SFEM, SGCA, and the inclusion of depth maps in the input, as shown in Tab. 2.

When only SFEM is added, the model exhibits improvements across all evaluation metrics, indicating that the saliency-based feature enhanced module effectively extracts information from salient regions, thereby enhancing the model's focus on key areas. Similarly, when only SGCA is added, the model demonstrates performance improvements in all evaluation metrics. This indicates that the saliency-guided cross attention module facilitates the exploration of complex interaction patterns between region features and grid features, further enriching the generated image captions.

When SFEM and SGCA are added simultaneously, the model exhibits more significant performance improvements across all evaluation metrics, indicating a synergistic effect between the two components. After incorporating depth maps, although there is a slight decrease in the BLEU-1 and BLEU-4 metrics, other metrics show substantial improvements. Overall, the introduction of depth maps leads to further enhancement in the model's performance. Particularly in the CIDEr metric, the score significantly increases to 134.5, indicating that the model better understands the semantic information of the images after incorporating depth maps. These results further validate the effectiveness of each component of SG-DLCT.

Table 2: Ablation studies on different components of our method. The bold number is the top score.

Image	Model	B-1	B-4	M	R	C
RGB	Baseline (ORT[10])	80.5	38.6	28.7	58.4	128.3
	Baseline+SFEM	81.8	39.7	29.1	58.9	130.2
	Baseline+SGCA	82.0	40.1	29.1	59.0	131.5
	Baseline+SFEM+SGCA	**82.1**	**40.3**	29.2	**59.2**	132.0
RGBD	Baseline+SFEM	81.2	39.6	28.9	58.7	132.3
	Baseline+SGCA	81.6	39.7	29.0	58.9	132.4
	Baseline+SFEM+SGCA	81.8	40.0	**29.5**	**59.2**	**134.5**

4.4.2 Different Schemes for Assigning Saliency Weights

We design three schemes for converting saliency maps into saliency weights, which are described as follows:

- $Discretization$: map the mean saliency value within the bounding box to five uniformly distributed discrete intervals ranging from 0 to 1 and get the corresponding saliency weights
- $Normalization$: normalize the saliency values within all bounding boxes to the range of [0,1]
- $HistogramTrans$: obtain the saliency weights by allocating intervals based on the saliency value histogram

To evaluate the performance of the above schemes, we compare the effects of applying SFEM and SGCA separately on the baseline under different schemes, as shown in Tab. 3. Based on the results in Tab. 3, the saliency weight allocation for SFEM follows the $Normalization$ scheme, as described by formula (2), while the saliency weight allocation for SGCA follows the $HistogramTrans$ scheme, as

described by formula (5).

Table 3: Effects of SFEM and SGCA under different saliency weight allocation schemes. The bold number is the top score within the respective module.

Model	Schme	B-1	B-4	M	R	C
Baseline+SFEM	$Discretization$	81.4	**39.8**	28.8	58.7	130.0
	$Normalization$	**81.8**	39.7	**29.1**	**58.9**	130.2
	$HistogramTrans$	81.7	**39.8**	28.8	58.7	129.8
Baseline+SGCA	$Discretization$	81.8	39.7	29.1	58.9	130.2
	$Normalization$	81.7	40.0	29.1	**59.0**	131.0
	$HistogramTrans$	**82.0**	**40.1**	29.1	**59.0**	**131.5**

5 Conclusion

In this paper, we propose a novel Saliency-Guided Dual-Level Collaborative Transformer to effectively capture key information in images. For the region features, we employ Saliency Feature Enhanced Module (SFEM) to assign higher weights to salient regions, guiding the model to focus more on areas that are more meaningful for the generated image captions. Moreover, we leverage Saliency-Guided Cross Attention (SGCA) under the guidance of saliency object ranking to model the interaction between region features and grid features. Extensive experiments validate the effectiveness of our proposed method. In the future, we plan to utilize saliency information to guide other tasks that involve significant differences in importance among different objects.

References

[1] Hu X, Gan Z, Wang J, et al. Scaling up vision-language pre-training for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 17980-17989.

[2] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.

[3] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. PMLR, 2015: 2048-2057.

[4] Lu J, Xiong C, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 375-383.

[5] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.

[6] Zhang X, Sun X, Luo Y, et al. RSTNet: Captioning with adaptive attention on visual and non-visual words[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15465-15474.

[7] Luo Y, Ji J, Sun X, et al. Dual-level collaborative transformer for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 2286-2293.

[8] Wang Y, Xu J, Sun Y. End-to-end transformer based model for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 2585-2594.

[9] You Q, Jin H, Wang Z, et al. Image captioning with semantic attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 4651-4659.

[10] Herdade S, Kappeler A, Boakye K, et al. Image captioning: Transforming objects into words[J]. Advances in Neural Information Processing Systems, 2019, 32.

[11] Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory transformer for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10578-10587.

[12] Ji J, Luo Y, Sun X, et al. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(2): 1655-1663.

[13] Jiang H, Misra I, Rohrbach M, et al. In defense of grid features for visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10267-10276.

[14] Bhat S F, Birkl R, Wofk D, et al. Zoedepth: Zero-shot transfer by combining relative and metric depth[J]. arXiv preprint arXiv:2302.12288, 2023.

[15] Fang H, Zhang D, Zhang Y, et al. Salient object ranking with position-preserved attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16331-16341.

[16] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Proceedings of the European Conference on Computer Vision. 2014: 740-755.

[17] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.

[18] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2002: 311-318.

[19] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005: 65-72.

[20] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. 2004: 74-81.

[21] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015: 4566-4575.

[22] Pan Y, Yao T, Li Y, et al. X-linear attention networks for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10971-10980.

[23] Huang L, Wang W, Chen J, et al. Attention on attention for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4634-4643.

[24] Shi Z, Zhou X, Qiu X, et al. Improving image captioning with better use of caption[C]. Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2020. 7454-7464.