Fine-Grained Image Captioning by Ranking Diffusion Transformer

Jun Wan, Min Gan, *Senior Member, IEEE,* Lefei Zhang, *Senior Member, IEEE,* Jie Zhou, Jun Liu, Bo Du, *Senior Member, IEEE,* C. L. Philip Chen, *Fellow, IEEE,*

***Abstract*—The CLIP visual feature-based image captioning models have developed rapidly and achieved remarkable results. However, they still struggle to generate descriptive and discriminative captions as they fail to fully exploit visual details and model complex visual-linguistic alignment relationships. To overcome these limitations, this paper proposes a Ranking Diffusion Transformer (RDT) which consists of a Ranking Visual Encoder (RVE) and a Ranking Loss (RL) for fine-grained image captioning. The RVE is designed to mine diverse and discriminative information from the visual features by proposing a new ranking attention. Meanwhile, the RL is proposed to optimize the diffusion process while strengthening the vision-language semantic alignment by using the ranking results of the generated caption sentence quality as an additional overall semantic supervisory signal. We show that by collaborating RVE and RL via the novel Ranking Diffusion Transformer, and gradually adding and removing noise in the diffusion process, more discriminative visual features are learned and precisely aligned with the language features. Experimental results on popular benchmark datasets demonstrate that our RDT surpasses existing state-of-the-art image captioning models in the literature. The code is publicly available at: https://github.com/junwan2014/RDT.**

***Index Terms*—image captioning, diffusion model, visual feature, language feature, fine-grained.**

I. Introduction

IMAGE captioning involves the automated generation of descriptive sentences for a given image, achieved by identifying objects, modeling their spatial and semantic relationships, and verbalizing them using natural language. **Fine-grained image captioning** [1], [2] refers to the task of generating detailed and specific descriptions for images and the generated captions may include specific attributes, characteristics, or relationships depicted in the image, such as the color, texture, shape, orientation, or arrangement of objects. The production of accurate and descriptive image captions is of paramount importance in numerous applications, such as visual intelligence [3], [4] in image search, conversational robots [5], photo sharing [6], [7], and assisting individuals with visual impairments [8].

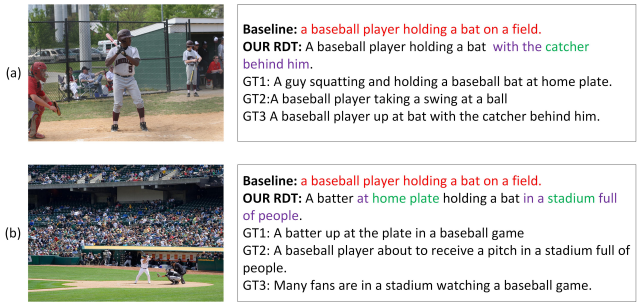

Fig. 1. The baseline (i.e., SCD-Net [9]) fails to distinguish images with similar content and tends to generate less descriptive and discriminative captions, e.g., the SCD-Net ignores some important visual information of images (a) and (b), and generates the same captions for them. By collaborating with Ranking Visual Encoder and Ranking Loss, our proposed RDT can mine more diverse and discriminative visual information (e.g., home plate, catcher, stadium, etc.) and achieve more precise vision-language semantic alignment (e.g., with the catcher behind him, and in a stadium full of people, etc) for generating fine-grained captions.

The general paradigm of image captioning is that: the visual features (e.g., region image feature [10] or grid image feature [11]) is fed into a Transformer [12] to generate captions. The Transformer is an encoder-decoder framework, wherein the encoder is used to enhance the visual features and the decoder generates the captions conditioned on the enhanced visual features. Given that grid features outperform region features in both performance and time cost, current captioning models [13], [14], [15] tend to employ grid features (e.g., the CLIP visual features [15]) to generate captions. However, current CLIP visual feature-based image captioning models [9], [16], [17], [18] often produce coarse and less discriminative captions (as shown in Fig. 1) due to: 1) the adoption of a visual encoder design with a self-attention mechanism and the sequential stacking way of multiple visual encoders, which enhances visual features by prioritizing important information while suppressing unimportant details, resulting in the inability to effectively model all visual information [19], [20], and 2) the use of cross-entropy loss as the objective function, which treats each word independently and overlooks the complex

This work is supported by the National Natural Science Foundation of China (Grant No. 62571555, 62002233 and 62476172), and the Natural Science Foundation of Hubei Province (Grant No. 2024AFB992). Corresponding author: Lefei Zhang and Bo Du.

J. Wan is with the School of Information Engineering, Zhongnan University of Economics and Law, Wuhan, 430073, China and with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail:junwan2014@whu.edu.cn).

M. Gan is with the College of Computer Science and Technology, Qingdao University, Qingdao 266071, China (e-mail: ganmin@aliyun.com).

L. Zhang and B. Du are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: zhanglefei@whu.edu.cn, dubo@whu.edu.cn).

J. Liu is with the School of Computing and Communications, Lancaster University, LA1 4YW Lancaster, U.K. (e-mail: j.liu81@lancaster.ac.uk).

J. Zhou is with the Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. (e-mail: jie_jpu@163.com).

C. L. Philip Chen is the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: philip.chen@ieee.org).

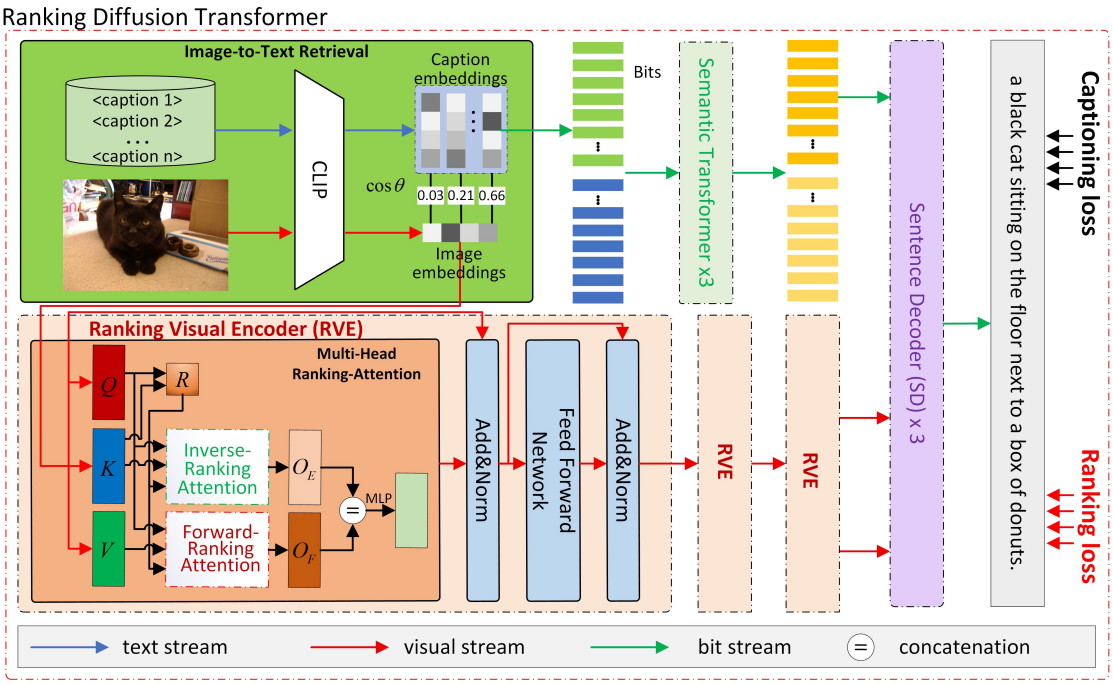

Fig. 2. The overall architecture of the proposed Ranking Diffusion Transformer (RDT). The proposed RDT can effectively enhance the visual representation by designing a novel Ranking Visual Encoder to mine diverse and discriminative information from the visual features, and then use a Ranking Loss as an additional supervision signal to guide the convergence of the diffusion process and the precise alignment between vision and language features, thereby achieving fine-grained image captioning.

vision-language alignment and overall semantic and contextual information [21], [9].

To address the aforementioned issue, this paper proposes a Ranking Diffusion Transformer (RDT) (as shown in Fig. 2) for fine-grained image captioning. The RDT takes the advantage of diffusion model in capturing details and diverse outputs, and incorporates a Ranking Visual Encoder (RVE) and a Ranking Loss (RL) to improve the distinguishability of generated captions. Specifically, the RVE is designed to focus on both important and previously disregarded features by proposing a novel ranking attention. The ranking attention contains a forward-ranking attention and an inverse-ranking attention. The forward-ranking attention aims to learn important visual features and mine correlations between them, which is further enhanced by a sequentially stacked encoder structure. The inverse-ranking attention module pays attention to previously disregarded visual features again and reactivates them to capture potential subtle differences. By expanding the model's focus to include not only important but also previously disregarded "unimportant" features, the ranking attention is able to mine more discriminative information to enhance visual features. Additionally, a novel RL is proposed to strengthen supervision on the diffusion process and promote visual-language semantic alignment by ranking the quality of generated captions corresponding to different noise timesteps. Therefore, the proposed RDT achieves fine-grained image captioning and its main contributions are summarized as follows:

1) We propose a novel Ranking Diffusion Transformer (RDT) to address fine-grained image captioning problems by building RVE and RL on a diffusion model, and RDT outperforms the state-of-the-art image captioning models on popular benchmark datasets COCO [22], Flickr30k [23] and Nocaps [24].

2) A well-designed Ranking Visual Encoder (RVE) is developed to mine diverse and discriminative information from visual features by combining forward-ranking attention and inverse-ranking attention.

3) A novel Ranking Loss is presented to provide a finer supervision signal to optimize the diffusion process and strengthen the vision-language semantic alignment by accurately ranking and distinguishing the generated less discriminative captions.

The remaining sections of this paper are organized as follows: Section **II** presents a comprehensive overview of the related work, while Section **III** introduces the proposed Ranking Diffusion Transformer (RDT). In Section IV, we provide the experimental setup, results, and evaluation of the RDT. Finally, Section **V** concludes the paper.

II. Related Work

Many approaches have been proposed in image captioning and have yielded promising results. Broadly, captioning models can be classified into two groups: autoregressive methods [25], [26], [14] and non-autoregressive methods [27], [21], [9].

Autoregressive methods. With the proliferation of deep learning techniques, the encoder-decoder framework [26], [28] has been extensively employed in captioning models. CNN and RNN are usually used as the encoder and decoder for learning visual features and generating the output descriptions,

respectively. Liu et al. [29] propose the NICVATP2L model to tackle the challenge of Chinese image caption generation. By integrating visual attention with topic modeling, NICVATP2L can generate more informative and natural Chinese captions. However, these models fail to produce accurate and fluent captions due to their sequential nature and inability to model complex relationships among distant objects. In recent years, the Transformer has emerged as a solution to this problem by replacing recurrence and convolutions with the attention mechanism, resulting in remarkable performance. By integrating region features and grid features, more effective visual representations are learned by Dual-Level Collaborative Transformer (DLCT) [13] to generate more accurate captions. By taking the segmentation features as the complement information to enhance grid features, DIFNet [14] generates captions that are more faithful to given images and achieve excellent captioning performance. Li et al. [15] present the Comprehending and Ordering Semantics Network (COS-Net) as a solution that integrates a semantic comprehender and ranker to enhance the sentence decoding process and ultimately improve captioning performance.

With the widespread dissemination of large-scale vision-language models [16], [30], [31], the incorporation of external knowledge has become increasingly important for enhancing visual features. For example, Nie et al. [32] propose a conversational image search framework (LARCH), which integrates visual features with multi-form knowledge to learn knowledge-enhanced representations. Meanwhile, CLIP-based visual features have emerged as the mainstream and most widely adopted choice for image captioning, and CLIP-based captioning methods have rapidly advanced. Mokady et al. [16] propose the ClipCap method, which converts the CLIP visual features into prefix embeddings of visual prompting by training a simple mapping network, then the GPT-2 is fine-tuned to generate image captions. Luo et al. [17] propose the I-Tuning method, which aims to automatically filter visual information in images to adjust the output hidden state of large language models. I-Tuning can achieve state-of-the-art results while reducing training parameters by half to three-quarters. Yu et al. [33] present the CoCa model which utilizes captioning loss and contrastive loss to combine visual pre-training and natural language supervision, thereby improving the captioning performance. Ramos et al. propose SmallCap [18], which uses the retrieved description as a task demonstration and language prompt, and then combines it with the CLIP visual features to improve captioning performance. Li et al. [34] propose a retrieval-augmented image captioning method, which improves image captioning accuracy by prompting LLMs with object names retrieved from External Visual–name memory. By using the retrieved text as visual prompts in the CLIP space, ViPCap [35] can effectively enhance lightweight image captioning performance. However, the above autoregressive models suffer from the limitation of generation speed and the accumulation of errors.

Non-autoregressive methods. Autoregressive methods typically generate sentences sequentially, word-by-word. In contrast, non-autoregressive methods generate all words simultaneously, enabling bidirectional text messaging and improving image captioning performance. The non-autoregressive model is first proposed to address the neural machine translation problem, which improves both the accuracy and inference speed and also promotes the development of image captioning [36]. By generating captions in parallel from a totally masked sequence to a totally non-masked sequence, the masked non-autoregressive model [37] enables more diverse and descriptive image captioning. Liu et al. [38] propose an Object-Oriented Non-Autoregressive (O2NA) approach, which involves generating a draft caption and then refining it to obtain a fluent final caption. By preserving the autoregressive property globally and generating words parallelly local, Zhou et al. [27] propose a semi-autoregressive Transformer for balancing its speed and quality. Recently, Chen et al. [21] use self-conditioning and asymmetric time intervals to improve the sample quality in Bit Diffusion, which achieves competitive results compared to autoregressive captioning models. By firstly searching semantically relevant sentences and then treating them as the semantic prior to generate captions in a diffusion process, Luo et al. [9] propose a semantic-conditional diffusion network, which shows promising potential for image captioning.

Our research is also classified within the domain of non-autoregressive methods employing diffusion models. Our RDT surpasses conventional diffusion models by strengthening the alignment of visual and language semantics through the proposed Ranking loss. Additionally, we introduce a novel Ranking Visual Encoder that extracts diverse and discriminative information from visual features to enhance visual representations for achieving fine-grained image captioning.

III. Ranking Diffusion Transformer

In this section, we present the proposed Ranking Diffusion Transformer (RDT) that aims to enhance the vision-language semantic alignment and facilitates the learning of diverse and discriminative visual features. Fig. 2 illustrates the overall framework of RDT.

A. Problem Formulation

Image captioning refers to generating a sentence to describe an image. Generally, an image I is described by a sentence Y, which consists of N_y words denoted by $Y = \{y_1, y_2, \cdots, y_{N_y}\}$. Then, each word is converted into $n = \lceil \log_2 \mathcal{W} \rceil$ binary bits (i.e., $\{0,1\}^n$) to trigger the diffusion model, where $\mathcal{W}$ denotes the vocabulary size. Images are typically represented by features extracted using a pre-trained detector/classifier, and the grid features are often chosen to represent the image information for image captioning. Assuming the grid feature is denoted by $\mathcal{G} = (g_1, g_2, ..., g_N)$ consisting N grids and $g_i \in \mathbb{R}^{D_g}$. Then, the diffusion model, which includes both a forward process and a reverse process, is used to achieve vision-language semantic alignment conditioned on the grid features.

Foward Process. In the forward process of the diffusion model [39], Gaussian noise is gradually added to the sentence data x_0, where x_0 denotes the bit representation of Y. Assum-

ing a total of T timesteps, the forward state transition can be defined as follows:

$$x_t = \sqrt{\sigma\left(-\gamma\left(t'\right)\right)}x_0 + \sqrt{\sigma\left(\gamma\left(t'\right)\right)}\epsilon, \tag{1}$$

where $t' = t/T$, ϵ follows the Gaussian distribution $\mathcal{N}(0,1)$, and t follows the uniform distribution $\mathcal{U}(0,T)$. $\mathcal{N}$ and $\mathcal{U}$ represent the normal distribution and the uniform distribution, respectively. $\gamma(t')$ and σ are the monotonically increasing function and the sigmoid function. Subsequently, a diffusion transformer $f(x_t, \gamma(t'), \mathcal{G})$ is trained to reconstruct x_0 with the guidance of $\mathcal{G}$ in a denoising process. The reconstruction process can be defined as follows:

$$\mathcal{L}_{bit} = \mathbb{E}_{t\sim\mathcal{U}(0,T),\epsilon\sim\mathcal{N}(0,1)}\left\|f\left(x_t, \gamma\left(t'\right), \mathcal{G}\right) - x_0\right\|^2 \tag{2}$$

Reverse Process. In the reverse process [39], the diffusion model samples a sequence of latent states x_t from $t = T$ to $t = 0$, and the reverse state transition x_{t-1} is defined as follows:

$$\alpha_s = \sqrt{\sigma\left(-\gamma\left(s'\right)\right)}, \alpha_t = \sqrt{\sigma\left(-\gamma\left(t'\right)\right)}, \tag{3}$$

$$\mu_s = \sqrt{\sigma\left(\gamma\left(s'\right)\right)}, c = -\mathrm{expm1}\left(\gamma\left(s'\right) - \gamma\left(t'\right)\right) \tag{4}$$

$$u\left(x_t; s', t'\right) = \alpha_s\left(x_t\left(1-c\right)/\alpha_t + cf\left(x_t, \gamma\left(t'\right), \mathcal{G}\right)\right), \tag{5}$$

$$\mu^2(s', t') = \mu_s^2 c, \tag{6}$$

$$x_{t-1} = u(x_t; s', t') + \mu(s', t')\epsilon \tag{7}$$

where Δ represents the time difference, $s = t - 1 - \Delta$ is calculated by discretizing time uniformly with a width of $1/T$, and $s' = s/T$. α_s and α_t are coefficients that control the noise level and are used to adjust the noise intensity through the functions $\gamma()$. c represents the difference between the current timestep and the next timestep, determining the extent of noise influence during the reverse diffusion process between timesteps. $u\left(x_t; s', t'\right)$ means computing an estimate of the next state x_{t-1} based on the current state x_t, time t' and s', which is derived from a linear combination of the current state x_t and the predicted caption sentence $f\left(x_t, \gamma\left(t'\right), g\right)$. $\mu^2(s', t')$ is the noise variance between the current timestep and the next timestep and is used to represent the noise intensity in the reverse process. By iteratively triggering the Diffusion Transformer starting from x_T, we can obtain the estimated value x_0.

B. Ranking Diffusion Transformer

The fundamental Diffusion Transformer adheres to a conventional encoder-decoder structure, comprising a visual encoder and a sentence decoder. Specifically, the visual encoder initially transforms visual features into visual tokens and enhances them. These enhanced visual tokens are subsequently combined with word tokens $x_t = \left\{y_0^t, y_1^t, ..., y_{N_s}^t\right\}$ at timestep t and inputted into the sentence decoder for generating captions. However, current visual encoders employ the self-attention mechanism to capture correlations among visual features and enhance them through sequentially stacked encoders. This characteristic would make the captioning models focus on the most salient common objects and ignore specific detailed aspects of an image that distinguish it from others

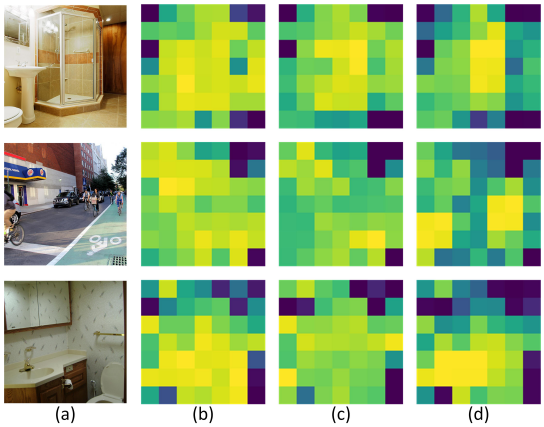

Fig. 3. Visualization of heatmaps generated by different encoders of the baseline (i.e., SCD-Net). (a) Image, (b) the first encoder, (c) the second encoder, and (d) the third encoder. The self-attention mechanism would make the baseline captioning models focus on the most salient common objects and may ignore specific detailed aspects of an image that distinguish it from others, e.g., the highlighted area is gradually decreasing.

(e.g., the highlighted area is gradually decreasing as shown in Fig. 3), thereby leading to coarse and less discriminative caption generation. To address this limitation, we introduce a novel Ranking Visual Encoder (RVE) (as depicted in Fig. 4), which expands the model's focus to include both important features and previously ignored "unimportant" ones, thereby enabling the mining of diverse and discriminative feature representations to help achieve fine-grained image captioning.

Ranking Visual Encoder. The RVE takes the grid feature as input and enhances it by proposing the multi-head ranking-attention (MHRA) layer (as shown in Fig. 2). Assuming there are N_v stacked RVEs and each RVE contains an MHRA layer, a feed-forward network (FFN), multiple layer normalization (LN) layers, and fully connected (FC) layers. The entire process of the RVE can be defined as follows:

$$\begin{aligned}\mathcal{G}^{i_v+1} &= RVE(\mathcal{G}^{i_v}) \\ &= FFN(LN(\mathcal{G}^{i_v} + MHRA(\mathcal{G}^{i_v}, \mathcal{G}^{i_v}, \mathcal{G}^{i_v}))),\end{aligned} \tag{8}$$

$$FFN(Z) = LN(Z + FC(\delta(FC(Z)))), \tag{9}$$

$$MHRA(Q, K, V) = Concat(h_1, h_2, ..., h_H)W^O, \tag{10}$$

$$h_j = Attention(QW_j^Q, KW_j^K, VW_j^V), \tag{11}$$

$$Attention(Q, K, V) = softmax(\frac{\mathrm{Q}K^T}{\sqrt{d}}V), \tag{12}$$

where H denotes the number of heads in the RVE and d is the dimension of each head. $Concat(\cdot)$ and δ are the concatenation operation and the activation function, respectively. After stacking multiple RVEs, the diverse and discriminative visual tokens $\hat{\mathcal{V}} = \mathcal{G}^{N_v}$ can be obtained.

After that, we follow SCD-Net [9] and introduce a **Semantic Transformer** to constrain the diffusion process by utilizing a semantically relevant sentence as the semantic condition. To obtain semantically related sentences, we first build a training sentence pool using training captions from COCO, and design an image-to-text retrieval block. The image-to-text retrieval block extracts image and caption representations using CLIP

[1] and selects semantically related sentences based on cosine similarity. After that, a **Sentence Decoder** is also designed to generate the final caption.

Ranking Loss. So far, the Bit-Diffusion-based image captioning [21], [9] has been achieved, i.e., they gradually perturb the original input data by adding Gaussian noise over successive steps in the forward process and then recovering the original input data from the diffused (noisy) data step by step in the reverse process. In this way, the diffusion model decomposes the challenging image captioning problem into multiple relatively simple tasks, with each task corresponding to one timestep denoising. However, this process [21], [9] overlooks the complex vision-language alignment and inherent sequential dependency among words during diffusion process, resulting in coarse and less accurate captions. Moreover, employing the cross-entropy loss [40], [19], [9] to fit the captioning model also results in the model generating descriptions with correct individual words but less overall semantic or contextual information as it treats each word independently and assigns equal penalties for incorrect predictions. To mitigate these issues, we further explore how to better optimize the diffusion process to precisely guide the semantic alignment between vision and language. Specifically, we propose a novel Ranking Loss, introducing the impact of the noises added at different timesteps on the quality of generated captions in the supervision signal. The overall new objective function is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{XE} + \lambda_2 \mathcal{L}_{bit} + \lambda_3 \mathcal{L}_{rank}, \tag{13}$$

where $\mathcal{L}_{XE}$, $\mathcal{L}_{bit}$ and $\mathcal{L}_{rank}$ denote the cross-entropy loss, the bit loss and the proposed Ranking Loss. λ_1, λ_2 and λ_3 are weights for balancing these losses. To further boost captioning performance, we follow SCD-Net [9] and use the Guided Self-Critical Sequence Training mechanism. The corresponding gradient is approximated as:

$$\nabla_\theta \mathcal{L}_R(\theta) \approx -\frac{1}{N_y} \sum_{j=0}^{N_y} \left(R\left(y_{1:N_s}^{s_j}\right) - R\left(\hat{y}_{1:N_s}\right) \right) \nabla_\theta \log p_\theta \left(y_{1:N_s}^{s_j}\right), \tag{14}$$

where $y_{1:N_s}^{s_j}$ represents the sampled caption, and N_y denotes the total number of randomly sampled captions plus one (i.e., the caption predicted by a standard autoregressive transformer teacher model, which shares the same structure as the RDT). $R\left(\hat{y}_{1:N_s}\right)$ is the baseline's sentence-level reward. This mechanism guarantees that high-quality sentences generated by the RDT are rewarded positively, thus encouraging the generation of high-quality sentences.

The details of our two main components, namely RVE and RL, are described in the following subsection.

C. Ranking Visual Encoder

The proposed Ranking Visual Encoder (RVE) consists primarily of a multi-head ranking-attention (MHRA) layer and a feed-forward network (FFN). Compared with the multi-head self-attention (MHSA) layer, our MHRA designs a novel ranking attention which facilitates the extraction of more discriminative feature representations for achieving fine-grained image captioning.

[1]CLIP-ViT-B/32, https://github.com/jianjieluo/OpenAI-CLIP-Feature

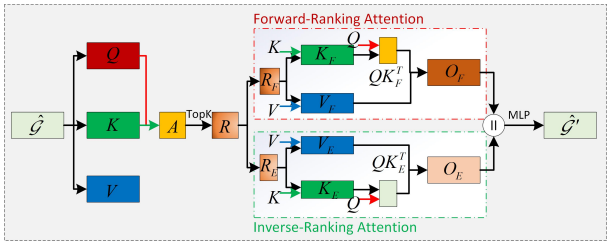

Fig. 4. The proposed Multi-Head Ranking-Attention (MHRA). By incorporating forward-ranking attention and inverse-ranking attention, the proposed MHRA can mine more diverse and discriminative information from visual features to enhance visual representation.

Ranking attention. RVE utilizes the CLIP visual features [16], [15] as the visual input. The CLIP visual features comprises the grid feature $\mathcal{G} = (g_1, g_2, ..., g_{N-1})$ (contains $N-1$ grids and $g_i \in \mathbb{R}^{D_g}$), and the CLS feature (denoted as g_c). Before feeding them into the RVE, these features are transformed into a new embedding space and concatenated as $\hat{\mathcal{G}} = [\hat{g}_c, \hat{g}_i |_{i=1}^{N-1}]$. Then, the query, key, and value tensors, $Q, K, V \in \mathbb{R}^{N \times D_g}$ can be calculated as:

$$Q = W^q \hat{\mathcal{G}}, K = W^k \hat{\mathcal{G}}, V = W^v \hat{\mathcal{G}}, \tag{15}$$

where W^q, W^k and $W^v \in \mathbb{R}^{N \times N}$ are embedding matrixes. Subsequently, we construct a directed graph to identify attended relationships, determining which grids should be attended to for each given grid. Specifically, we obtain the grid-to-grid affinity graph's adjacency matrix, denoted as $\mathcal{A} \in \mathbb{R}^{N \times N}$, through matrix multiplication between Q and the transpose of K. $\mathcal{A}$ quantifies the semantic relation between two grids. In MHSA, all grid-to-grid affinity relationships will be used to update the features and each query grid should associate with all attended grids. However, according to the visualization of pretrained ViT [41], queries in different semantic regions actually attend to quite different key-value pairs. Hence, forcing all queries to attend to the same set of grids may be suboptimal and this paper seeks a dynamic, query-aware attention mechanism, i.e., the **ranking attention**. The proposed ranking attention includes a forward-ranking attention and an inverse-ranking attention. The forward-ranking attention aims for each query to attend to the highly relevant key-value pairs and discard the others for learning the main information, while the inverse-ranking attention only focuses on the weakly relevant key-value pairs (e.g., those discarded by the forward-ranking attention) and reactivate them for mining diverse and discriminative information. These key-value pairs can be selected by ranking attended grids according to their attended scores in $\mathcal{A}$, and we use the top-$\mathcal{K}$ method to calculate the ranked affinity routing index $\mathcal{R}$ for all query grids, which can be summarized as follows:

$$\mathcal{R} = top\mathcal{K}(\mathcal{A}), \tag{16}$$

where $\mathcal{K}$ is set to N, and the i-th row of $\mathcal{R}$ consists of N indices representing the relevant grids for the i-th grid, arranged in descending order of relevance.

Forward-Ranking Attention. In forward ranking attention, we set $\mathcal{K} = \lambda_F \cdot N$, where λ_F represents the ratio of attended key-value pairs. Consequently, we can obtain the forward-ranking routing matrix $\mathcal{R}_F$ using the following formulation:

$$\mathcal{R}_F = \mathcal{R}[\ :,\ :\lambda_F N]. \tag{17}$$

Then, we gather forward-ranking key and value with routing index matrix R_F:

$$K_F = \text{gather}(K, \mathcal{R}_F), V_F = \text{gather}(V, \mathcal{R}_F), \tag{18}$$

where K_F and $V_F \in \mathbb{R}^{\lambda_F N \times D_g}$, gather operation can be implemented by directly using torch.gather() funcion. Then, self-attention is applied to the gathered key-value pairs as:

$$O_F = Attention(Q, K_F, V_F) = softmax(\frac{QK_F^T}{\sqrt{d}})V_F, \tag{19}$$

where $\sqrt{d}$ is used to avoid concentrated weight and gradient vanishing.

Inverse-Ranking Attention. In the inverse-ranking attention, the ratio of discarded key-value pairs is denoted as λ_E. It should be noted that due to the small weights assigned by the forward-ranking attention to low-ranked attended key-value pairs, these pairs will be reused in the inverse-ranking attention, i.e., $\lambda_E > (1 - \lambda_F)$. The resulting inverse-ranking routing matrix is denoted as $\mathcal{R}_E$ and can be calculated as:

$$\mathcal{R}_E = \mathcal{R}[\ :, (1 - \lambda_E)N\ :]. \tag{20}$$

Then, we gather the inverse-ranking key and value with routing index matrix $\mathcal{R}_E$, and then the self-attention is applied to the gathered key-value pairs as:

$$K_E = gather(K, \mathcal{R}_E), V_E = gather(V, \mathcal{R}_E), \tag{21}$$

$$O_E = Attention(Q, K_E, V_E) = softmax(\frac{QK_E^T}{\sqrt{d}})V_E. \tag{22}$$

The discarded features can be reweighted for extracting potential diverse and discriminative feature information using the softmax operation in Eq. (22). Subsequently, we concatenate O_F and O_E along the channel dimension, followed by a Multi-Layer Perception (MLP) for fusion. The overall process can be defined as:

$$MHRA(\mathcal{G}, \mathcal{G}, \mathcal{G}) = MLP(Concat(O_F, O_E)). \tag{23}$$

Therefore, with the well-designed RVE, more diverse and discriminative information from visual features can be recollected and mined to enhance feature representation for achieving fine-grained image captioning.

D. Ranking Loss

Besides proposing the Ranking Visual Encoder (RVE) to mine diverse and discriminative information to enhance visual features for image captioning, this paper further explores how to promote visual-language semantic alignment by strengthening supervision on the diffusion process, i.e., we propose a novel Ranking Loss (RL), which ranks the quality of captions generated at different timesteps and treats them as fine supervisory signals, thereby helping achieve fine-grained image captioning.

In our RDT, we follow [21] and also use analog bits to represent the sentences for achieving continuous diffusion model. Since analog bits are continuous variables and inspired by the idea that the average can provide useful summary information, we define the sentence quality score as the mean of all analog bits (before thresholding [21]), which will be used to characterize the semantic information of a sentence.

Assuming that pairs of images (e.g., images $I_{k,a}$ and $I_{k,b}$ with the same content) are input to our RDT, k is the index of images in datasets. In the forward process of our RDT, their corresponding input sentences (i.e., the ground-truth one) are the same and Gaussian noises corresponding to different timesteps will be added to them. Although both are optimized towards the same goal, i.e., the ground-truth sentence, the generated caption sentence qualities $q_{k,a}$ and $q_{k,b}$ should differ. Generally, the more Gaussian noise is added to the input sentence, the greater the difference between the generated caption sentence quality score and the ground-truth caption sentence quality score will be, as the added Gaussian noise disrupts the semantics of the input sentence, thereby enlarging the score difference. This motivates us to further distinguish the generated captions by ranking them according to their added Gaussian noise intensity during training. Additionally, considering a more general case, i.e., the single case (not the pairwise case), we can also achieve this goal with the help of the ground-truth caption sentence quality score. The proposed Ranking Loss can be defined as:

$$\begin{aligned} &\mathcal{L}_{rank}^{i,j} = \\ &\max\left(0, \left|q_i^* - q_j^*\right| - \text{sign}\left(q_i^*, q_j^*\right)\left(q_{i,t_i} - q_{j,t_j}\right)\right) \end{aligned} \tag{24}$$

$$\text{sign}\left(q_i^*, q_j^*\right) = \begin{cases} 1, & q_i^* > q_j^* \\ 0, & q_i^* = q_j^* \\ -1, & q_i^* < q_j^* \end{cases}. \tag{25}$$

where i and j denote two image indexes in a mini-batch, with q_i^* and q_j^* representing their ground-truth caption quality scores. The notations q_{i,t_i} and q_{j,t_j} denote the predicted quality scores for the generated captions of images i and j, respectively, when noise is injected at timesteps t_i and t_j. The term $\text{sign}(q_i^* - q_j^*)$ encodes the relative ranking order of the ground-truth captions. When the model predicts rankings consistent with the ground truth–i.e., $\text{sign}(q_i^* - q_j^*)$ and $\left(q_{i,t_i} - q_{j,t_j}\right)$ share the same sign–our RDT effectively pushes q_{i,t_i} toward q_i^* and q_{j,t_j} toward q_j^*. In this case, $\text{sign}(q_i^* - q_j^*)\left(q_{i,t_i} - q_{j,t_j}\right)$ approximates $|q_i^* - q_j^*|$, and the pairwise ranking loss $\mathcal{L}_{rank}^{i,j}$ approaches zero. Conversely, when the predicted ranking conflicts with the ground truth, the discrepancy between $|q_i^* - q_j^*|$ and $\text{sign}(q_i^* - q_j^*)\left(q_{i,t_i} - q_{j,t_j}\right)$ becomes large, producing a positive loss that signals the need for adjustment. Moreover, even when q_i^* and q_j^* are close–indicating highly similar ground-truth captions and visual content–our ranking loss remains effective in distinguishing subtle differences, thereby enabling fine-grained image captioning.

To further distinguish the less discriminative captions and achieve fine-grained image captioning, our RL introduces

Algorithm 1: Sampling of RDT

Input: grid feature $\mathcal{G}$, conditional sentence x_c
Output: predicted sentence x_0.
Randomly initialize x_T, $x_{start} = None$
for t = T, ..., 1 do
  $s' = (t-1-\Delta)/T$, $t' = t/T$
  $x_{start} = \text{embed}\left(\text{cat}\left(\left(x_t, \gamma\left(t'\right)\right), -1\right)\right)$
  $x_{start} = \text{cat}\left(\left(x_{start}, x_c\right), 1\right)$
  $x_{start} = \text{ST}(x_{start})$
  $x_{start} = x_{start}[:, : seq\_len, :]$ # cut operation
  $x_{start} = \text{SD}(x_{start}, \mathcal{G})$
  $\alpha_s = \sqrt{\sigma\left(-\gamma\left(s'\right)\right)}, \alpha_t = \sqrt{\sigma\left(-\gamma\left(t'\right)\right)}$
  $\mu_s = \sqrt{\sigma\left(\gamma\left(s'\right)\right)}, c = -\text{expm1}\left(\gamma\left(s'\right) - \gamma\left(t'\right)\right)$
  $u\left(x_t; s', t'\right) = \alpha_s\left(x_t\left(1-c\right)/\alpha_t + c * x_{start}\right),$
  $\mu^2(s', t') = \mu_s^2 c,$
  $x_{t-1} = u(x_t; s', t') + \mu(s', t')\epsilon$
return x_0

constraints on timestep difference and sentence quality score difference. For constraints on timestep difference, we assume that qualities of captions generated by adding noise corresponding to close timesteps are regarded to have close quality scores to their corresponding ground-truth ones, and only when their timestep difference is less than δ_{tim} (i.e., they are less discriminative), the corresponding Ranking Loss is used to adjust the model. Also for the sentence quality score difference, we only need to adjust the model when their ground-truth caption sentence quality score difference is less than δ_{sen}. The new RL can be defined as:

$$\mathcal{L}_{rank} = \frac{1}{n^2} M_{tim}^{i,j} M_{sen}^{i,j} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{L}_{rank}^{i,j}, \tag{26}$$

$$M_{tim}^{i,j} = \begin{cases} 1, |t_i - t_j| \leq \delta_{tim} \\ 0, \qquad\quad else \end{cases}, \tag{27}$$

$$M_{sen}^{ij} = \begin{cases} 1, \left|q_i^* - q_j^*\right| \leq \delta_{sen} \\ 0, \qquad\quad else \end{cases}, \tag{28}$$

With the above timestep mask M_{tim} and sentence difference mask M_{sen}, the proposed RL accurately ranks and distinguishes the generated less discriminative captions, thereby providing finer supervision signals to optimize the diffusion process and helping achieve fine-grained image captioning. Moreover, since the diffusion process is carried out step by step, a certain step may focus on a specific sub-semantic. By adjusting and optimizing in the intermediate steps, the accuracy and fluency of the generated captions can be gradually improved, and the image details and contextual information (e.g., color, shape of an object) are effectively captured. Hence, the proposed RL can promote the diffusion model to pay more attention to visual information related to certain sub-semantics at a certain timestep, thereby mining diversity and discriminative visual information to strengthen vision-language semantic alignment.

IV. Experiments

This section begins by introducing the datasets and implementation details. Subsequently, the comparison of our proposed RDT and state-of-the-art image captioning models

TABLE I
Notation used in the Ranking Diffusion Transformer (RDT).

Symbol	Meaning
I	Input image
$Y = \{y_1, \ldots, y_{N_s}\}$	Ground-truth caption tokens
$x_0,\ x_t$	Clean / noised sentence at timestep t
$T,\ t,$	Number of timesteps, current step
$\gamma(\cdot),\ \sigma(\cdot)$	Noise schedule and sigmoid function
ε	Gaussian noise $\mathcal{N}(0,1)$
$\mathcal{G} = (g_1, \ldots, g_N)$	CLIP grid features (N grids, dim D_g)
$\hat{\mathcal{V}}$	Enhanced visual tokens after RVEs
Q, K, V	Query/Key/Value in attention
$\mathcal{A}$	Grid-to-grid affinity matrix
$\mathcal{R}$	Ranked routing indices from A
$\mathcal{R}_F,\ \mathcal{R}_E$	Forward-/Inverse-ranking routes
K_F, V_F	Gathered KV after forward-ranking
K_E, V_E	Gathered KV after inverse-ranking
$O_F,\ O_E$	Attended outputs (forward / inverse)
$\lambda_F,\ \lambda_E$	Keep/reuse ratios for ranking attention
$H,\ d$	Number of heads; per-head dimension
$\text{MHRA}(\cdot)$	Multi-Head Ranking-Attention layer
$\alpha_s, \alpha_t, \mu_s, c$	Reverse process coefficients
$u(\cdot)$	Mean update in reverse transition
$ST,\ SD$	Sentence Transformer; Sentence Decoder
$\mathcal{L}_{\text{XE}},\ \mathcal{L}_{\text{bit}},\ \mathcal{L}_{\text{rank}}$	Cross-entropy, bit loss, ranking loss
$\lambda_1, \lambda_2, \lambda_3$	Loss weights in the total objective
i, j	Indices of two images in a mini-batch
q_i^*, q_j^*	Ground-truth caption quality scores
q_{i,t_i}, q_{j,t_j}	Predicted quality scores at timesteps t_i, t_j
$\text{sign}(q_i^* - q_j^*)$	Ground-truth ranking relation
$\mathcal{L}_{\text{rank}}^{i,j}$	Pairwise ranking loss
$M_{\text{tim}}^{i,j},\ M_{\text{sen}}^{i,j}$	Timestep mask; sentence-difference mask
$\delta_{\text{tim}}, \delta_{\text{sen}}$	Thresholds for masks
$D@2, D@3$, Voc-u	Diversity evaluation metrics

[45], [46], [26], [13], [40] is given. Ablation studies and self-evaluations are then conducted. Finally, we present the experimental results and engage in discussions.

A. Datasets and Implementation Details

Datasets. 1) MS-COCO [22]. This dataset consists of a total of 164,062 images, with 82,783 images for training, 40,504 images for validation, and 40,775 images for testing. Each image in the dataset is annotated with 5 captions. We adopt the Karpathy split [52] with 113,287 training images and 5,000 each for validation and testing. 2) Flickr30k [23]. It includes 31,000 images collected from the Ficker website, along with 158 thousand captions written by humans. In our experiments, we use the Karpathy split [52] for Flickr30k. 3) Nocaps [24]. It is divided to three parts, in-domain contains images portraying only COCO classes, near-domain contains both COCO and novel classes, and out-of-domain consists of only novel classes. We evaluate RDT on the validation set.

Evaluation Metrics. We follow the standard evaluation protocol and report results on widely used captioning metrics, including BLEU-N (B@N) [53], METEOR (M) [54], ROUGE (R) [55], CIDEr (C) [56], and SPICE (S) [57]. In addition, we employ Dist-2 (D@2) and Dist-3 (D@3) [58], as well as vocabulary usage (Voc-u) [59], to assess the diversity of the generated captions.

TABLE II
PERFORMANCE COMPARISONS WITH THE STATE-OF-THE-ART IMAGE CAPTIONING MODELS ON COCO KARPATHY TEST SPLIT, WHERE B@N, M, R, C AND S ARE SHORT FOR BLEU@N, METEOR, ROUGE-L, CIDER AND SPICE SCORES. ◊ MEANS USING THE CLIP VISUAL FEATURES AS INPUT.

Method	Cross-Entropy Loss								CIDEr Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
Autoregressive																
RSTNet$_{\text{CVPR21}}$ [40]	-	-	-	-	-	-	-	-	81.1	-	-	39.3	29.4	58.8	133.3	23.0
TF-Complete$_{\text{CVPR22}}$ [19]	-	-	-	-	-	-	-	-	80.2	-	-	38.8	29.0	58.3	129.5	22.7
CIIC$_{\text{CVPR22}}$ [20]	-	-	-	-	-	-	-	-	81.7	-	-	40.2	29.5	59.4	133.1	23.2
X-Transformer$^{\diamondsuit}{}_{\text{CVPR20}}$ [26]	78.3	62.9	49.3	38.2	29.2	58.3	124.5	22.6	82.0	67.2	53.1	41.2	30.2	60.0	137.2	24.2
ClipCap$^{\diamondsuit}$ [16]	-	-	-	33.5	27.5	-	113.1	21.1	-	-	-	-	-	-	-	-
I-Tuning$^{\diamondsuit}$ [17]	-	-	-	35.5	28.8	-	120.0	22.0	-	-	-	-	-	-	-	-
SMALLCAP$^{\diamondsuit}{}_{\text{CVPR23}}$ [18]	-	-	-	37.2	28.3	-	121.8	21.5	-	-	-	-	-	-	-	-
DTNet$_{\text{TNNLS24}}$ [42]	-	-	-	-	-	-	-	-	81.5	-	-	40.0	29.5	59.2	134.9	-
I2OA$_{\text{TIP25}}$ [43]	77.4	-	-	37.7	28.9	57.8	121.3	22.1	81.9	-	-	40.5	29.9	59.9	136.2	23.5
ViPCap$_{\text{AAAI25}}$ [35]	-	-	-	37.7	28.6	-	122.9	21.9	-	-	-	-	-	-	-	-
EVCAP$_{\text{CVPR24}}$ [34]	-	-	-	41.5	31.2	-	140.1	24.7	-	-	-	-	-	-	-	-
Non-Autoregressive																
CMAL$_{\text{IJCAI21}}$ [44]	78.5	-	-	35.3	27.3	56.9	115.5	20.8	80.3	-	-	37.3	28.1	58.0	124.0	21.8
SATIC$_{\text{ICCV21}}$ [27]	77.3	-	-	32.9	27.0	-	111.0	20.5	80.6	-	-	37.9	28.6	-	127.2	22.3
BitDiffusion$_{22}$ [21]	-	-	-	34.7	-	58.0	115.0	-	-	-	-	-	-	-	-	-
SCD-Net$_{\text{CVPR23}}$ [9]	79.0	63.4	49.1	37.3	28.1	58.0	118.0	21.6	81.3	66.1	51.5	39.4	29.2	59.1	131.6	23.0
SCD-Net$^{\diamondsuit}{}_{\text{CVPR23}}$ [9]	79.8	-	-	37.8	29.0	58.7	121.3	22.1	82.1	-	-	40.8	29.9	59.8	135.9	23.8
Our RDT$^{\diamondsuit}$	81.2	-	-	38.9	29.7	59.4	125.9	22.7	**82.7**	-	-	**41.5**	**30.5**	**60.4**	**139.3**	**24.4**

TABLE III
LEADERBOARD OF THE PUBLISHED STATE-OF-THE-ART IMAGE CAPTIONING MODELS ON THE COCO ONLINE TESTING SERVER, WHERE B@N, M, R AND C ARE SHORT FOR BLEU@N, METEOR, ROUGE-L AND CIDER SCORES. ◊ MEANS USING THE CLIP VISUAL FEATURES AS THE INPUT.

Method	B@1		B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Ensemble Model														
ETA$_{\text{ICCV19}}$ [47]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
AoANet$_{\text{ICCV18}}$ [45]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
M2 Transformer$_{\text{CVPR20}}$ [46]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer$_{\text{CVPR20}}$ [26]	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
DLCT$_{\text{AAAI21}}$ [13]	82.0	96.2	66.9	91.0	52.3	83.0	40.2	73.2	29.5	39.1	59.4	74.8	131.0	133.4
RSTNet$_{\text{CVPR21}}$ [40]	81.7	96.2	66.5	90.9	51.8	82.7	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4
VCT$_{\text{TCVST23}}$ [48]	82.2	96.2	67.2	91.2	52.7	83.5	40.6	73.8	29.6	39.3	59.6	75.0	132.0	134.5
Liu et al. $_{\text{TMM24}}$ [49]	82.5	96.7	68.3	92.3	54.2	84.8	42.2	75.5	30.2	40.2	61.0	77.0	136.3	138.0
Single Model														
CMAL$_{\text{IJCAI21}}$ [44]	79.8	94.3	63.8	87.2	48.8	77.2	36.8	66.1	27.9	36.4	57.6	72.0	119.3	121.2
CAVP$_{\text{ACMMM19}}$ [50]	80.1	94.9	64.7	88.8	50.0	79.7	37.9	69.0	28.1	37.0	58.2	73.1	121.6	123.8
SGAE$_{\text{CVPR19}}$ [51]	80.6	95.0	65.0	88.9	50.1	79.6	37.8	68.7	28.1	37.0	58.2	73.1	122.7	125.5
CIIC$_{\text{CVPR22}}$ [20]	-	-	-	-	-	-	38.5	70.1	29.1	38.4	58.6	74.0	126.4	129.2
SCD-Net$_{\text{CVPR23}}$ [9]	80.2	95.1	64.9	89.3	50.1	80.1	38.1	69.4	29.0	38.2	58.5	73.5	126.2	129.2
SCD-Net$^{\diamondsuit}{}_{\text{CVPR23}}$ [9]	81.2	95.9	66.2	90.3	51.7	82.6	39.2	71.9	29.8	39.2	59.2	74.3	130.9	133.2
Our RDT$^{\diamondsuit}$	**82.3**	**96.5**	**67.1**	**91.3**	**52.5**	**83.4**	**40.3**	**73.2**	**30.6**	**39.8**	**59.9**	**75.2**	**134.1**	**136.9**

Implementation Details. Our RDT implementation is based on SCD-Net [9], with two cascaded RDT models to further enhance captioning performance. The visual input is the CLIP feature[2] , whose grid feature dimension is transformed to 512. Following [21], [9], 14 bits are used to represent each word. We set three Ranking Visual Encoders, and both the Semantic Transformer and Sentence Decoder consist of 3 Transformer blocks with hidden size 512. In the RVE, λ_F and λ_E are set to 0.75 and 0.55, while the Ranking Loss parameters δ_{tim} and δ_{sen} are set to 0.12 and 0.15. To balance cross-entropy ($\mathcal{L}_{XE}$), binary ($\mathcal{L}_{bit}$), and ranking losses ($\mathcal{L}_{rank}$), λ_1, λ_2, and λ_3 are set to 1, 1, and 5, respectively.

The image-to-text retrieval block is built on CLIP ViT-B/32[3] features of images and captions, where captions are precomputed from COCO training data and indexed with FAISS [60] (IndexFlatIP, normalized, no training) for efficient nearest-neighbor search. For each image, the top 20 captions with the highest cosine similarity are retrieved. During training, 5 retrieval sentences are randomly sampled from the 5 ground-truth captions per image, while in inference the sentence with the highest similarity is used as the conditional input.

The training of RDT consists of two stages. In the first stage, the model is optimized with Adam on a single RTX 4090 GPU using ℓ_2, ranking, and cross-entropy losses for 60 epochs with a batch size of 16. In the second stage, parameters are initialized from the best first-stage model (highest CIDEr score) and further trained for 60 epochs using the guided self-critical sequence training mechanism [9], with a fixed learning rate of 0.00001 and batch size of 16. During training, $t' = t/T$ is sampled from $\mathcal{U}(0, 0.999)$. Algorithm 1 outlines the

[2]CLIP-RN101-448, https://github.com/jianjieluo/OpenAI-CLIP-Feature
[3]CLIP-ViT-B/32, https://github.com/jianjieluo/OpenAI-CLIP-Feature

TABLE IV
PERFORMANCE COMPARISONS WITH THE STATE-OF-THE-ART IMAGE CAPTIONING MODELS ON FLICKR30K AND NOCAPS.

Method	Flickr30k		NoCaps Val			
	Test		In	Near	Out	Entire
	C	S	C	C	C	C
Autoregressive						
ClipCap$^{\diamondsuit}$ [16]	-	-	84.9	66.8	49.1	65.8
I-Tuning$_{base}$ [17]	61.5	16.9	83.9	70.3	48.1	65.8
I-Tuning$_{medium}$ [17]	72.3	19.0	89.6	77.4	58.8	75.4
SMALLCAP$^{\diamondsuit}{}_{\text{CVPR23}}$ [18]	60.6	-	87.6	78.6	68.9	77.9
ViPCap$_{\text{AAAI25}}$ [35]	66.8	17.2	93.8	81.6	71.5	81.3
Non-Autoregressive						
SCD-Net$_{\text{CVPR23}}$ [9]	58.4	15.4	84.3	68.2	55.4	67.9
SCD-Net$^{\diamondsuit}{}_{\text{CVPR23}}$ [9]	61.5	16.7	87.9	76.4	68.1	76.3
Our RDT$^{\diamondsuit}$	**64.2**	**17.1**	**93.9**	**80.3**	**70.4**	**80.2**

TABLE V
IMPACT ON DIFFERENT PROPOSED MODELS ON COCO KARPATHY SPLIT.

base	RVE	RL	B@1	B@4	M	R	C	S	D@2	D@3	Voc-u
✓			82.1	40.8	29.9	59.8	135.9	23.8	11.6	24.5	10.3
✓	✓		82.5	41.2	30.3	60.2	138.1	24.2	12.4	25.8	10.9
✓		✓	82.4	41.4	30.3	60.1	138.4	24.1	12.5	26.0	11.0
✓	✓	✓	82.7	41.5	30.5	60.4	139.3	24.4	12.7	26.4	11.1

TABLE VI
ABLATION OF RANKING ATTENTION ON COCO KARPATHY TEST SPLIT.

λ_F	λ_E	B@1	B@4	M	R	C	S
0.6	0	79.9	37.9	29.1	58.7	122.2	22.0
0.7	0	80.0	37.9	29.2	58.9	123.8	22.2
0.75	**0**	**80.5**	**38.3**	**29.4**	**59.1**	**124.4**	**22.4**
0.8	0	80.4	38.2	29.3	59.9	124.2	22.3
0.9	0	80.3	38.1	29.2	59.0	123.8	22.3
1	**0**	**80.3**	**38.1**	**29.2**	**58.9**	**123.6**	**22.3**
0.75	0.25	80.9	38.6	29.5	59.2	125.1	22.5
0.75	0.4	81.1	38.7	29.6	59.2	125.5	22.5
0.75	0.5	81.1	38.8	29.6	59.3	125.7	22.6
0.75	**0.55**	**81.2**	**38.9**	**29.7**	**59.4**	**125.9**	**22.7**
0.75	0.6	80.5	38.4	29.5	59.1	124.8	22.4
0.75	1	80.3	38.2	29.3	59.0	123.9	22.4

sampling process, where $T = 50$ and $\Delta = 0$, and ST and SD denote the Sentence Transformer and Sentence Decoder, respectively. Table I provides an overview of the mathematical symbols used in this paper along with their definitions.

Detailed comparisons with state-of-the-art captioning methods are presented below.

B. Performance Comparison

In-domain. The in-domain evaluation is conducted on COCO (Tables II, III) and Flickr30k (Table IV). For COCO, we report both offline (Table II) and online (Table III) results, with separate reporting for the two training stages to ensure fair comparison. Since our RDT belongs to non-autoregressive models, we compare it against both autoregressive and non-autoregressive approaches. Overall, RDT achieves superior performance over state-of-the-art region-based [10], [45], [46], [26] and grid-based models [11], [13], [14]. Specifically, RDT$^{\diamondsuit}$ attains 139.3% CIDEr and 24.4% SPICE, surpassing strong autoregressive [46], [26], [13], [40], [19], [20], non-autoregressive [44], [27], [21], [9], and large-scale vision-language models [16], [17], [26]. Although EVCAP [34] slightly outperforms RDT, it leverages an external visual–name memory and large-scale pretrained LLMs (Vicuna–13B). Online results (Table III) further show that a single RDT model outperforms all single captioning baselines [44], [50], [51], [20], [9] and even ensemble models [46], [26], [13], [40]. On Flickr30k (Table IV), RDT also surpasses recent state-of-the-art models [18], [9], [35]. These results demonstrate that RDT generates more descriptive captions by enhancing visual features and aligning them effectively with language representations.

Cross-domain. The cross-domain evaluation is conducted on Nocaps using the COCO-trained first-stage model. As shown in Table IV, RDT surpasses state-of-the-art models [18], [9], [35] on in-domain data and achieves competitive results on other subsets. In particular, it improves over the baseline by 5.1% on near-domain and 3.4% on out-of-domain data, underscoring its effectiveness for zero-shot and real-world scenarios. These improvements are mainly attributed to 1) the RVE, which leverages forward- and inverse-ranking attention to recover or suppressed cues for more domain-invariant visual representations, and 2) the RL, which applies sentence-level ranking across diffusion timesteps to enhance semantic alignment, together fostering robust generalization to unseen domains.

Qualitative Analysis. We present image captioning results of three methods–the Transformer, the baseline (SCD-Net) [9], and our RDT–in Fig. 5(a). GT1, GT2, and GT3 denote the ground-truth captions. All three methods use the same grid features, and while the Transformer and SCD-Net produce semantically relevant but less descriptive captions, RDT leverages the RVE and RL to achieve fine-grained captioning. As shown in Fig. 5, RDT not only discovers new objects (e.g., two men behind him", red coat") but also captures detailed attributes (e.g., dry leaves", white frisbee", left handed baseball player", open umbrella"), leading to more accurate and descriptive captions.

C. Ablation Study

Ablation Study. To comprehensively evaluate our model, we start with the baseline SCD-Net [9] and progressively integrate the RVE and RL. As shown in Table V, adding RVE improves captioning performance, and incorporating both RVE and RL yields the full RDT, which achieves further gains. These results indicate that RVE enhances visual representations by extracting diverse and discriminative features, RL strengthens vision–language alignment by ranking caption quality across timesteps, and their integration enables fine-grained captioning that surpasses state-of-the-art models.

Ablation on Ranking Attention. By ranking and optimizing attended key-value pairs, the proposed RVE enhances visual representations for fine-grained captioning. To study the impact of λ_F and λ_E, we conduct experiments using CLIP features without Guided Self-Critical Sequence Training. As illustrated in Table VI, when $\lambda_F = 1$ and $\lambda_E = 0$, i.e., the original MHSA with the RL, the model outperforms SCD-Net$^{\diamondsuit}$ (Table II), confirming the effectiveness of the proposed RL. To

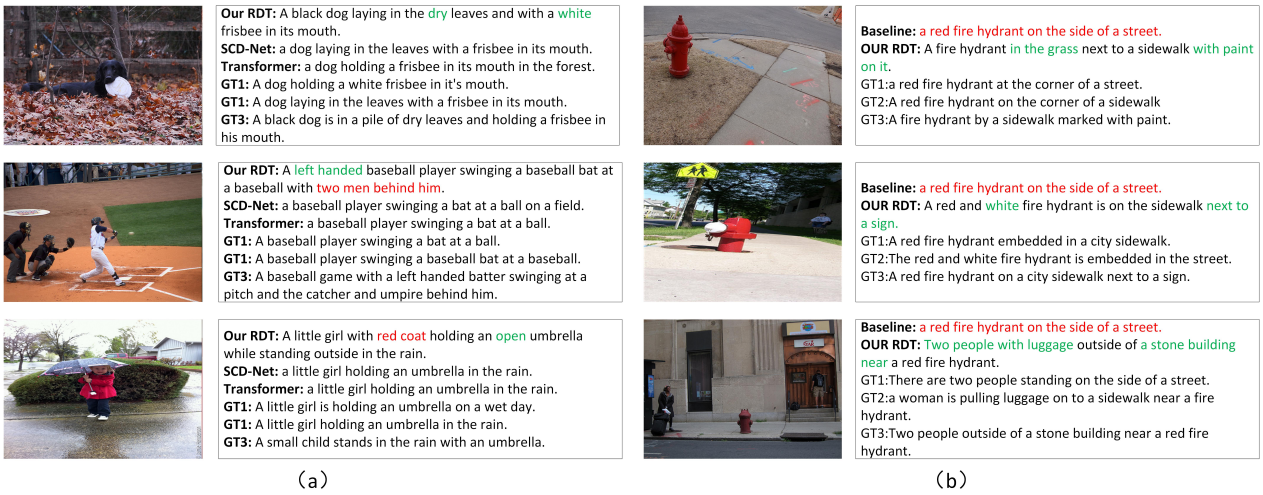

Fig. 5. Examples of image captioning results. (a) RDT identifies new objects (red) and detailed attributes (green), producing more descriptive captions. (b) RDT generates more discriminative captions for visually similar images.

determine the optimal λ_F, we fix $\lambda_E = 0$ and vary λ_F from 0.6 to 1, finding the best performance at $\lambda_F = 0.75$, which suggests that emphasizing important visual features improves captioning. Moreover, introducing inverse-ranking attention ($\lambda_E > 0$) further boosts performance, indicating that CLIP features ignore some visual details that can be reactivated and reweighted. When $0.25 < \lambda_E < 0.55$, RDT achieves better results than at $\lambda_E = 0.25$, highlighting the effectiveness of inverse attention. However, scaling λ_E to 1, i.e., concatenating O_F and all tokens O, yields worse results than using only forward-ranking attention ($\lambda_F = 0.75$, $\lambda_E = 0$), suggesting that the concatenation of O will hinder and interfere with the learning of more effective features. Overall, these results validate the effectiveness of both forward and inverse ranking attention, as well as their balanced combination.

D. Self Evaluations

Evaluation of image token visualization. The enhanced visual features has a dimension of 50×512, from which we select 49 grid features (excluding the CLS feature) of size 49×512. For visualization, we compute the mean along the last dimension, apply a softmax, and resize to 7×7. Fig. 6 compares the image tokens of SCD-Net (second column) and our RDT (third column). The highlighted tokens indicate the model's attention, revealing that RDT captures more detailed and discriminative features, enabling more descriptive captions. For example, in the first row, RDT attends to the background mountain and generates the phrase "with a cloud-mountain in the background," while the baseline neglects this detail.

Evaluation of Caption Sentence Discriminability. To evaluate the discriminability of generated captions, we select images with similar contents and present their captions in Fig. 5(b). For such images, the baseline (SCD-Net) produces nearly identical captions, indicating its failure to distinguish them. In contrast, RDT generates more fine-grained captions by extracting diverse and discriminative visual information and precisely aligning it with language features through the collaboration of the Ranking Visual Encoder and Ranking Loss within the proposed Ranking Diffusion Transformer.

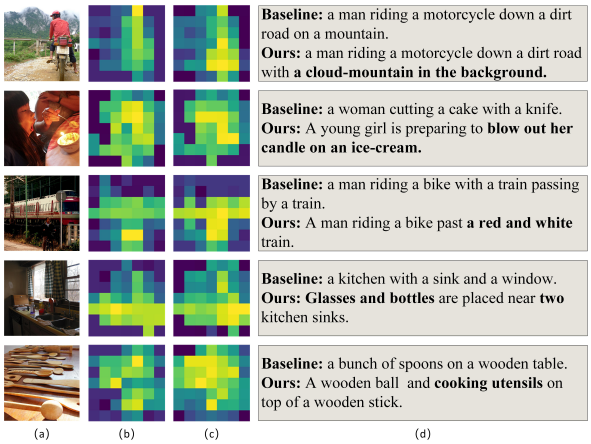

Fig. 6. (a) Image, (b) Image token visualization of the baseline–SCD-Net, (c) Image token visualization of our proposed RDT, (d) generated caption sentence. We can find that the highlighted tokens in the heatmap correspond to the regions where the captioning model focuses its attention, thereby capturing more detailed and discriminative visual features and producing more descriptive captions.

Evaluation of Caption Sentence Diversity. To evaluate caption diversity, we report Dist-2 (D@2), Dist-3 (D@3) [58], and vocabulary usage (Voc-u) [59]. Existing methods typically use top-k sampling, i.e., selecting the top 5 captions from beam search, while SCD-Net and our RDT generate multiple captions by randomly selecting 5 conditional sentences. As shown in Table VII, RDT outperforms both autoregressive and non-autoregressive methods, demonstrating its ability to produce more diverse captions.

Impact of timestep threshold δ_{tim} and sentence difference threshold δ_{sen}. The proposed Ranking Loss introduces a timestep threshold δ_{tim} and a sentence difference threshold δ_{sen} to generate masks M_{tim} and M_{sen}, which differentiate caption quality and provide fine-grained supervisory signals

TABLE VII
DIVERSITY COMPARISONS WITH THE STATE-OF-THE-ART IMAGE CAPTIONING MODELS ON COCO KARPATHY TEST SPLIT. ◊ MEANS USING THE CLIP VISUAL FEATURES AS THE VISUAL INPUT.

Method	Diversity		
	D@2 ↑	D@3 ↑	Voc-u ↑
	Autoregressive		
M2 Transformer$_{\text{CVPR20}}$ [46]	7.9	16.3	8.3
DLCT$_{\text{AAAI21}}$ [13]	8.1	17.1	8.6
DIFNet$_{\text{CVPR22}}$ [14]	9.3	19.5	9.1
CapDec[61]	8.3	14.9	1.9
ClipCap[16]	11.3	21.7	2.6
	Non-Autoregressive		
SCD-Net$_{\text{CVPR23}}$ [9]	10.1	22.6	9.7
SCD-Net$^{\diamondsuit}{}_{\text{CVPR23}}$ [9]	11.6	24.5	10.3
RDT$^{\diamondsuit}$ [9]	12.7	26.4	11.1

TABLE VIII
IMPACT OF TIMESTEP THRESHOLD δ_{tim} AND SENTENCE DIFFERENCE THRESHOLD δ_{sen} ON COCO KARPATHY TEST SPLIT.

δ_{tim}	δ_{sen}	B@1	B@4	M	R	C	S
0.05	-	80.2	38.3	29.3	59.0	123.2	22.3
0.10	-	80.5	38.6	29.5	59.2	124.8	22.4
0.12	-	**80.9**	**38.7**	**29.6**	**59.3**	**125.2**	**22.5**
0.15	-	80.5	38.5	29.6	59.2	124.9	22.5
0.20	-	80.4	38.4	29.5	59.1	124.5	22.4
0.12	0.05	80.4	38.4	29.4	59.1	123.7	22.3
0.12	0.1	81.0	38.7	29.6	59.2	125.3	22.5
0.12	0.12	81.1	38.9	29.6	59.3	125.7	22.6
0.12	**0.15**	**81.2**	**38.9**	**29.7**	**59.4**	**125.9**	**22.7**
0.12	0.2	81.1	38.8	29.6	59.3	125.4	22.5

for optimizing the diffusion process. To assess their impact, we first set $\delta_{sen} = 2$ (excluding M_{sen}) and vary δ_{tim} from 0.05 to 0.20, with the best performance at $\delta_{tim} = 0.12$ (Table VIII). We then fix $\delta_{tim} = 0.12$ and vary δ_{sen} from 0.05 to 0.20, achieving optimal performance at $\delta_{sen} = 0.15$. These results demonstrate that proper selection of δ_{tim} and δ_{sen} is crucial for providing effective supervisory signals to optimize the diffusion process for fine-grained captioning.

Time and Memory Analysis About MHRA. The time complexity of multi-head self-attention (MHSA) consists of similarity calculation ($O(n^2d)$), softmax ($O(n^2)$), and weighted summation ($O(n^2d)$), resulting in $O(n^2d)$ overall, where n is the number of grids in the CLIP feature and d is the feature dimension. For our RDT, $n = 50$ and $d = 512$. The proposed multi-head ranking attention (MHRA) adds extra operations (e.g., matrix addition, outer product, top-K ranking, gathering), with a final complexity of $O(n^2d + n^3)$. Thus, both MHSA and MHRA scale linearly with d, but MHRA has higher cost; since $n < d$, this increase is limited. In summary, RDT improves captioning performance at the expense of additional computation. The space complexity of both MHSA and MHRA is $O(n^2 + nd)$, with MHRA incurring slightly higher cost for better performance.

V. CONCLUSION

Discriminative and descriptive image captioning remains a very challenging task due to limited visual information and inefficient vision-language alignment. In this work, we propose a Ranking Diffusion Transformer to achieve fine-grained image captioning by seamlessly integrating the Ranking Visual Encoder (RVE) and Ranking Loss (RL). The well-designed RVE effectively strengthens visual representations by mining diverse and discriminative information from them through ranking and optimizing the attended key-value pairs. Then, a novel RL is proposed to optimize the diffusion process while boosting the vision-language semantic alignment by taking the quality difference ranking results of the generated captions as additional supervisory signals. Hence, by cooperating RVE and RL via a novel Ranking Diffusion Transformer, more representative and discriminative visual features can be learned and precisely aligned with language features to achieve fine-grained image captioning. In the future, we plan to conduct research on endowing human-like controllability to the captioning model and extend our model to other related topics.

REFERENCES

[1] Z. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 710–722, 2019.

[2] J. Cho, S. Yoon, A. Kale, F. Dernoncourt, T. Bui, and M. Bansal, "Fine-grained image captioning with clip reward," *ArXiv*, vol. abs/2205.13115, 2022.

[3] J. Wan, Z. Lai, J. Liu, J. Zhou, and C. Gao, "Robust face alignment by multi-order high-precision hourglass network," *IEEE Transactions on Image Processing*, vol. 30, pp. 121–133, 2020.

[4] J. Wan, H. Xi, J. Zhou, Z. Lai, W. Pedrycz, X. Wang, and H. Sun, "Robust and precise facial landmark detection by self-calibrated pose attention network," *IEEE Transactions on Cybernetics*, vol. 53, pp. 3546–3560, 2021.

[5] A. Iovine, F. Narducci, M. Degemmis, and G. Semeraro, "Humanoid robots and conversational recommender systems: a preliminary study," *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pp. 1–7, 2020.

[6] M. J. Amon, R. Hasan, K. Hugenberg, B. I. Bertenthal, and A. Kapadia, "Influencing photo sharing decisions on social media: A case of paradoxical findings," *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 1350–1366, 2020.

[7] F. Li, Z. Sun, A. Li, B. Niu, H. Li, and G. Cao, "Hideme: Privacy-preserving photo sharing on social networks," *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 154–162, 2019.

[8] M. Afif, R. Ayachi, E. E. Pissaloux, Y. Said, and M. Atri, "Indoor objects detection and recognition for an ict mobility assistance of visually impaired people," *Multimedia Tools and Applications*, vol. 79, pp. 31 645 – 31 662, 2020.

[9] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei, "Semantic-conditional diffusion networks for image captioning*," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23 359–23 368, 2022.

[10] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2017.

[11] H. Jiang, I. Misra, M. Rohrbach, E. G. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 264–10 273, 2020.

[12] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[13] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," *ArXiv*, vol. abs/2101.06462, 2021.

[14] M.-K. Wu, X. Zhang, X. Sun, Y. Zhou, C. Chen, J. Gu, X. Sun, and R. Ji, "Difnet: Boosting visual information flow for image captioning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 999–18 008, 2022.

[15] Y. Li, Y. Pan, T. Yao, and T. Mei, "Comprehending and ordering semantics for image captioning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 969–17 978, 2022.

[16] R. Mokady and A. Hertz, "Clipcap: Clip prefix for image captioning," *ArXiv*, vol. abs/2111.09734, 2021.

[17] Z. Luo, Y. Xi, R. Zhang, and J. Ma, "I-tuning: Tuning language models with image for caption generation," *ArXiv*, vol. abs/2202.06574, 2022.

[18] R. P. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva, "Smallcap: Lightweight image captioning prompted with retrieval augmentation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2840–2849, 2022.

[19] Z. Fei, X. Yan, S. Wang, and Q. Tian, "Deecap: Dynamic early exiting for efficient image captioning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 206–12 216, 2022.

[20] B. Liu, D. Wang, X. Yang, Y. Zhou, R. Yao, Z. Shao, and J. Zhao, "Show, deconfound and tell: Image captioning with causal inference," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 020–18 029, 2022.

[21] T. Chen, R. Zhang, and G. E. Hinton, "Analog bits: Generating discrete data using diffusion models with self-conditioning," *ArXiv*, vol. abs/2208.04202, 2022.

[22] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[23] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *International Journal of Computer Vision*, vol. 123, pp. 74 – 93, 2015.

[24] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "nocaps: novel object captioning at scale," *International Conference on Computer Vision*, pp. 8947–8956, 2019.

[25] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Neural Information Processing Systems*, 2019.

[26] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 968–10 977, 2020.

[27] Y. Zhou, Y. Zhang, Z. Hu, and M. Wang, "Semi-autoregressive transformer for image captioning," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3132–3136, 2021.

[28] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, "Improving image captioning by leveraging intra- and inter-layer global representation in transformer network," *ArXiv*, vol. abs/2012.07061, 2020.

[29] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, "Chinese image caption generation via visual attention and topic modeling," *IEEE Transactions on Cybernetics*, vol. 52, pp. 1247–1257, 2020.

[30] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022.

[31] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," *ArXiv*, vol. abs/2004.06165, 2020.

[32] L. Nie, F. Jiao, W. Wang, Y. Wang, and Q. Tian, "Conversational image search," *IEEE Transactions on Image Processing*, vol. 30, pp. 7732–7743, 2021.

[33] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, vol. 2022, 2022.

[34] J. Li, D. M. Vo, A. Sugimoto, and H. Nakayama, "Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 733–13 742, 2023.

[35] T. Kim, S. Lee, S.-W. Kim, and D.-J. Kim, "Vipcap: Retrieval text-based visual prompts for lightweight image captioning," in *AAAI Conference on Artificial Intelligence*, 2025.

[36] J. Gu, J. Bradbury, C. Xiong, V. O. K. Li, and R. Socher, "Non-autoregressive neural machine translation," *ArXiv*, vol. abs/1711.02281, 2017.

[37] J. Gao, X. Meng, S. Wang, X. Li, S. Wang, S. Ma, and W. Gao, "Masked non-autoregressive image captioning," *ArXiv*, vol. abs/1906.00717, 2019.

[38] F. Liu, X. Ren, X. Wu, B. Yang, S. Ge, and X. Sun, "O2na: An object-oriented non-autoregressive approach for controllable video captioning," in *Findings*, 2021.

[39] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *ArXiv*, vol. abs/2107.00630, 2021.

[40] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "Rstnet: Captioning with adaptive attention on visual and non-visual words," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 460–15 469, 2021.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.

[42] Y. Ma, J. Ji, X. Sun, Y. Zhou, X. Hong, Y. Wu, and R. Ji, "Image captioning via dynamic path customization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, pp. 6203–6217, 2024.

[43] X. Zhang, A. Jia, J. Ji, L. Qu, and Q. Ye, "Intra- and inter-head orthogonal attention for image captioning," *IEEE Transactions on Image Processing*, vol. 34, pp. 594–607, 2025.

[44] L. Guo, J. Liu, X. Zhu, X. He, J. Jiang, and H. Lu, "Non-autoregressive image captioning with counterfactuals-critical multi-agent learning," in *International Joint Conference on Artificial Intelligence*, 2020.

[45] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4633–4642, 2019.

[46] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 575–10 584, 2019.

[47] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8927–8936, 2019.

[48] S. Cao, G. An, Z. Zheng, and Z. Wang, "Vision-enhanced and consensus-aware transformer for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 7005–7018, 2023.

[49] A. Liu, Y. Zhai, N. Xu, H. Tian, W. zhi Nie, and Y. Zhang, "Event-aware retrospective learning for knowledge-based image captioning," *IEEE Transactions on Multimedia*, vol. 26, pp. 4898–4911, 2024.

[50] D. Liu, Z. Zha, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for sequence-level image captioning," *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[51] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 677–10 686, 2018.

[52] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, 2014.

[53] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Annual Meeting of the Association for Computational Linguistics*, 2002.

[54] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *WMT@ACL*, 2007.

[55] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Annual Meeting of the Association for Computational Linguistics*, 2004.

[56] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2014.

[57] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," *ArXiv*, vol. abs/1607.08822, 2016.

[58] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan, "A diversity-promoting objective function for neural conversation models," *ArXiv*, vol. abs/1510.03055, 2015.

[59] B. Dai, S. Fidler, and D. Lin, "A neural compositional paradigm for image captioning," *ArXiv*, vol. abs/1810.09630, 2018.

[60] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, pp. 535–547, 2017.

[61] J. Lee, E. Mansimov, and K. Cho, "Deterministic non-autoregressive neural sequence modeling by iterative refinement," *ArXiv*, vol. abs/1802.06901, 2018.