This ACCV 2024 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2024
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

M-RAT: a Multi-grained Retrieval Augmentation Transformer for Image Captioning

Jiayan Song[1,2]*, Renjie Pan[1,2]*, Jun Zhou[1,2], and Hua Yang[1,2]†

[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2] Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai 200240, China

Abstract. Current encoder-decoder methods for image captioning mainly consist of an object detection module (two-stage), or rely on big models with large-scale datasets to improve the effectiveness, which leads to increasing computation costs and cannot introduce new external knowledge. In this paper, we propose a novel end-to-end method Multi-grained Retrieval Augmentation Transformer (M-RAT) that innovatively fuses retrieved text derived from a changeable datastore with input visual feature through a Multi-modal Aligned Encoder, and introduce a specialized attention mechanism, Multi-MSA, to exploit both local and global interactions for delicate fine-grained details. Additionally, we enhance the decoder generation ability by employing low-level and high-level fused embeddings. Experiments demonstrate that M-RAT achieves comparable performance to state-of-the-art baselines with remarkable accuracy and details, as well as showing excellent domain adaptability for novel objects.

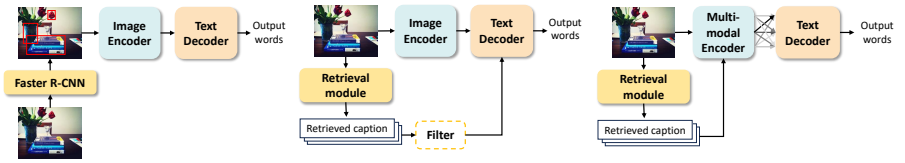
Keywords: Image Captioning · Retrieval Augmentation · Computer Vision

1 Introduction

Driven by the rapid development in Computer Vision (CV) and Natural Language Processing (NLP), multi-modal learning[6, 2, 35] has garnered significant attention in both research and industrial community. Image captioning[49, 21, 44], as one of the fundamental cross-modal tasks, aims to automatically describe the visual content of a given image with fluent and coherent sentences.

Existing image captioning methods can be roughly categorized into two categories. (1) **Encoder-decoder Transformer-based methods**[4, 7, 18, 37, 51] consist of a two-stage learning process. Most of the existing methods first extract region-level features through Object Detection module[15, 14, 43], which are then fed into Transformer-based[47] image encoder and text decoder with various backbones[10, 32, 8] to generate textual descriptions, as shown in fig. 1(a).

*Equal contribution. †Corresponding author.



(a) Encoder-Decoder models **(b)** Retrieval Augmentation based models **(c)** Multi-grained Retrieval Augmentation Transformer

Fig. 1: Three paradigms of image captioning models. Our Multi-grained Retrieval Augmentation Transformer (M-RAT) enhances the traditional encoder-decoder architecture by incorporating additional linguistic information and distinguishes itself from other retrieval augmentation-based models by employing a multi-modal encoder.

However, it remains challenging for Encoder-decoder Transformer-based methods to learn in an end-to-end manner. Beside, it is observed that the region-level features may not include sufficient local-to-global information from a holistic perspective, leading to a lack of cross-grained interaction.With the development of Vision Language Pre-train (VLP)[39, 25, 27, 17, 52, 26], image captioning has been improved by scaling up models with large-scale datasets. However, it inevitably leads to exponentially increasing computation costs. More importantly, VLP is unable to incorporate new knowledge and novel objects after pre-training stage, resulting in difficulty in achieving an up-to-date image captioning with constantly updated knowledge. Therefore, (2) **Retrieval Augmentation-based methods**[42, 54, 41, 45, 29, 24] has emerged, which add to a retrieval module to retrieve potentially relevant textual information from a datastore, thereby facilitating the generating process, as shown in fig. 1(b). Although Retrieval-Augmentation based methods may achieve a trade-off between efficiency and effectiveness, the input visual features cannot be fully exploited and interacted with the retrieved information due to their direct incorporation into the decoder[42, 41]. Apart from this, the uncertainty in the retrieval information can potentially affect the accuracy of the generated output, necessitating an additional filter to mitigate these effects[29]. Based on this fact, the key to an effective and efficient image captioning lies in a dynamic multi-modal interaction in the encoder and decoder.

To this end, we proposed a novel end-to-end method **Multi-grained Retrieval Augmentation Transformer (M-RAT)**, as shown in fig. 1(c), and further delve into the impact of retrieval augmentation on image captioning. Unlike previous Retrieval Augmentation-based methods[42, 45, 54], we consider the retrieved text derived from a changeable datastore as an extra input, and employ local and global-grained fusion with image feature through a specific Multi-modal Aligned Encoder with Multi-MSA. With the supervision of the joint low-level and high-level representations obtained by the Multi-modal Aligned Encoder, the decoder adopt beam-search to generate captions word-by-word, considering external related semantic information apart from the image alone.

Extensive experiments on COCO[31] "Karpathy"[22] test split demonstrate that our M-RAT achieves comparable performance to state-of-the-art baselines

with remarkable accuracy and fine-grained details. The retrieved text contributes greatly to an excellent domain adaptability for novel objects on Nocaps[1] and Flickr30K[56], even without fine-tuning or re-training. Furthermore, we find that a simple Importance Sample and aligned-fusion encoding during training stage of the retrieval-augmentation based models is beneficial to exploiting crucial information from retrieval results while eliminating redundant noise. Therefore, M-RAT eliminates the need for additional selection of retrieval results, which is indispensable for ensuring the accuracy of the generated sentences in general retrieval-augmentation based models[45]. Also, M-RAT possesses significant potential in end-to-end transformer-based model for novel objects captioning, further promoting the development of multi-modal encoders in vision-language generative tasks.

Our main contributions are summarized as follows:

- We proposed a novel end-to-end method Multi-grained Retrieval Augmentation Transformer (M-RAT) that innovatively fuses changeable retrieved text with input image through a Multi-modal Aligned Encoder, which can enable M-RAT to extract the retrieved essence and discard the irrelevant information effectively.
- We introduce a specialized attention mechanism, Multi-MSA, to exploit both local and global interactions in order to ensure that M-RAT pays adequate attention to delicate fine-grained details. Additionally, we employ joint low-level and high-level representations for supervision during the decoder generation stage.
- Experiments demonstrate that our M-RAT achieves comparable performance to state-of-the-art baselines with remarkable accuracy and details. Furthermore, M-RAT shows excellent domain adaptability for novel objects, even without fine-tuning or re-training, further verifying the flexibility, relevance, and efficiency of our model.

2 Related Work

2.1 Image Captioning

Early methods usually regarded image captioning[46] as a sequence-to-sequence problem, requiring an encoder-decoder[20, 55, 21, 49, 4] or parallel encoders[33] framework with specific backbone[19, 43], to completely extract image features and generate captions word-by-word under the guidance of visual information. The swift progress of the Transformer-based architecture has propelled a massive number of competitive methods in exploring various attention mechanism and its variants[7, 18, 37], as well as the research on a range of backbones with powerful representational capacity[10, 32, 8]. Nevertheless, the initial exploration of deep learning methods exhibited a simplistic architecture, albeit with limited adaptability in general tasks.

Recently, thanks to the high efficiency of parallel computing of Transformer, advanced image captioning based on Vision-Language Pre-train (VLP) methods[27, 52, 39, 23, 17] focuses on big models and large-scale datasets with some

strategies for efficient fine-tuning[9] and data denoising[26, 25]. Practice illustrate that the availability of massive data provides rich prior knowledge and the expansion of model size further boosts cross-modal fusion. However, the VLP model heavily relies on the model size, posing challenges to the computational costs.

2.2 Retrieval Augmentation based Models

Some efforts[36, 34, 42] have been made to strike a balance between the efficiency and effectiveness of image captioning by utilizing the off-the-shelf pre-trained models[39, 40]. Researchers have gradually discovered that retrieval augmentation, which has been broadly used in NLP[53, 16], appears to be a reliable alternative to help models maintain knowledge updates and fully uncover open-world comprehension for flexible application scenarios. The aim of retrieval augmentation is to empower the language model by adding extra textual information retrieved from an external memory, rather than relying exclusively on input visual features.

In previous works, image-image[45], image-text[41] and object category retrieval[24] are commonly employed to obtain external information, which was then utilized as a prompt to the decoder with the assistance of cross-attention for generation[45, 29, 54]. However, the noise carried alongside the retrieval information undoubtedly introduces interference and impact the accuracy of the results. Additionally, an efficient construction of the retrieval database is crucial, as storing visual knowledge may increase memory consumption than textual knowledge. Hence, there still remains to be explored on how to retrieve and what to retrieve.

3 Our Approach

3.1 Overview

The overall framework of M-RAT is shown in fig. 2, which is composed of a Retrieval Augmentation module based on CLIP, a Feature Extraction module (section 3.2), a Multi-modal Aligned Encoder and a Decoder made of stacks of attention layers(section 3.3). Retrieval Augmentation module is employed for image-text retrieval from a changeable datastore full of ample text information and diverse domain classes. Then the initial visual features and word embedding vectors can be obtained through the feature extraction module. The above two parts can be regarded as the pre-processing for V&L Alignment. All intra-modality and cross-modality interactions are constructed via a special scaled dot-product attention mechanism in Multi-modal Aligned Encoder, particularly the fusion of local (word-word, region-region, region-word) and global (image-text) relationships. Finally, a Decoder leverages both low-level and high-level joint representations of encoder output simultaneously with a meshed connection.

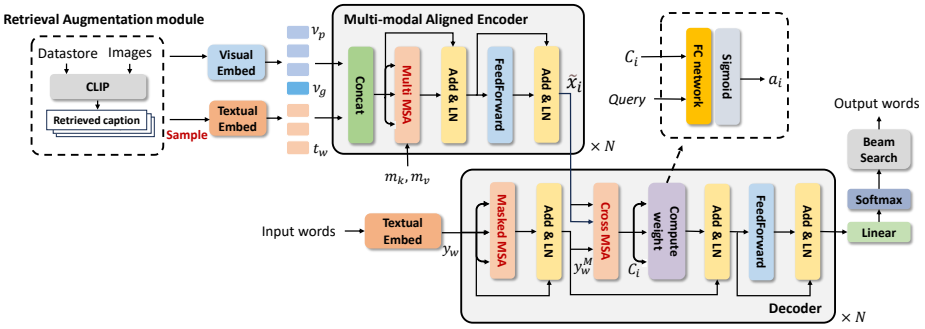


Fig. 2: **Architecture of M-RAT.** The retrieved caption derived from retrieval augmentation module is co-encoded with image features in Multi-modal Aligned Encoder. Through a meshed connection, the Decoder adopt beam-search to generate captions with the supervision of the low-level and high-level fused embedding.

3.2 Pre-process for V&L Alignment

Retrieval Augmentation module. We choose to employ image-text retrieval based on the comparison between uni-modal and multi-modal retrieval mentioned in section 2.2. As shown in fig. 2, a parameter-frozen CLIP[39] is utilized to map two modalities into a joint embedding space and match the top-k captions by nearest neighbor search based on cosine similarity, which are closest to the input image from a chosen datastore. It is worth mentioning that to prevent the model from "cheating", we filter out the *ground-truth* to make sure that all the retrieved results consist exclusively of hard negative samples.

In order to improve the retrieval efficiency, we leverage *Faiss*[11] for fast retrieval, which is proved to be faster than brute force retrieval under large-scale datasets, with an average of 0.023 seconds per image. The text categories and styles in the retrieval datastore can be dynamically adjusted during training or inference based on the user's application scenarios and generation requirements. This greatly enhances the real-time update capability of the model to adapt to new knowledge. The impact between generated captions and retrieval datastore are discussed in detail in section 4.4.

Importance Sample. We observe that for the same input image, the retrieval results always remain consistent without modifying the datastore. However, solely relying on fixed retrieval captions leads to the model's insensitivity, as well as resulting in inaccurate generations. We speculate this is because the embeddings of the retrieved text in datastore and input image features are separated by a domain gap. To bridge this gap, we employ an Importance Sampling approach to "inject noise", which can be expressed as follows:

$$\mathrm{Top}_3(S) = \{S_i, i \in \mathrm{argmax}_3(V^\top T_i)\}, \tag{1}$$

$$\mathrm{RetrCap} = S \sim \mathrm{Top}_3(S). \tag{2}$$

The simple design of Importance Sample enables the model to effectively mitigate the domain gap during training effectively with distinct retrieval captions.

Feature Extraction module. Feature Extraction module is employed for the preliminary encoding and standardizing dimensions, laying the foundation for subsequent joint representation learning. Vision Transformer [10] is employed to extract a set of patch features and global features $V = \{v_1, v_2, \ldots, v_m, v_g\}$ from a given input image as the initial visual feature, where the i-th patch feature $v_i \in \mathbb{R}^D, i \in \{1, m\}$, and global feature $v_g \in \mathbb{R}^D$. D is the embedding dimension of each vector. Different from existing methods that apply global features, we observe that, for image captioning, local-grained patch information is more advantageous than global image information. Captioning facilitates models to incorporate more details and spatial relationships, mitigating the risk of misjudgment due to incomplete target overlap or low pixel resolution. To represent words, we use one-hot decoding and linearly project them to the same input dimension of image features D. We also sum it with sinusoidal positional encoding to obtain the word-level text embedding sequence $T = \{t_1, t_2, \ldots, t_n\}$ where the i-th word feature $t_i \in \mathbb{R}^D$, too.

3.3 Design of Encoder and Decoder

As for the design of Encoder and Decoder, We take into full consideration the importance of local and global features as well as low-level and high-level representations. To be more specific, we employ Multi-MSA, Masked-MSA, and Cross-MSA at the corresponding positions of encoder and decoder. As shown in Fig. 2, different colors represent different structures.

Multi-modal Aligned Encoder. To comprehensively consider global (image-text) interaction and other external prior knowledge, we incorporate a learnable matrix concatenation into the attention mechanism to strengthen the flexibility of encoding.

The Multi-modal Aligned Encoder is responsible for aligned-fusion between image features and retrieval text embeddings. For two dimensionally concatenated features, both intra-modality and inter-modality interactions can be modeled via traditional scaled dot-product attention. However, the fused embeddings obtained by this method solely rely on the input features, which are derived from image patches and text words. Therefore, it is challenging to encode the global relationships between vision and language during the representation learning. To handle this, we introduce a specific MultiMSA.

Specifically, given a set of input $X = \{x_1, x_2, \ldots, x_i\}$, where $x_i = [V_i, T_i] \in \mathbb{R}^{(m+1+n)\times D}$ is formed by concatenating initial image feature V_i and text embedding T_i, i means i-th image-text pair. Then MultiMSA is expressed as follows:

$$\text{MultiMSA}(Q, K, V) = \text{MSA}(W_q X, K, V), \tag{3}$$

$$K = [W_k X, M_k], V = [W_v X, M_v], \tag{4}$$

where $\mathrm{MSA}(\cdot,\cdot,\cdot)$ is the normal attention mechanism, M_k, M_v are additional learnable parameter matrices independent of the input data. In this way, we can flexibly insert additional "slots" to encode extra key information, such as global relationships, retrieval similarity, and prior knowledge. The overall structure of the encoder is composed of stacked layers. The final output is $\widetilde{X} = (\widetilde{X}^1, \widetilde{X}^2, \ldots, \widetilde{X}^N)$, where N is the number of encoder layers and $\widetilde{X}^j$ refers to the output of j-th layer, including low-level and high-level aligned multi-modal features.

Decoder. Taken the previously generated word embedding vector Y and multi-modal aligned feature $\widetilde{X} = (\widetilde{X}^1, \widetilde{X}^2, \ldots, \widetilde{X}^N)$ into account, the probability distribution of word prediction at each timestep is influenced by the preceding text information, as well as the encoding characteristics of input image described by retrieval captions.

To fully exploit the low-level and high-level representations acquired from the encoder, we refine a weight-matching decoder structure. Specifically, within a decoder layer, Y is fed into a MaskedMSA to acquire contextual semantic information $\widehat{Y}$, followed by CrossMSA to obtain a weight α_j in the output of the j-th encoder layer. The detailed expression can be denoted as follows:

$$\mathrm{CrossMSA}(\widetilde{X}^j, \widehat{Y}) = \mathrm{MSA}(W_q\widehat{Y}, W_k\widetilde{X}^j, W_v\widetilde{X}^j), \tag{5}$$

$$\alpha_j = \sigma\mathrm{FC}[W_q\widehat{Y}, \mathrm{CrossMSA}(\widetilde{X}^j, \widehat{Y})], \tag{6}$$

where $\mathrm{CrossMSA}(\cdot)$ represents cross attention, $[\cdot,\cdot]$ represents the tensor concatenation, σ is sigmoid activation. The weight α_j not only mark the individual contribution of $\widetilde{X}^j$ but also illustrate the relative attention difference between decoder layers when performing prediction. Finally, the decoder outputs the vocabulary probability distribution and obtains the current predicted word by maintaining a fixed-size candidate queue according to Beam Search[4] algorithm, which can be expressed as follows:

$$S_t = \text{Top-B}\{(y_{1:t-1}, y_t) | y_{1:t-1} \in S_{t-1}, y_t \in \text{Vocab}\}. \tag{7}$$

3.4 Training strategy

Following previous practice[4], we first apply Cross-Entropy loss to initially complete pre-training.

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* | y_{1:t-1}^*, \widetilde{X})), \tag{8}$$

where $y_{1:T}^*$ is the target ground truth sequence, and θ denotes the parameters of our model. Then the model is fine-tuned according to the Self-Critical Sequence Training(SCST)[44], which is a commonly used Reinforcement Learning algorithm in sequence generation tasks and facilitates the utilization of non-differentiable metrics as optimization objectives.

$$L_{SCST}(\theta) = -\mathbf{E}_{y_{1:T}\sim p_\theta}[r(y_{1:T})], \tag{9}$$

$$\nabla_\theta L_{SCST}(\theta) \approx -(r(y_{1:T}^s) - b)\nabla_\theta \log p_\theta(y_{1:T}^s), \tag{10}$$

where $r(\cdot)$ is the score of CIDEr because it exhibits a strong correlation with human judgment. $y^s_{1:T}$ is a sampled caption and $b = (\sum_i^k r(y^s_{1:T}))/k$ depicts the baseline by calculating the average reward of the sampled sequence.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on COCO dataset[31] with *Karpathy*[22] split, and Flickr30K [56] is used in the test stage. To evaluate the domain adaption of novel objects, the widely adopted NOCAPS dataset [1] is utilized. Furthermore, we also employ self-built image datasets to assess and compare model's captioning in challenging data qualitatively. Self-built datasets comprise the Pedestrian Group dataset, which is recorded by traffic cameras, and the Virtual Game dataset consists of several simulated city scenes. Self-built datasets can effectively mitigate the risk of *data leakage*. The distribution of images in the Virtual Game dataset is different from that of real-world scenes, which makes them challenging to label. The low resolution further exacerbates their difficulty. We construct a datastore for retrieval, consisting of 802K captions. The detail is elaborated in table 1 and section 4.4.

Metrics. For the purpose of measuring the Fluency, Precision and Recall of generated captions from different perspectives, we use the full set of captioning metrics entailing BLEU[38], METEOR[5], ROUGE[30], CIDEr[48] and SPICE[3]. Among them, CIDEr is more emphasized in image captioning because it can capture key information.

Implementation Details. All experiments are performed by using the PyTorch toolkit. Retrieval Augmentation module consists of pre-trained CLIP[39] with ResNet-50x64 backbone. We apply a pre-trained ViT-B/32 [10] to extract image features, the dimension of each patch D is 768, and the number of patches m is 7×7. To obtain the word-level text embeddings, we linearly map the input words with one-hot decoding and sinusoidal position embedding. In M-RAT, We set the dimension of the Transformer layer to 512 and the head to 8, with 2048-dimension hidden layer of the feed-forward network, and the dropout rate is set to 0.1. We follow the commonly used learning rate tuning strategy[47], where the parameter warm-up consists of 10000 iterations. During SCST, the learning rate is fixed to 3×10^{-7}. We adopt Adam optimizer for optimization with batch-size of 64. To strike a balance between performance and efficiency, we ultimately design 5 beams and 2 decoder layers. The entire training process can be executed on a single A100 GPU.

Table 1: Statistics of used Datasets for Experiments and Retrieval Datastore.

Source	COCO all	Flickr30K test,retrieval	Nocaps test	Self-built test	VATEX retrieval	Clotho retrieval	TGIF retrieval	VizWiz retrieval
Image/Text	112k/560k	31k/155k	4.5k/45k	10k/–	–/349k	–/14k	–/125k	–/117k

Table 2: Quantitative comparison with SOTA methods on common image captioning benchmark COCO. * denotes using **CIDEr Optimization**. We report the size of parameters and training data, with BLEU@1 (B@1), BLEU@4 (B@4), METEOR (M), ROUGE(R), CIDEr (C), and SPICE (S) scores on COCO test set. score indicates the best results among compared methods, score indicates the second best results.

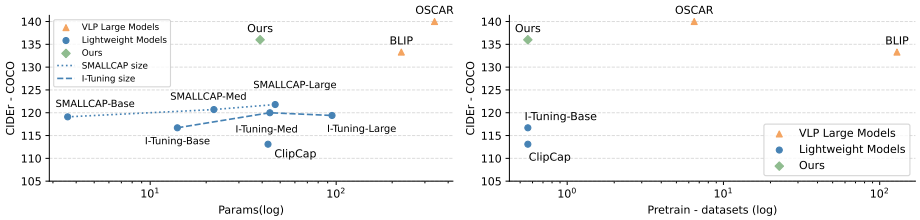
Method	**Training**		**COCO**[a]					
	Para.	Data	Test					
			B@1	B@4	M	R	C	S
Encoder-Decoder models								
SCST* [44]	–	–	–	34.2	26.7	55.7	114.0	–
Up-Down [4]	–	–	79.8	36.3	27.7	56.9	120.1	21.4
RFNet [20]	–	–	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM [55]	–	–	80.5	38.2	28.5	58.3	127.6	22.0
AoANet* [18]	–	–	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer* [7]	–	–	80.8	39.1	29.2	58.6	131.2	22.6
X-Transformer* [37]	–	–	80.9	39.7	29.5	59.1	132.8	23.4
RSTNet*$_{\text{ResNext101}}$ [57]	–	–	81.1	39.3	29.4	58.8	133.3	23.0
Retrieval Augmentation based models								
SmallCap$_{\text{Base}}$ [42]	7M	COCO	–	37.0	27.9	–	119.7	21.3
Re-ViLM$_{\text{Base}}$ [54]	158M	119M	–	37.8	–	–	129.1	–
EXTRA*$_{\text{K=5}}$ [41]	–	COCO	–	36.4	28.2	–	131.1	21.3
Ours*	39M	COCO	81.8	39.9	29.7	59.4	136.1	23.6
Lightweight and VLP models								
ClipCap [36]	43M	COCO	–	33.5	27.4	–	113.1	21.1
I-Tuning$_{\text{Base}}$ [34]	14M	COCO	–	35.2	28.5	–	118.3	22.0
Oscar*$_{\text{Large}}$ [27]	338M	6.5M	–	41.7	30.6	–	140.0	24.5
BLIP$_{\text{CapFlit-L}}$ [25]	224M	129M	–	39.7	–	–	133.3	–

[a] COCO with *Karpathy* split contains 0.56M image-text pairs for training, 5k images for validation and 5k images for test.

4.2 Analysis of Results

Comparison with State-of-the-art. We conduct performance comparisons with several state-of-the-art image captioning methods, as shown in table 2, where the size of training datasets, model parameters, and various metrics are included. When comparing M-RAT with encoder-decoder models and retrieval augmentation based models, it can be observed that *M-RAT outperforms most of the baselines in terms of all metrics*. This highlights the superiority of the design of M-RAT and confirms the effectiveness of similar external contextual information derived from Retrieval Augmentation module.

Regarding other advanced image captioning methods from table 2 and fig. 3, it is evident that VLP based methods typically exhibit enormous model scales. Lightweight models commonly leverage large language models to help generate captions. In contrast, M-RAT achieves a balance between model size and data



(a) CIDEr scores on different parameters. (b) CIDEr scores on different datasets.

Fig. 3: Our model's performance on the COCO test dataset, compared to other approaches in terms of trainable parameters and pre-train datasets. Our model is competitive with other lightweight models and is comparable to VLP models, even with a smaller model size.

size compared with VLP methods (39M vs 224/338M para., 0.56M vs 129/6.5M data), as well as achieving excellent performance. Furthermore, M-RAT even surpasses quite a few VLP methods in terms of BLEU-4 (39.9) and CIDEr (136.1) scores, signifying the outstanding performance of M-RAT on general dataset. Notably, M-RAT achieves superior generation results of CIDEr score (136.1 vs 113.1) among lightweight models.

Zero-shot Transfer on Novel Objects. NoCaps and Flickr30K datasets always serve as dedicated test set for evaluating zero-shot transfer ability of models trained on COCO, because these images contain a substantial number of novel objects that have not been seen during the training. The performance results

Table 3: Quantitative comparison with other methods on two zero-shot benchmarks. † denotes using **Parameter-Frozen module**, and * denotes including **Retrieval Augmentation module**. We report C and S scores on in-domain, near-domain, out-domain and overall data of NoCaps validation set; C and S scores on Flickr30K test set. Higher score is better. score indicates the best results among these methods, score indicates the second best results.

Method	**NoCaps val**								**Flickr30K**	
	In-domain		Near-domain		Out-domain		Overall		Test	
	C	S	C	S	C	S	C	S	C	S
$\text{Oscar}_{\text{Large}}$ [27]	79.9	12.4	68.2	11.8	45.1	9.4	65.2	11.4	–	–
ViECap† [13]	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5	47.9	13.6
ClipCap† [36]	84.8	12.1	66.8	10.9	49.1	9.6	65.8	10.9	–	–
SmallCap*† [42]	87.6	–	78.6	–	68.9	–	77.9	–	–	–
$\text{I-Tuning}^{\dagger}_{\text{Base}}$ [34]	83.9	12.4	70.3	11.7	48.1	9.5	67.8	11.4	61.5	16.9
Transformer [47, 7]	78.0	11.0	–	–	29.7	7.8	54.7	9.8	–	–
$\mathcal{M}^2$ Transformer [7]	85.7	12.1	–	–	38.9	8.9	64.5	11.1	–	–
Human [1]	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	–	–
$\text{Ours}^{\dagger}_{\text{w/o RA}}$	81.8	12.0	79.5	11.8	58.9	9.3	74.6	11.0	55.6	13.7
$\text{Ours*}^{\dagger}_{\text{w/ coco-datastore}}$	89.4	12.8	84.3	12.3	60.9	9.4	78.9	11.4	–	–
$\text{Ours*}^{\dagger}_{\text{w/ all-datastore}}$	93.9	13.0	87.1	12.5	66.4	10.3	81.5	11.6	63.4	15.2

on the NOCAPS validation set and Flickr30K test set are shown in table 3. It can be observed that M-RAT surpasses most of the baselines on Flickr30K and all the sub-domains of Nocaps, meanwhile approaches or even surpasses human-labeled captions, especially in In-domain (+9.5 CIDEr) and Near-domain (+2.1 CIDEr). These results emphasize M-RAT's advantages of zero-shot transfer in practical applications without fine-tuning or re-training. Additionally, M-RAT exhibits significant improvement (66.4 CIDEr) in out-domain evaluation compared with other encoder-decoder methods which lack Retrieval Augmentation module, such as Transformer (29.7 CIDEr) and $\mathcal{M}^2$ Transformer (38.9 CIDEr).

The bottom of table 3 is the comparison between M-RAT w/ and w/o RA module (81.5 vs 74.6 CIDEr). It can be observed that the RA module contributes greatly to adopting new knowledge. Moreover, the assistance of retrieved information is closely related to the content of the datastore, as proved by experiments conducted using the COCO datastore and all datastore, in which the former significantly improves performance in the in-domain and near-domain scenarios, and a relative big-scale datastore demonstrates more improvement across all sub-domains.

Qualitative Results and Visualization. In this section, we perform a detailed qualitative analysis of M-RAT. fig. 4 showcases how M-RAT effectively take advantage of the retrieved textual information to enhance both the diversity and accuracy of the generated output for the same image while eliminating the redundant noise, with the help of the aligned encoding and importance sample method. M-RAT is capable of automatically recognizing and aligning with relevant visual information (Bold parts in text-pair). Conversely, M-RAT can also accurately discern and choose to disregard the noise (e.g. "a helmet", "a man", "a black collar", etc.). More generated captions are shown in fig. 5 on the COCO, Flickr30K, and Self-built datasets (Virtual Game, Pedestrian Group), compared with two other SOTA methods (SMALLCAP is an outstanding RA based method and BLIP is a competitive VLP method). The generated captions of M-RAT possess superior accuracy with more detailed descriptions in a

Fig. 4: Generated results of M-RAT under different retrieval captions. □ corresponds to the retrieval captions; △ corresponds to the generated captions for input image. Dashed lines are used to separate different retrieval-generated caption pairs, and **Bold** parts indicate the relevance between them.

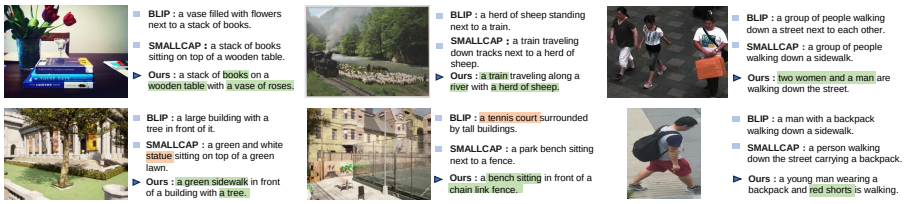


Fig. 5: Captioning result of our model and two chosen baselines on COCO, Flickr30K, and Self-built datasets. Incorrect objects in captions are highlighted in orange, while correct ones are in green.

concise sentence. As for Virtual Game pictures, both SMALLCAP and BLIP have experienced some misunderstandings (describing "building" as "statue", describing "basketball court" as "tennis court"), showing that virtual pictures pose challenges for captioning. For SMALLCAP, directly using the retrieval text as prompt for the decoder is susceptible to noise interference and fails to comprehensively incorporate visual information. M-RAT exhibits greater flexibility towards specific details (such as "chain link fence") while maintaining accuracy, which is greatly attributed to the advantage of our multi-grained mining and multi-level interaction design. For Pedestrian Group dataset, we can conclude that BLIP and SMALLCAP still lack attention to some details, while M-RAT not only summarizes these fine-grained attributes of pedestrians, but also are more conducive to the clothing information ("red shorts") and group clustering relationship ("two women and a man" instead of "a group of people").

Inference Costs and Efficiency. We measured the average inference time on an NVIDIA A100 GPU across 100 randomly sampled images. The resulting values of M-RAT, SMALLCAP and BLIP are 0.23, 0.26 and 0.48 seconds per image, respectively(including retrieval time). Due to the size of total parameters, M-RAT exhibits faster inference speed. Regarding the computational costs and efficiency when scaling datastore, We have included table 4 to further illustrate that M-RAT does not rely on high-quality large-scale datastore as a forced dependency for most general applications. When the domain gap is small(in-domain), increasing the size does not benefit a lot. However, when applied to zero-shot tasks(out-domain), doubling COCO with other datasets leads to significant improvements, but still maintains brief inference time and low GPU memory.

4.3 Ablation Study

To validate the effectiveness of M-RAT, we perform ablation studies on different components to explore the respective contribution to the overall performance,

Table 4: Computational costs and Nocaps CIDEr of M-RAT with different datastore sizes.

Datastore	Mem	Time(s)		Nocaps CIDEr	
		retr.	infe.	in.	out.
None	–	–	0.190	81.8	58.9
COCO	1.58 G	0.032	0.196	89.4	60.9
ALL	3.06 G	0.035	0.204	93.9	66.4

Table 5: Ablation study: performances of different module types on COCO test set. We report the Module and Type, with BLEU@4 (B@4), METEOR (M), ROUGE(R), CIDEr (C).

Module	Type	Metrics			
		B@4	M	R	C
Retrieval Aug.	w/o RA	39.0	28.8	58.4	127.9
	w/o Sample	38.6	29.3	58.7	132.0
Image Feat.	Regional	39.1	29.3	58.3	134.6
	Global	39.2	29.4	58.9	134.9
	Local+Global	39.9	29.7	59.4	135.5
Encoder-Decoder	$\text{Dec only}_{3\text{layers}}$	38.4	28.6	58.3	129.5
	$\text{Dec only}_{6\text{layers}}$	38.4	28.7	58.1	129.8
	w/o Multi-MSA	39.2	29.5	59.0	134.1
	w/o Low-level rep.	39.4	29.6	59.0	133.6
M-RAT	–	**39.9**	**29.7**	**59.4**	**136.1**

as shown in table 5. It can be observed that RA module plays an important role in M-RAT. Besides, Importance Sample is able to effectively improve the fused embedding. For image feature backbones, the global image feature derived from ViT is slightly better than the regional features extracted by pre-trained Faster R-CNN. In addition, applying both benefits a lot in captioning, which highlights the combined effect of local and global features in providing more coherent visual information. Furthermore, only regarding retrieval captions as a prompt input for decoder can still provide a certain auxiliary effect. However, the upper limit of decoder-only M-RAT is not high, even by increasing the number of decoder layers. Therefore, we can conclude that the effect of retrieved information can be amplified by multi-modal aligned encoder, compared with methods based on RA module (solely rely on a single-modal image encoder and utilize the retrieved caption as decoder input).

Fig. 6: Top-3 retrieval captions by applying distinct datastores(COCO, Flickr30K) for the same image. the useful information that matches the image is in green, while the interference noise is in orange. We also show the groundtruth, presented in **bold**.

Table 6: Evaluation on COCO test set when trained M-RAT performing captioning with specific retrieval texts we provide, including blank caption, random caption, retrieval caption and human-labeled caption.

Text content	BLEU-4	METEOR	ROUGE	CIDEr
Blank caption	39.0	28.7	58.4	128.0
Retrieval caption	39.9	29.7	59.4	135.5
Human-labeled caption	40.8	30.1	60.0	138.4

4.4 The Effect of Retrieval Augmentation Module

To further explore the effect of Retrieval Augmentation module on image captioning, we conduct a thorough analysis from the following perspectives, primarily aimed at guiding users on how to customize the content of datastore flexibly and serving as inspiration for the application of retrieval augmentation-based methods.

As shown in fig. 6, the retrieval results contain both useful information and redundant noise. By applying distinct datastores (COCO, Flickr30K) for the same image, we can observe that if the domain gap between datastore and input image is relatively narrow, then the proportion of useful information in the retrieved caption is likely to increase. We enforce M-RAT to generate captions with specific retrieval texts we provide. table 6 shows that using Human-labeled caption (with almost 100% useful information) as retrieval result yielded the best, outperforming Blank caption by 10.4 CIDEr. This observation proves: *The auxiliary impact of retrieval captions exhibits a positive correlation with the quantity of useful information, which can be directly promoted by the construction of datastore.* It can assist us in sensibly constructing datastore during inference to achieve optimal results. In experiments, we select several commonly used datasets from different research fields(i.e. VATEX[50] for video Captioning, Clotho[12] for Audio Captioning, TGIF[28] for GIF Captioning etc., as shown in table 1) and build a general datastore for most captioning scenarios.

5 Conclusion

Based on the advantages and disadvantages of existing research in image captioning, we propose a novel Multi-grained Retrieval Augmentation Transformer (M-RAT) for image captioning. Specifically, we introduce a Multi-Modal Aligned Encoder to effectively fuse extra retrieved information with the input visual feature, which greatly improves the accuracy of caption generation, and enhances the capacity of domain adaption. In addition, we realize an end-to-end structure of encoder-decoder with low-level and high-level meshed connection, which is proved to be productive by extensive experiments. In the future, we will further explore the structural design and embedding methods for retrieval augmentation method to better broaden its application.

Acknowledgments. This research was partly supported by grants of National Natural Science Foundation of China (NSFC, Grant No. 62171281), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 20DZ1200203, 2021SHZDZX0102).

References

1. Agrawal, H., Anderson, P., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S.: nocaps: novel object captioning at scale. In: ICCV. pp. 8947–8956. IEEE (2019)
2. Alberti, C., Ling, J., Collins, M., Reitter, D.: Fusion of detected objects in text for visual question answering. In: EMNLP/IJCNLP (1). pp. 2131–2140. Association for Computational Linguistics (2019)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: ECCV (5). Lecture Notes in Computer Science, vol. 9909, pp. 382–398. Springer (2016)
4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR. pp. 6077–6086. Computer Vision Foundation / IEEE Computer Society (2018)
5. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: IEEvaluation@ACL. pp. 65–72. Association for Computational Linguistics (2005)
6. Chen, F., Chen, X., Shi, J., Zhang, D., Chang, J., Tian, Q.: Hivlp: Hierarchical vision-language pre-training for fast image-text retrieval. CoRR **abs/2205.12105** (2022)
7. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: CVPR. pp. 10575–10584. Computer Vision Foundation / IEEE (2020)
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1). pp. 4171–4186. Association for Computational Linguistics (2019)
9. Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H., Chen, J., Liu, Y., Tang, J., Li, J., Sun, M.: Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat. Mac. Intell. **5**(3), 220–235 (2023)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)
11. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. CoRR **abs/2401.08281** (2024)
12. Drossos, K., Lipping, S., Virtanen, T.: Clotho: an audio captioning dataset. In: ICASSP. pp. 736–740. IEEE (2020)
13. Fei, J., Wang, T., Zhang, J., He, Z., Wang, C., Zheng, F.: Transferable decoding with visual entities for zero-shot image captioning. In: ICCV. pp. 3113–3123. IEEE (2023)
14. Girshick, R.B.: Fast R-CNN. In: ICCV. pp. 1440–1448. IEEE Computer Society (2015)
15. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. pp. 580–587. IEEE Computer Society (2014)
16. Gu, J., Wang, Y., Cho, K., Li, V.O.K.: Search engine guided neural machine translation. In: AAAI. pp. 5133–5140. AAAI Press (2018)

17. Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L.: Scaling up vision-language pre-training for image captioning. CoRR **abs/2111.12233** (2021)
18. Huang, L., Wang, W., Chen, J., Wei, X.: Attention on attention for image captioning. In: ICCV. pp. 4633–4642. IEEE (2019)
19. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR **abs/1508.01991** (2015)
20. Jiang, W., Ma, L., Jiang, Y., Liu, W., Zhang, T.: Recurrent fusion network for image captioning. In: ECCV (2). Lecture Notes in Computer Science, vol. 11206, pp. 510–526. Springer (2018)
21. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137. IEEE Computer Society (2015)
22. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137. IEEE Computer Society (2015)
23. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 5583–5594. PMLR (2021)
24. Li, J., Vo, D.M., Sugimoto, A., Nakayama, H.: Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. CoRR **abs/2311.15879** (2023)
25. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (2022)
26. Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., Hoi, S.C.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS. pp. 9694–9705 (2021)
27. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (30). Lecture Notes in Computer Science, vol. 12375, pp. 121–137. Springer (2020)
28. Li, Y., Song, Y., Cao, L., Tetreault, J.R., Goldberg, L., Jaimes, A., Luo, J.: TGIF: A new dataset and benchmark on animated GIF description. In: CVPR. pp. 4641–4650. IEEE Computer Society (2016)
29. Li, Z., Liu, D., Wang, H., Zhang, C., Cai, W.: Exploring annotation-free image captioning with retrieval-augmented pseudo sentence generation. CoRR **abs/2307.14750** (2023)
30. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004)
31. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (5). Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014)
32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 9992–10002. IEEE (2021)
33. Lou, L., Lu, K., Xue, J.: Improved transformer with parallel encoders for image captioning. In: ICPR. pp. 4072–4075. IEEE (2022)
34. Luo, Z., Hu, Z., Xi, Y., Zhang, R., Ma, J.: I-tuning: Tuning frozen language models with image for lightweight image captioning. In: ICASSP. pp. 1–5. IEEE (2023)
35. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: ICMR. pp. 19–27. ACM (2018)

36. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: CLIP prefix for image captioning. CoRR **abs/2111.09734** (2021)
37. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: CVPR. pp. 10968–10977. Computer Vision Foundation / IEEE (2020)
38. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318. ACL (2002)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
40. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
41. Ramos, R., Elliott, D., Martins, B.: Retrieval-augmented image captioning. In: EACL. pp. 3648–3663. Association for Computational Linguistics (2023)
42. Ramos, R., Martins, B., Elliott, D., Kementchedjhieva, Y.: Smallcap: Lightweight image captioning prompted with retrieval augmentation. In: CVPR. pp. 2840–2849. IEEE (2023)
43. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
44. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR. pp. 1179–1195. IEEE Computer Society (2017)
45. Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Retrieval-augmented transformer for image captioning. In: CBMI. pp. 1–7. ACM (2022)
46. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 539–559 (2023)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
48. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575. IEEE Computer Society (2015)
49. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. pp. 3156–3164. IEEE Computer Society (2015)
50. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: ICCV. pp. 4580–4590. IEEE (2019)
51. Wang, Y., Xu, J., Sun, Y.: End-to-end transformer based model for image captioning. In: AAAI. pp. 2585–2594. AAAI Press (2022)
52. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. In: ICLR. OpenReview.net (2022)
53. Weston, J., Dinan, E., Miller, A.H.: Retrieve and refine: Improved sequence generation models for dialogue. In: SCAI@EMNLP. pp. 87–92. Association for Computational Linguistics (2018)
54. Yang, Z., Ping, W., Liu, Z., Korthikanti, V., Nie, W., Huang, D., Fan, L., Yu, Z., Lan, S., Li, B., Shoeybi, M., Liu, M., Zhu, Y., Catanzaro, B., Xiao, C., Anandkumar, A.: Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. In: EMNLP (Findings). pp. 11844–11857. Association for Computational Linguistics (2023)

55. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: ECCV (14). Lecture Notes in Computer Science, vol. 11218, pp. 711–727. Springer (2018)
56. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics **2**, 67–78 (2014)
57. Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R.: Rstnet: Captioning with adaptive attention on visual and non-visual words. In: CVPR. pp. 15465–15474. Computer Vision Foundation / IEEE (2021)