Cross-modal Clustering-based Retrieval for Scalable and Robust Image Captioning

Jingyi You, Hiroshi Sasaki, Kazuma Kadowaki
The Japan Research Institute
{you.jingyi,sasaki.hiroshi,kadowaki.kazuma}@jri.co.jp

Abstract

Recent advances in retrieval-augmented generative image captioning (RAG-IC) have significantly improved caption quality by incorporating external knowledge and similar examples into language model-driven caption generators. However, these methods still encounter challenges when applied to real-world scenarios. First, many existing approaches rely on bimodal retrieval datastores that require large amounts of labeled data and substantial manual effort to construct, making them costly and time-consuming. Moreover, they simply retrieve the nearest samples to the input query from datastores, which leads to high redundancy in the retrieved content and subsequently degrades the quality of the generated captions.

In this paper, we introduce a novel RAG-IC approach named ***Cro**ss-modal **Di**versity-promoting **Ret**rieval technique* (CODIRET), which integrates a text-only unimodal retrieval module with our unique cluster-based retrieval mechanism. This proposal simultaneously enhances the scalability of the datastore, promotes diversity in retrieved content, and improves robustness against out-of-domain inputs, which eventually facilitates real-world applications. Experimental results demonstrate that our method, despite being exclusively trained on the COCO benchmark dataset, achieves competitive performance on the in-domain benchmark and generalizes robustly across different domains without additional training.

1 Introduction

Retrieval-augmented generative image captioning (RAG-IC) combines information retrieval with language model-based caption generation (Mallen et al., 2023; Cornia et al., 2020; Zhou et al., 2020; Shi et al., 2021) to leverage external knowledge or contextually relevant information to the input image and produce more accurate and informative image descriptions. This technology mitigates overdependence on the internal knowledge encoded in language models and instead incorporates external real-world data, thereby enhancing the semantic alignment between the generated captions and the visual content of the input images.

Figure 1: An example from MS COCO (Lin et al., 2014) of retrieved content containing redundant and semantically irrelevant terms with respect to the query image. We highlight the topic-deviant words in different colors from the correct keywords for clarity of presentation.

Although remarkable successes have been achieved in image captioning with the aid of retrieval techniques, several issues still hinder its application in real world scenarios. First, many existing RAG-IC approaches primarily perform unimodal retrieval (Sarto et al., 2022; Radford et al., 2021; Zhou and Long, 2023; Wu et al., 2024), where image-text pairs are selected based on the visual similarity between the retrieved and input images to augment contextual information. However, constructing such retrieval datastores requires a finely annotated corpus of image-text pairs, which is costly and labor-intensive, thereby limiting the scalability and adaptability of these methods in practical applications.

Secondly, traditional approaches typically rely

Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025), pages 47–58

on nearest-neighbor search to retrieve datastore contents based on the proximity of embedding representations extracted by pre-trained models (Khandelwal et al., 2021; Lewis et al., 2020). Therefore, as shown in Fig. 1, the retrieved texts tend to be highly repetitive and lack semantic diversity (Li et al., 2024b; Hoang et al., 2022), which in turn leads captioning models to overproduce these high-frequency words. In addition, such retrieval strategies are prone to retrieving irrelevant samples when the input falls outside the domain of the pre-trained model, which limits generalizability of the captioning system across domains.

To address the aforementioned limitations, we introduce a novel cross-modal retrieval approach that leverages a text-only datastore constructed without manual image-text annotations, thereby improving the scalability of the method. Furthermore, our proposed cluster-based retrieval strategy selects instances based on clustering in the embedding space, which not only improves the informativeness but also reduces semantic redundancy in the retrieved content. Specifically, we finetune the embedding function (encoder) by jointly incorporating a triplet contrastive loss and a nuclear norm regularization into the training objective to simultaneously reinforce alignment across modalities and capture the clustering structure of retrieved content in the embedding space (Nie et al., 2017; You et al., 2021).

We highlight our contributions as follows:

- We propose a novel RAG-IC framework that integrates cluster-wise selection with cross-modal retrieval. Our approach does not require an image-text paired datastore, thereby increasing the diversity of retrieved content and the robustness to out-of-domain inputs, which is critical for real-world applications.
- We introduce a specialized training paradigm that simultaneously addresses the gap between different modalities and encourages cluster formation among the embedding features of datastore samples by combining triplet contrastive loss and nuclear norm-based clustering regularization.
- Our analysis shows that CODIRET reduces retrieval redundancy and outperforms existing competitors in captioning quality, particularly in cross-domain inference settings, highlighting the effectiveness and robustness of our methodology.

2 Related Work

Robust retrieval-augmented generation. Retrieval-Augmented Generation (RAG) enhances text generation by incorporating externally retrieved knowledge as additional input (Lewis et al., 2020). Despite its success, particularly in natural language processing (NLP) (Mialon et al., 2023; Yasunaga et al., 2023), it has an overreliance on repetitive information in the retrieved content, which degrades the robustness to out-of-domain data and noisy inputs (Li et al., 2024b). To overcome the issue of practicality and generalizability in real-world applications, recent research focuses on strengthening RAG models to mitigate unstable retrievals and hallucination. One popular strategy is to dynamically adjust the training process in response to noisy retrievals (Zheng et al., 2021) with adversarial training (Fang et al., 2024) and relevance-aware evaluation of a given query (Yu et al., 2024) to facilitate the model to recognize and cope with various forms of retrievals. Another direction focuses on employing learnable filters or discriminators to effectively identify and eliminate redundant and misleading information (Zhu et al., 2024; Hong et al., 2024; Wu et al., 2024; Yoran et al., 2024). Additionally, methods such as random shuffling of retrieved content during training have been shown to boost the model's tolerance to domain mismatches and reduce overfitting to high-frequency patterns (Hoang et al., 2022; Li et al., 2024b; Hao et al., 2023).

Retrieval-augmented generative image captioning. Image captioning is the task of automatically generating descriptive textual captions for images (Herdade et al., 2019; Xu et al., 2015), combining techniques from computer vision and NLP. Recently, RAG-integrated image captioning has garnered increasing interest due to its prominent ability to improve accuracy, diversity, and factual consistency. Sarto et al. (2022); Ramos et al. (2023a); Sarto et al. (2024); Li et al. (2024a) propose to retrieve captions associated with visually similar images and develop encoder-decoder models that attend to both image features and retrieved caption embeddings. Rather than encoding images directly, Ramos et al. (2023b); Yang et al. (2023) enable "image-blind" decoding by utilizing only retrieved captions, allowing the model to focus on text-based reasoning without relying on direct visual understanding, which proves beneficial in zero-shot scenarios. Ramos et al. (2023c);

Algorithm 1 Traditional RAG-IC

Input: $\mathbf{I}$ // query image
k // number of samples to retrieve
$\mathcal{D} = \{(v_i, t_i)\}_{i=1}^{N}$ // external datastore containing image-captiong pairs
Output: $\mathbf{C} = \{c_i\}_{i=1}^{T}$ // output caption

/* extract features of the query image */
$v^q \leftarrow f_v(\mathbf{I})$
/* retrieve k image features from $\{v_i\}$ based on similarity to v^q
r_i: indices of the retrieved samples */
$\{v_{r_i}\}_{i=1}^{k} \leftarrow \text{Rtr}_k(v^q; \{v_i\})$
/* generate a caption for $\mathbf{I}$ using $\{t_{r_i}\}$ corresponding to $\{v_{r_i}\}$ */
$\mathbf{C} \leftarrow f_{\text{LLM}}(v^q, \{t_{r_i}\}_{i=1}^{k})$

Algorithm 2 Our cross-modal RAG-IC

Input: $\mathbf{I}$, k, $\mathcal{D} = \{t_i\}_{i=1}^{N}$
l // number of clusters
Output: $\mathbf{C}$

/* cluster $\{t_i\}$ by CoDiRet
c_i: indices of the clusters */
$\{g_{c_i}\}_{i=1}^{l} \leftarrow \text{Clu}_l(\{t_i\})$
/* extract features of the query image */
$v^q \leftarrow f_v(\mathbf{I})$
/* retrieve k cluster centroids from $\{g_{c_i}\}$ */
$\{g_{r_i}\}_{i=1}^{k} \leftarrow \text{Rtr}_k(v^q; \{g_{c_i}\})$
/* randomly select one text from each $\{g_{r_i}\}$ */
$\{t_{r_i}\} \leftarrow \text{RndSmp}_k(\{g_{r_i}\})$
/* generate a caption for $\mathbf{I}$ using $\{t_{r_i}\}$ */
$\mathbf{C} \leftarrow f_{\text{LLM}}(v^q, \{t_{r_i}\}_{i=1}^{k})$

Zeng et al. (2024) successfully implement retrieval over a unimodal textual datastore and adopt a lightweight architecture that integrates pre-trained CLIP (Radford et al., 2021) and GPT-2 (Radford et al., 2019) through retrieval-based prompting. We adopt SmallCap (Ramos et al., 2023c) as the baseline for training and evaluating our proposal due to its minimal trainable parameters for fine-tuning.

3 Methodology

Fig. 2 presents the overall architecture of our proposed CoDiRet framework, which is built upon two primary strategies: a **cross-modal alignment strategy** and a **cluster-based retrieval strategy**. Hereafter, we will present formal notations of variables and task definitions related to RAG-IC in Sec.3.1, and introduce each component subsequently in detail.

3.1 Preliminaries

Let $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ be an input image, where H, W, and C denote the height, width, and number of channels, respectively. As described in Alg.1, RAG-IC involves the following steps: 1) employing a *pre-trained visual encoder*, f_v, such as ViT (Dosovitskiy et al., 2021) or CLIP (Radford et al., 2021), to extract patch representations $\mathbf{X}$ from $\mathbf{I}$; 2) leveraging *retriever* Rtr_k to collect k semantically relevant instances R from an external database $\mathcal{D}$ by conducting feature-based nearest neighbor search between the query image and $\mathcal{D}$; and 3) utilizing a pre-trained large language model (LLM) as a *decoder* to generate a caption sequence $\mathbf{C}$ autoregressively by integrating the extracted visual embedding of $\mathbf{I}$ along with the retrieved textual knowledge R.

3.2 Cross-modal aligner

Given an image $\mathbf{I}$ organized in a 2-dimensional format as input, traditional RAG-IC approaches, as shown in Alg.1, rely on an external datastore consisting of image-caption pairs $\{(v_i, t_i)\}_{i=1}^{N}$ to retrieve similar images. In contrast, we exclusively construct the datastore from textual information in the target modality, denoted as $\mathcal{D} = \{t_i\}_{i=1}^{N}$, and retrieve captions based on the distance between features of the query image and the datastore captions by leveraging a shared multimodal representative space (Alg.2). This design facilitates efficient domain adaptation and scalability, as the datastore can be easily modified by replacing the textual corpus with off-the-shelf domain-specific data without requiring large-scale manually annotated datasets.

Triplet contrastive learning Although CLIP (Radford et al., 2021) aligns image and text representations in a shared multimodal embedding space by training a vision-language model in a contrastive learning manner, Mistretta et al. (2025) reveal a remaining modality gap between image and text representations, which causes inaccurate retrieval in the text modality. Specifically, even when captions describe different images, some sentences tend to cluster together in the 2-dimensional projection space, while, conversely, an image and its corresponding caption may be mapped far apart.

To solve this problem, we propose a triplet-based cross-modal alignment constraint aimed at mini-

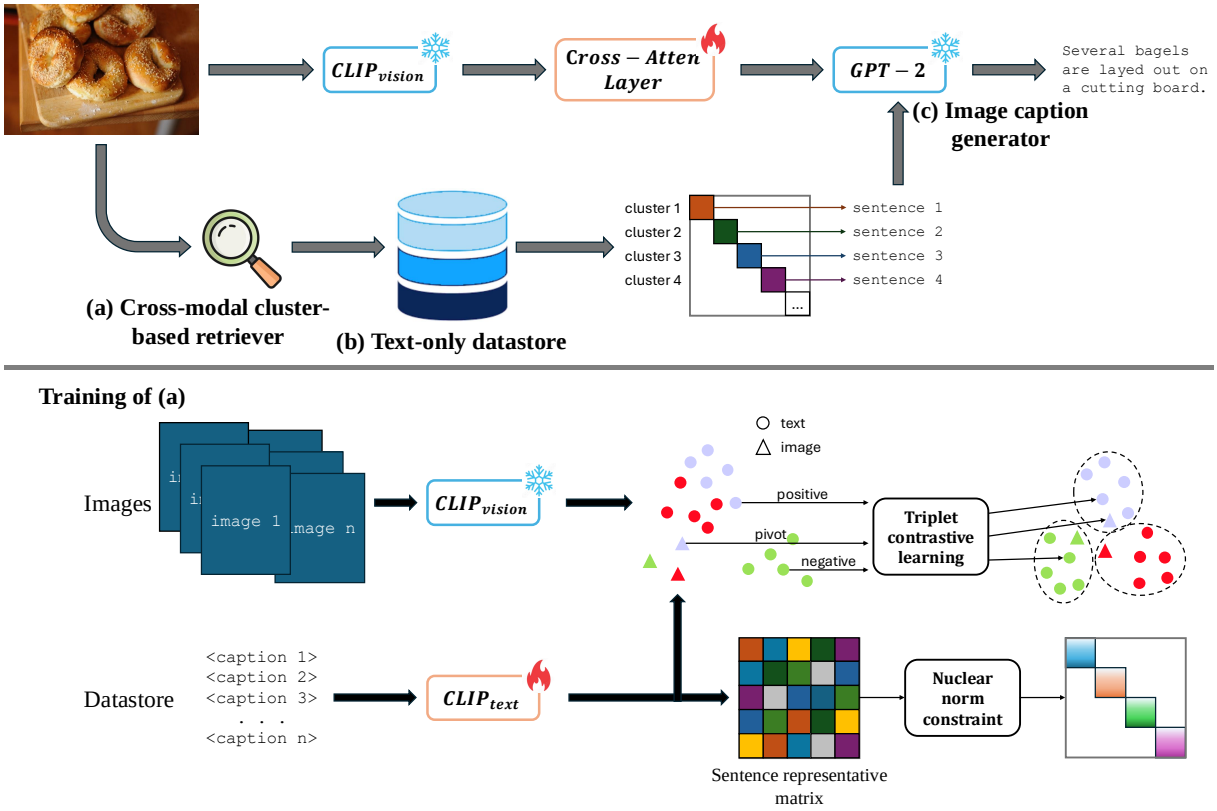

Figure 2: Model overview. CoDiRet comprises three chief components: (a) a cross-modal cluster-based retriever, (b) a text-only datastore, and (c) an image caption generator. Component (a) is trained using contrastive learning and nuclear norm regularization to mitigate misalignment between images and texts, while also clustering texts within the datastore. Subsequently, we utilize (a) to directly retrieve relevant text clusters from (b) based on the input image and randomly select one text from each cluster as supplementary input for (c).

mizing the modality gap and ensuring semantically relevant retrieval in a shared latent space, which is achieved by leveraging contrastive learning with a triplet loss formulation. Formally, for each particular image-caption example, the image serves as the pivot data point i^*, while one of its associated captions is randomly sampled as the positive example c^+. A caption from a different image is then randomly chosen as the negative example c^-. Subsequently, both the image and text are encoded into a shared embedding space using the CLIP model, as described below:

$$\mathbf{e}^* = f_{\text{clip}}^{\text{vision}}(i^*) \in \mathbb{R}^d, \tag{1}$$

$$\mathbf{e}^+ = f_{\text{clip}}^{\text{text}}(c^+) \in \mathbb{R}^d, \quad \mathbf{e}^- = f_{\text{clip}}^{\text{text}}(c^-) \in \mathbb{R}^d, \tag{2}$$

where d refers to the dimension of the CLIP embedding space.

We then conduct triplet noise-contrastive estimation (Gutmann and Hyvärinen, 2010) with a ranking loss to minimize the l_2 distance between the pivot and positive examples, while maximizing the distance between the pivot and negative ones:

$$\mathcal{L}_{\text{triplet}} = \max(0, \|\mathbf{e}^* - \mathbf{e}^+\|_2 - \|\mathbf{e}^* - \mathbf{e}^-\|_2). \tag{3}$$

By optimizing this objective, the model learns to group semantically similar image-text pairs while pushing apart unrelated ones, thereby ensuring that the retrieved text better matches the input image.

3.3 Cluster-based retriever

A simple ranking and selection of the top-k nearest neighbors based on similarity scores has long been dominant in the RAG field. However, this method often overlooks the underlying structure of the datastore, leading to captioning models receiving highly resembled information and repeated terms. As a result, the model is prone to copying these redundant words, regardless of their relevance (Hoang et al., 2022; Li et al., 2024b) and is then easily contaminated by noise. To equip models with diverse and informative supplementary data, we propose a cluster-based retriever that chooses texts from the nearest "clusters" detected

through a clustering operation beforehand, as illustrated in Alg.2. This approach reduces the occurrence of repeated words in the retrieved content and increases the possibility of including relevant words when processing out-of-domain images, by selecting from distant clusters.

Nuclear norm regularization Detecting clustering structures in $\mathcal{D}$ by directly applying K-means to the sentence representation matrix, H_t, can effectively reduce redundancy when retrieving captions. However, K-means is sensitive to initialization and outliers, which often leads to unstable results (Ding and Li, 2007). Additionally, the clustering performance is suboptimal due to the independent nature of triplet contrastive representation learning (Eq. (3)) and sentence clustering. Nie et al. (2017) pave the way for better capture of the clustering structure of H_t by transforming the clustering task into a matrix-rank problem. The theoretical basis behind the clustering structure learning comes from the following theorem:

Theorem 1 (Chung and Graham, 1997) The multiplicity of eigenvalue 0 of the normalized Laplacian matrix of H_t is equal to the number of clusters in H_t.

Haeffele and Vidal (2020); Piao et al. (2019) propose the nuclear norm and prove that the constraint on the Laplace matrix of H_t is mathematically equal to the constraint on sentence representation matrix H_t as

$$\mathcal{L}_{cluster} = \sum_{i=1}^{l} \lambda_i^{H_t}, \tag{4}$$

where $\lambda_i^{H_t}$ represents the i-th smallest eigenvalue of H_t (Piao et al., 2019). By suppressing $\|H_r\|_*^l$ to 0, l clusters (determined by elbow method (Bholowalia and Kumar, 2014)) in H_r can be obtained by reorganizing its columns or rows and converting it into a block-diagonal form with l blocks, as shown in Fig. 2. To incorporate this clustering into our training process, we define the training objective as:

$$\mathcal{L}_{cluster} = \|H_r\|_*^l. \tag{5}$$

3.4 Joint learning

We adopt a joint learning framework that optimizes both cross-modal alignment and modality-specific structure preservation. The overall objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{triplet}} + \lambda \mathcal{L}_{\text{cluster}}, \tag{6}$$

DATASET	Train	Validation	Test	
	MS COCO	MS COCO	MS COCO	NoCaps
IMAGES	113,287	5,000	5,000	4,500
CAPTIONS	566,747	25,010	25,010	45,000
Avg. Caps.	5	5	5	10
DOMAIN	in			out

Table 1: Basic dataset statistics. Avg. Caps. refers to average captions for each image.

where λ is a balancing coefficient that regulates the trade-off between enforcing cross-modal alignment and maintaining intra-modality cluster structures.

With the cluster structure of the text representation, we compute the centroid of each cluster by simply averaging the representations of sentences within the cluster. We then retrieve the top-k most relevant centroids and randomly sample one sentence from each retrieved cluster for the training of our captioning model, as shown in Alg. 2. Random sampling is adopted here to promote diversity and prevent the model from overfitting to highly prototypical or redundant sentences that may dominate each cluster.

4 Experiments

4.1 Datasets and Evaluation Metric

We carried out our experiments on the MS COCO Caption (Lin et al., 2014) and NoCaps (Agrawal et al., 2019) datasets to assess our approach's accuracy on in-domain data and its robustness to out-of-domain inputs, respectively. MS COCO Caption is a widely used benchmark that contains diverse image-caption pairs, while NoCaps focuses on novel object descriptions not present in the COCO training set, making it suitable for evaluating generalization to unseen concepts. The statistics of the datasets are summarized in Table 1.

For evaluation, we employ four standard automatic metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), which measure various aspects of caption quality, including n-gram overlap, semantic relevance, and compositionality.

4.2 Implementation Details

Our CODIRET retriever is first initialized using CLIP-ViT-B/32 (Radford et al., 2019) as both the image and text encoder and finetuned by triplet clustering learning outlined in Sec. 3.4 with LoRA

Metrics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	SPICE
SmallCap	76.5	60.2	44.3	31.7	23.1	74.4	13.4
CoDiRet	**77.8**	**61.2**	**45.6**	**32.9**	**24.2**	**76.3**	**14.2**
- w/o Tpl	75.4	58.9	43.0	31.1	22.7	72.4	12.7

Table 2: Robustness evaluation on the test set of NoCaps while the models are still trained on MS COCO. Best results among the generated captions are marked in bold.

(Low-Rank Adaptation) (Hu et al., 2022) to reduce computational cost and improve training efficiency. The scaling factor in Eq. (6) is set to $\lambda = 0.2$, as we found it yields the best empirical performance. As for the main image captioning model, we follow the SmallCap (Ramos et al., 2023c) setup, using CLIP-ViT-B/32 as the encoder and GPT-2 (Radford et al., 2019) as the decoder, with the parameters of both fixed, connected by a 12-head trainable cross-attention layer between the vision and language modalities to facilitate information fusion. Both the retrieval model and the main captioning model are trained exclusively on the MS COCO dataset using the standard Karpathy splits (Karpathy and Fei-Fei, 2015). The training procedures follow a batch size of 64, optimized with AdamW (Loshchilov and Hutter, 2019) and a learning rate of $1e-4$, using mixed-precision training with 16-bit floating-point precision (FP16). The training process runs for 5 epochs on CoDiRet and another 10 epochs on the captioning model on a single NVIDIA A100 GPU with 16GB of the available memory, taking approximately 13 hours to converge. During training, we retrieve $k = 4$ textual prompts per image by first identifying the top-k most similar clusters to the query image. The centroids of clusters and the query image embeddings are computed in the high-dimensional space by our CoDiRet. A single sentence is randomly sampled from each cluster and incorporated as a prompt for training. We employ the product quantizer with an inverted file system based on Faiss (Johnson et al., 2021) for efficient datastore quantization and nearest-neighbor search. Captions are decoded by beam search with a beam size of 3 at inference.

4.3 Baselines

The following excellent baselines are used for comparison to demonstrate the effectiveness of CoDiRet: **non-RAG** lightweight training method, including ClipCap (Mokady et al., 2021); **Img.→Img.** retrieval methods using image-text datastores such as EXTRA (Ramos et al., 2023a) and Re-ViLM (Yang et al., 2023); and **Img.→Txt.** retrieval method using text-only datastores like SmallCap (Ramos et al., 2023c). All methods are finetuned on the same training dataset as our method for a fair comparison.

5 Results and Discussion

5.1 Out-of-domain Robustness

To assess the robustness of our model under domain shift, we evaluate both CoDiRet and our baseline on NoCaps where the test set contains out-of-domain objects not present in the training distribution. From Table 2, we can observe that our model consistently outperforms the baseline SmallCap in terms of all metrics. The superior performance of our model in out-of-domain settings can be attributed to its ability to navigate the retrieval uncertainty and adapt to novel objects, which is a key limitation in conventional RAG-IC approaches.

The model's strong generalization ability indicates that it is less prone to overfitting. During training, it is provided with retrieval information from a broader range, which likely includes a small amount of noise. This prevents the model from simply copying or memorizing the content retrieved. Instead, it learns to flexibly apply the retrieved textual information in conjunction with the input visual data to generate accurate and fluent captions. In contrast, traditional kNN-based retrieval methods return captions associated with images that are similar to the query image, often resulting in a large amount of redundant information and repeated words. This redundancy causes the model to overfit specific patterns in the training data, thereby reducing its generalization ability on new data.

5.2 In-domain Performance

Table 3 lists the results for the non-RAG method at the top, with the ones with uni-modal retrieval in the middle, and cross-modal retrieval methods at the bottom. We can observe that when tested

Metrics	$\|\theta\|$	BLEU-4	METEOR	CIDEr	SPICE
ClipCap	43	33.5	27.5	113.1	21.1
EXTRA	45	37.5	28.5	120.9	21.7
Re-ViLM	158	37.8	-	129.1	-
SmallCap	7	37.0	27.9	119.7	21.3
CoDiRet	7.1	36.9	27.9	119.5	21.0
- w/o Tpl	7.1	34.9	26.8	117.6	20.5

Table 3: Results on the Karpathy COCO test split. $|\theta|$ refers to the number of trainable parameters in the model (in millions).

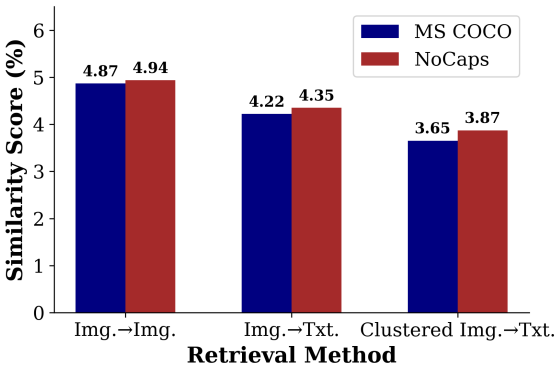

Figure 3: Comparison of proportion of duplicated key objects of image-to-image retrieval method, nearest neighbor-based image-to-text retrieval strategy, and our cluster structure-based CoDiRet.

with in-domain data, CoDiRet achieves a comparable performance to all state-of-the-art baselines in terms of all metrics, even with a small number of trainable parameters.

We also observed a consistent superiority of RAG-based models over ClipCap, which underscores the importance of external knowledge retrieval in image captioning. Without access to external descriptions, ClipCap is restricted to visual and linguistic knowledge already embedded in the pretrained LLM and often generates captions based on visual priors rather than factual correctness, leading to plausible but inaccurate descriptions.

In addition, we noticed that with the retrieval performed directly in the image modality, both EXTRA and Re-ViLM achieved better performance. We consider several possible reasons for this phenomenon. First, using the captions from the most visually similar images to the query image makes the models highly effective at preserving key visual details. Furthermore, these methods tend to achieve higher retrieval accuracy but greater keyword redundancy (as we show in Sec. 5.4) in the retrieved captions. This, in turn, allows the model to copy frequently repeated phrases from the retrieved text (as shown in Fig. 4), reinforcing consistency in generated captions.

In contrast, CoDiRet retrieves captions by directly searching for the most textually similar ones with a structured control over redundancy. By selecting a single representative caption per related text group, our approach promotes diversity in retrieved contents. However, since the entities in the images are limited, this diversity may introduce noise, which can lead to the model being slightly misled. Moreover, since the evaluation metrics used in this experiment, such as BLEU, cannot assess diversity, our model shows a minor performance decrease on in-domain data.

5.3 Ablation Study

We further investigate the contribution of the triplet contrastive learning module to CoDiRet through an ablation study conducted on each dataset. In the table, "- w/o Tpl" indicates the removal of the triplet contrastive learning module, where the retriever is trained solely with the nuclear norm constraint. We observe a significant performance drop of approximately 2 points on both in-domain and out-of-domain data compared with CoDiRet. This result suggests that triplet contrastive learning plays a crucial role in bridging the performance gap between different modalities, as it aligns image and text features more effectively.

Moreover, while the nuclear norm constraint primarily promotes representation compression and simplification by reducing the rank of the datastore sentence matrix, this process may inadvertently cause the model to overlook the intricate semantic differences between images and texts. As a consequence, important information encoded in the joint image-text space may be lost, weakening the quality of the retrieved captions and impairing the model's ability to generate accurate and contextually appropriate descriptions.

5.4 Duplicated Keywords and Redundancy

To better quantify the redundancy of retrieved contents in different retrieval strategies, we measure the lexical overlap of key objects across retrieved captions. Specifically, we select all nouns and proper nouns as candidate key objects from each caption using a spaCy-based part-of-speech tagger (Cutting et al., 1992). For each image, we then compute global object overlap among all retrieved

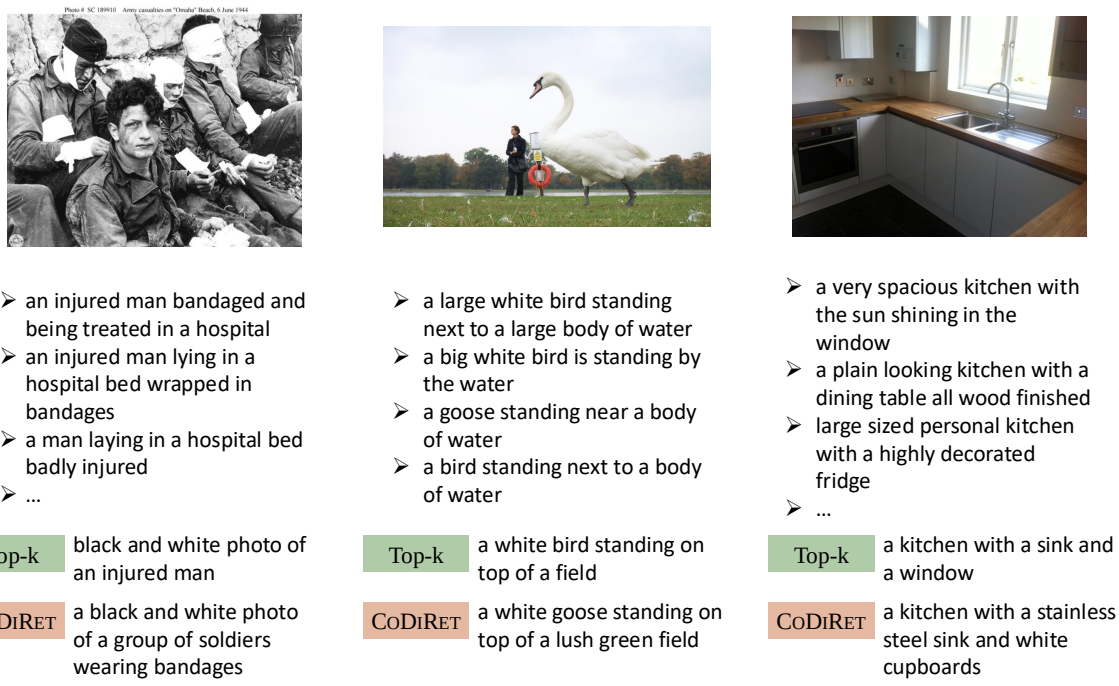

Figure 4: Examples of captions generated for NoCaps out-of-domain samples where the retrieved captions for the query image can be irrelevant.

captions by calculating the Jaccard similarity between the union and intersection of extracted object sets, defined as:

$$\text{Similarity}_{\text{object}} = \frac{|O_{\text{intersection}}|}{|O_{\text{union}}|}, \tag{7}$$

where $O_{\text{intersection}}$ is the set of objects appearing in all retrieved captions for a given image, and O_{union} is the set of all unique objects across the same set. A higher score indicates greater object repetition and thus higher lexical redundancy, while a lower score reflects increased content diversity. While simpler alternatives such as stopword removal could be used to filter non-content words, we adopt noun-based extraction to focus specifically on concrete entities that are most representative of the image content. This approach reduces the noise from abstract or generic terms that may still remain after stopword removal, and ensures that the resulting object sets more accurately reflect the semantic overlap of key visual concepts across captions.

We report the average object similarity score across all images in each retrieval setting on the two datasets separately as a proxy for topic-level redundancy in Fig. 3. The results demonstrate a clear trend in redundancy, where image-to-image retrieval exhibits the highest redundancy, while cluster-based image-to-text retrieval yields the most diverse references. We analyze the underlying reasons for these observations as follows. First, in image-to-image retrieval, since the retrieval is based purely on visual similarity, the captions tend to describe nearly identical content, often differing only in minor details or wording, which leads to a high degree of content repetition. Image-to-text retrieval bypasses the intermediate step of retrieving images and instead retrieves the most semantically similar captions directly from the text corpus, which offers greater flexibility by leveraging multi-modal embeddings to match text descriptions. However, our proposal introduces an additional step of clustering the text corpus before retrieval, ensuring that retrieved references come from different semantic groups. This enforces topic-level diversity among the retrieved references, as a result, preventing the model from receiving multiple variations of the same description.

5.5 Case Study

We demonstrate the quality of captions generated by CoDiRet through a case study. The examples shown in Fig. 4 are randomly sampled from the NoCaps dataset. We show captions retrieved from the datastore, along with comparison between captions generated by the traditional RAG-IC model and those produced by our approach.

A high-quality caption is typically characterized by (i) semantic relevance to the image content, (ii) specificity-inclusion of fine-grained details such as object attributes, actions, or materials, and (iii) fluency and coherence at the sentence level. Captions that satisfy these criteria are more informative and useful in downstream tasks such as image search or human-computer interaction.

From these examples, we observe that when certain words appear frequently in the retrieved content, models trained with standard nearest-neighbor-assisted information tend to copy those words verbatim into the generated caption. For instance, in the second example, the word "bird" is directly copied into the output caption. While such behavior may produce captions that are broadly accurate, they often lack specificity and fail to describe fine-grained visual attributes. In contrast, our model is better able to aggregate and distill informative content from the retrieved results, allowing it to produce more descriptive and contextually enriched details. For example, in the third image, the caption generated by CODIRET correctly includes the material "stainless steel" when describing the sink, offering a level of detail absent in the baseline output. Such specific terms are especially valuable for distinguishing similar scenes or objects, and thus contribute to a more effective and high-quality caption.

6 Conclusion

We addressed several fundamental problems concerning RAG-IC and proposed a joint learning framework called CODIRET, which trains a retriever by leveraging contrastive learning and clustering techniques to enhance cross-modal retrieval. The proposed model facilitates more semantically relevant retrieval results by minimizing the modality gap between image and text representations. Meanwhile, by incorporating a cluster constraint, the model effectively reduces redundancy in the retrieved content, ensuring better adaptation to out-of-domain scenarios. Experimental results, including those of the analysis of retrieved contents, demonstrated the effectiveness of CODIRET.

Limitations

In this study, we introduced diversity to enhance the model's robustness on unseen data, particularly by expanding the variety of retrieved content to avoid over-reliance on high-frequency samples. While this strategy significantly improved the model's performance on out-of-domain data, it led to a decline in performance on in-domain data. This phenomenon may be attributed to the increased diversity leading to the retrieval of content that is only partially relevant to the input image, thus affecting the accuracy and consistency of the model's outputs on known data. While the added diversity enhances the model's adaptability to unseen domains, it also causes a trade-off with its performance in specific domains. Therefore, balancing diversity with precision, ensuring strong robustness without compromising performance on in-domain data, remains a challenge that warrants further investigation. We consider this trade-off an important area for future work, aiming to explore how to achieve an optimal balance between the two.

Currently, most image captioning models rely on English-centric datasets such as COCO, which limits their effectiveness in multilingual and multicultural contexts. These models may struggle with linguistic and cultural differences, as well as diverse visual concepts. Future research should focus on multilingual image captioning datasets that include data from various languages and cultures, enabling models to perform better across different settings and promoting global application of image captioning technology.

References

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Purnima Bholowalia and Arvind Kumar. 2014. Ebk-

means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9):17–24.

Fan RK Chung and Fan Chung Graham. 1997. *Spectral graph theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Soc., Fresno State University.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. Computer Vision Foundation / IEEE.

Douglas R. Cutting, Julian Kupiec, Jan O. Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *3rd Applied Natural Language Processing Conference, ANLP 1992, Trento, Italy, March 31 - April 3, 1992*, pages 133–140. ACL.

Chris H. Q. Ding and Tao Li. 2007. Adaptive dimension reduction using discriminant analysis and K-means clustering. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 521–528. ACM.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10028–10039. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 297–304. JMLR.org.

Benjamin D. Haeffele and René Vidal. 2020. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(6):1468–1482.

Hongkun Hao, Guoping Huang, Lemao Liu, Zhirui Zhang, Shuming Shi, and Rui Wang. 2023. Rethinking translation memory augmented neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2589–2605. Association for Computational Linguistics.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11135–11145.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2022. Improving robustness of retrieval augmented translation via shuffling of suggestions. *CoRR*, abs/2210.05059.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2474–2495. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2024a. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13733–13742. IEEE.

Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. 2024b. Understanding retrieval robustness for retrieval-augmented image captioning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9285–9299. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *Trans. Mach. Learn. Res.*, 2023.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. 2025. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. *CoRR*, abs/2502.04263.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734.

Feiping Nie, Xiaoqian Wang, Cheng Deng, and Heng Huang. 2017. Learning A structured optimal bipartite graph for co-clustering. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4129–4138.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Xinglin Piao, Yongli Hu, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2019. Double nuclear norm based low rank representation on grassmann manifolds for clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12075–12084. Computer Vision Foundation / IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023a. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3648–3663. Association for Computational Linguistics.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023b. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1635–1651. Association for Computational Linguistics.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023c. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2840–2849. IEEE.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-augmented transformer for image captioning. In *CBMI 2022: International Conference on Content-based Multimedia Indexing, Graz, Austria, September 14 - 16, 2022*, pages 1–7. ACM.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. 2024. Towards

retrieval-augmented architectures for image captioning. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(8):242:1–242:22.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 269–277. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Hao Wu, Zhihang Zhong, and Xiao Sun. 2024. DIR: retrieval-augmented image captioning with comprehensive understanding. *CoRR*, abs/2412.01115.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Mohammad Shoeybi, Ming-Yu Liu, Yuke Zhu, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11844–11857. Association for Computational Linguistics.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jingyi You, Chenlong Hu, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2021. Robust dynamic clustering for temporal networks. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 2424–2433. ACM.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14672–14685. Association for Computational Linguistics.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024. Meacap: Memory-augmented zero-shot image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14100–14110. IEEE.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 368–374. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.

Yucheng Zhou and Guodong Long. 2023. Style-aware contrastive learning for multi-style image captioning. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2212–2222. Association for Computational Linguistics.

Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng Chu, Jingchang Chen, and Bing Qin. 2024. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1044–1069. Association for Computational Linguistics.