Exploring Annotation-free Image Captioning with Retrieval-augmented Pseudo Sentence Generation

Anonymous ACL submission

Abstract

Recently, training an image captioner without annotated image-sentence pairs has gained traction. Previous methods face limitations due to either using mismatched corpora for inaccurate pseudo pairs or relying on resource-intensive pre-training. To alleviate these challenges, we propose a new strategy where the prior knowledge from large pre-trained models (LPMs) is distilled and leveraged as supervision, and a retrieval process is integrated to further reinforce its effectiveness. Specifically, we introduce **R**etrieval-**a**ugmented **P**seudo **S**entence **G**eneration (RaPSG), which can efficiently retrieve highly relevant short region descriptions from mismatching corpora and use them to generate a variety of high-quality pseudo sentences via LPMs. Additionally, we introduce a fluency filter to eliminate low-quality pseudo sentences and a CLIP guidance objective to enhance contrastive information learning. Experimental results show that our method outperforms SOTA captioning models in zero-shot, unsupervised, semi-supervised, and cross-domain scenarios. Moreover, we observe that generating high-quality pseudo sentences may offer better supervision than the crawling sentence strategy, highlighting future research opportunities.

1 Introduction

Recent advancements in image captioning have been driven by Transformer-based models (Cornia et al., 2020; Luo et al., 2021). However, the reliance on high-quality human-annotated image-text pairs limits these fully-supervised approaches, increasing interest in annotation-free alternatives, such as unsupervised and pre-training strategies. Unsupervised approaches (Guo et al., 2020; Zhou et al., 2021) align crawled sentences with target images as pseudo annotations, but face issues with sentence diversity (Li et al., 2022) and content accuracy (Honda et al., 2021). Pre-training strategies show strong performance but require massive resources (Wang et al., 2021) and are affected by noisy data from coarse LPM-led selection (Byeon et al., 2022), as shown in Figure 1(a), leading to poor sample efficiency (Li et al., 2022).

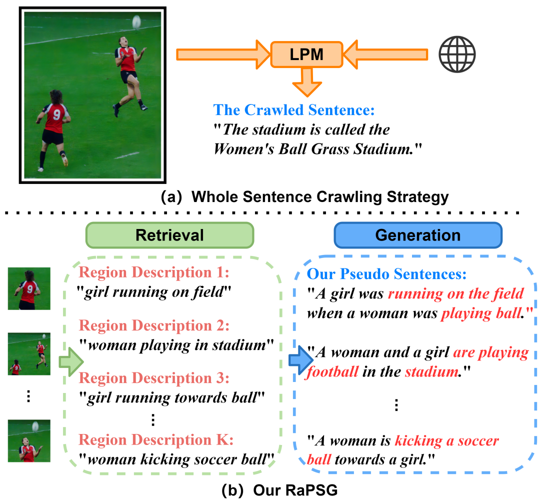

Figure 1: The comparison between whole sentence crawling strategy (Byeon et al., 2022) and our generation-based RaPSG method.

To alleviate these problems, recent methods transfer prior knowledge from frozen LPMs to vision-language (VL) tasks. Notable architectures like Flamingo (Alayrac et al., 2022) and BLIP2 (Li et al., 2023a) use trainable mapper modules to bridge LPMs with vision encoders, keeping LPMs frozen to reduce computational cost and avoid catastrophic forgetting. Similarly, LLaVA (Liu et al., 2023b) and MiniGPT4 (Zhu et al., 2023a) employ projection layers to integrate visual encoders with language decoders, innovating through fine-tuning on multimodal instructions. However, despite these advancements, all these methods still rely on billions of external image-text pairs for "mapper" learning and remain susceptible to the challenge of the noisy image-text pairs problem.

In this paper, we propose an efficient Retrieval-augmented Pseudo Sentence Generation framework (RaPSG) that leverages prior knowledge from frozen LPMs as supervision by generating high-quality pseudo sentences without the need of external image-text pairs or instruction tuning for optimization. Specifically, a retrieval-based pipeline is designed to generate multiple sentences for each target image, as shown in Figure 1(b). To address the challenge of noisy image-text pairs and improve the quality of generated pseudo sentences, we propose a refinement strategy based on a ranking of high-relevance region descriptions. For each target image, we employ the pre-trained model CLIP (Radford et al., 2021) to retrieve the top-k most correlated region descriptions from the Visual Genome (VG) dataset (Krishna et al., 2017) (We eliminate the overlapping parts between COCO and VG). Then, we further group region descriptions into multiple comprehensive and distinct long sentences using summarization LPMs, such as BART (Lewis et al., 2019) and LLaMA (Touvron et al., 2023). After this, we then introduce a self-supervised framework to facilitate the retrieval-augmented captioner, using original images and generated pseudo sentences as supervision. Additionally, we design two mechanisms to enhance the plain pseudo-labeling strategy. Firstly, a fluency filter removes imperfect descriptions to mitigate the impact of noisy image-text pairs. Second, a CLIP-based optimization strategy improves the model's comprehension of image-text pairs, offsetting the lack of external image-text pairs.

To demonstrate the capability of our RaPSG approach, we evaluate its performance on the MSCOCO (Chen et al., 2015) and Flickr30k (Plummer et al., 2015) benchmarks across various settings. The results show that our method outperforms the SOTA captioner Flamingo3B with fewer trainable parameters and consistently surpasses other models in pre-training, unsupervised, weakly supervised, and unpaired settings. This highlights its effectiveness and efficiency. Additionally, we validate its robustness in semi-supervised and cross-domain settings, where our model also achieves SOTA performance, underscoring its versatility.

Our contributions are summarized as four folds: (1) We propose an inference-only approach that distils knowledge from frozen LLMs by retrieving highly relevant region descriptions and generating a variety of distinct pseudo sentences for each target image. (2) A fluency filter and CLIP guidance are further introduced to strengthen the retrieval-augmented learning of the captioner for better prediction. (3) Experimental findings reveal that our approach surpasses current SOTA captioning models in a range of scenarios, including zero-shot, unsupervised, semi-supervised and cross-domain settings. (4) In our experiments, we also find that using high-quality generated pseudo sentences is more beneficial for captioner training than retrieved complete sentences, even if they are unpaired and sourced directly from the original dataset.

2 Related Work

Large Pre-trained Models for Image Captioning. In recent years, the appearance of a series of high-performance LPMs such as ViT (Dosovitskiy et al., 2020), GPT-2 (Radford et al., 2019), and CLIP (Radford et al., 2021) has widely extended the possibility of getting prior knowledge. Kuo and Kira (2022) used CLIP to mine missing attributes and relationships as auxiliary inputs in a fully supervised captioning task. Cho et al. (2022) used CLIP to build a CLIP score replacing the traditional cross-entropy loss, which can avoid references in strength learning of captioning tasks. Additionally, some works start to explore leveraging from the frozen LPMs. Flamingo (Alayrac et al., 2022) builds a trainable architecture that bridges the vision encoder and the large language model, efficiently accepting arbitrarily interleaved visual data and text as input, and generating text in an open-ended manner. BLIP-2 (Li et al., 2023a) bridges the modality gap with a lightweight querying Transformer and is more efficient in the pre-training strategy. However, all these methods still need pre-training on large-scale datasets for model optimization.

Retrieval-augmented Models with LPMs. Retrieval-augmented methods have been widely applied in VL tasks in recent years. In visual question answering, retrieving the outside knowledge for question answering has become the new trend (Lin and Byrne, 2022). In text-to-image generation, Chen et al. (2022b) propose a generative model that uses retrieved information to produce high-fidelity images for uncommon entities. Currently, few works apply a retrieval-augmented idea with LPMs for image captioning. Zhu et al. (2023b) use CLIP to extract the semantic prompt for more accurate caption prediction under the adversarial learning framework. Re-ViLM (Yang et al., 2023) builds upon the Flamingo but supports using CLIP to retrieve relevant knowledge from the external database. Compared with their methods, our approach gets knowledge from high-quality generated pseudo sentences and is

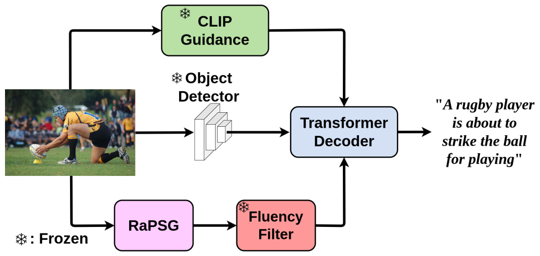

Figure 2: The overview of our proposed framework. It is structured around three core components: RaPSG, fluency filter, and CLIP guidance.

more data-efficient, which avoids using unpaired human annotation (Zhu et al., 2023b) or large-scale image-text corpus for pre-training (Yang et al., 2023).

3 Method

In this section, we introduce our proposed framework RaPSG, whose overview is shown in Figure 2. The retrieval-augmented pseudo sentence efficient generation module is proposed to learn knowledge from the LPMs (Section 3.1). To reduce the appearance of unnatural pseudo sentences, we innovatively design a fluency filter (Section 3.2). Finally, the self-supervised training with generated pseudo image-text pairs is guided by a CLIP-based loss to improve the prediction accuracy (Section 3.3).

3.1 Retrieval-Augmented PSG Module

To address the absence of human annotation, we propose RaPSG, a two-stage retrieval-augmented pseudo sentence generation method. It leverages the prior knowledge in LPMs to generate high-quality pseudo sentences for effective training supervision. Specifically, our method is based on the text processing capabilities from different aspects of LPMs including region-level matching with CLIP, global-level summarization through BART, and LLaMA for further enhancement. Stage-I transforms region-level information into global-level sentences to establish context, while Stage-II distills and refines these sentences with detailed content. This approach ensures high-quality pseudo sentences through comprehensive and robust text processing.

In Stage-I, we focus on utilizing the summarization capability of BART (Lewis et al., 2019) to condense short high-relevant region descriptions into pseudo sentences (Figure 3), capturing essential information from regions concisely. To begin, we retrieve local-level region descriptions from the Visual Genome (VG) dataset (a public dataset comprises region descriptions). However, since 47% of VG images overlap with MSCOCO, we apply a duplicate-removal scheme (Kuo and Kira, 2022) to refine region descriptions. After annotating region descriptions, we utilize the pre-trained CLIP to retrieve proper region descriptions for each image. Given an image I, we apply the cosine similarity function to calculate the matching score for each region description, then rank these descriptions according to their scores in descending order, forming the ordered set of region descriptions $\hat{D}$. Subsequently, the top-k most relevant descriptions are chosen based on their scores for the following steps, with the selection of k detailed in Figure 6. These selected top-k region descriptions for the given image are denoted as $\hat{D}^k$. However, as illustrated in Figure 1 (b), the region descriptions lack modifying phrases typically found in standard sentences. Previous research, such as Feng et al. (2019), indicates that concepts with minimal semantic content can lead to failures in image captioning training.

To cope with missing information, we refine local-level descriptions by summarizing them into global-level descriptions using BART. From the set $\hat{D}^k$, we select the top-m descriptions with the criteria for choosing m detailed in Figure 6. Then, these descriptions are summarized into the first single sentence, c_1, by removing repeated words and leveraging the text summarization ability of BART (comparisons across different summarization models also depicted in Figure 6). To enhance the diversity of pseudo sentences, instead of repeating the summarization process above, we group the remaining regions descriptions based on greater semantic differences. Specifically, a similarity score is calculated between each of the rest region descriptions $\hat{D}^{[k-m]}$ and the first pseudo sentence c_1. Next, these descriptions are grouped into n comprehensive summarization sentences based on scores (i.e., $n = \frac{k-m}{m}$, the top m for the c_2, the second top m for the c_3, and ...). In this way, descriptions sharing more similarities would be grouped together to avoid arranging too many objects in a single sentence generation process. The issue of grouping complex objects together will be discussed in Section 3.2. According to this setting, our method can generate a high-quality pseudo sentence group $\{c_i\}_{i=1}^{k/m}$ per image in the first stage.

In Stage-II, we distill crucial information from the preceding sentence group to generate more appropriate pseudo sentences. We refine the gener-

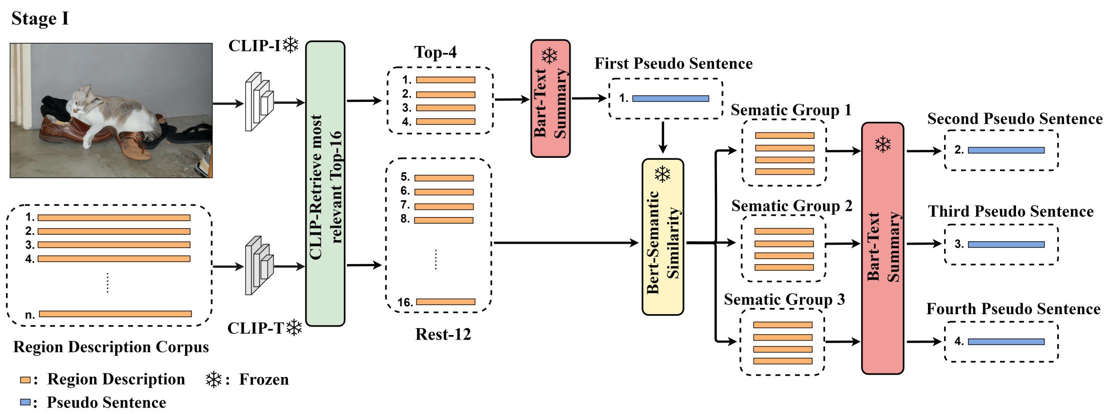

Figure 3: The Stage-I of RaPSG framework. Firstly, we retrieve top-k region descriptions from VG (Krishna et al., 2017) according to their matching scores computed by CLIP (Radford et al., 2021) model. Then, we use Sent-BERT (Reimers and Gurevych, 2019) model to divide them into four groups by their semantic similarity. Finally, BART (Lewis et al., 2019) model is used to summarize the grouped descriptions for four pseudo sentences.

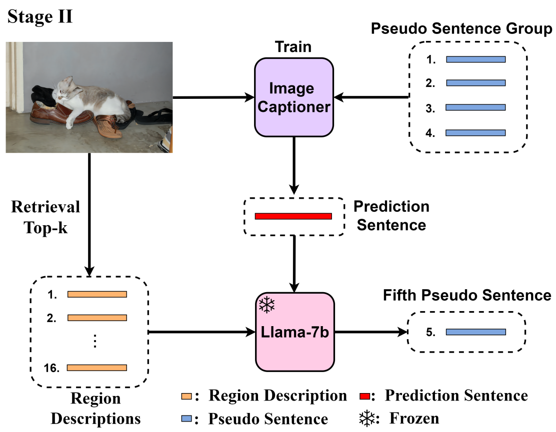

Figure 4: The Stage-II of RaPSG framework. Initially, we utilize the provided image in conjunction with the preceding four pseudo sentences as supervision to train the image captioner. Once trained, we freeze the captioner and generate a prediction sentence. To enhance the generation process, we incorporate the top-k most relevant region descriptions as supplementary material to get the fifth output.

ated pseudo sentences in Stage-I using the expressive power of large generative models, producing fluent and contextually relevant sentences for supervision. Initially, we pair the set $\{c_i\}_{i=1}^{k/m}$ with the I for captioner training, as shown in the top part of Figure 4. This process establishes a reconnection between the sentences and the visual content, enabling the captioner's accuracy in both image and text domains. However, the supervision by pseudo sentences could lead the captioner to learn repeated information, potentially resulting in a lack of specific details within the context.

To address this limitation, we propose incorporating a large-size generative model, LLaMA-7B, to generate pseudo sentences with more detailed information. In our approach, we refine the sentences by using the predictions from the frozen captioner as well as the $\hat{D}^k$. By combining these elements, LLaMA learns the core ideas from the predictions and incorporates the detailed information from the region descriptions. This integration enables us to generate superior pseudo sentences that encompass a greater level of detail. Consequently, we obtain a more appropriate sentence as our another output denoting as $c_{k/m+1}$. With these two stages completed, we successfully generate a group of pseudo sentences $\{c_i\}_{i=1}^{k/m+1}$ that are ready for further use.

3.2 Fluency Filter

The fluency filter is designed to sift the generated sentences to remove low-quality pseudo captions. For each given image I, the filter carefully selects the best sentence among $\{c_i\}_{i=1}^{k/m+1}$ to ensure a precise match. Figure 5 compares two generated pseudo sentences from BART based on two groups of region descriptions in the first stage of the RaPSG module. The first case shows that the model successfully comprehends the relationship between the skateboard and the trick in the inference process. By contrast, the second sentence does not capture the important information to describe the image because the model recognizes the metallic-element different from the skateboard. Due to the limited discernment of LPMs, varying appellations for the same object in region descriptions can cause confusion, potentially fragmenting the generated sentence into multiple semantic parts and reducing its coherence and accuracy.

We propose to filter out the low-quality pseudo sentences via CIDEr metric (Vedantam et al., 2015) (an image description evaluation based on human preference) because these low-quality pseudo sentences are also made up of highly relevant phrases but in an unnatural arrangement and can deceive

Figure 5: A comparison of two pseudo sentences in RaPSG process. The first sentence appears more fluent than the second sentence from the human view. Best viewed by zooming in.

the common evaluation methods. Since real annotations are unavailable, we use the model's predictions as references. To this end, we propose that the $\{c_i\}_{i=1}^{k/m+1}$ are examined by the CIDEr metric, and the one graded the highest is chosen as follows:

$$c_{cider} = \arg\max_{c} CIDEr(c_i, f_c(I)), \tag{1}$$

where c_i^j is the j-th pseudo sentence among five. $f_c(I_i)$ is the model prediction sentence and f_c is the basic captioning model.

3.3 CLIP Guidance

The CLIP guidance module is proposed to encourage the sentence prediction to semantically match image content in CLIP embedding space as we abandon pre-training on external large-scale datasets. The InfoNCE (Oord et al., 2018) is employed to reduce cross-modal information loss. The frozen image encoder CLIP-I and text encoder CLIP-T are used to embed a dozen original images and corresponding predictions into a shared semantic space. Then, the pairwise affinities are computed based on the encoded features. The learning process can be formulated as minimizing the contrastive information loss:

$$L_I = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)}, \tag{2}$$

where q is a visual embedding for an image extracted from the CLIP-I, k^+ is the text embedding for this image (positive key), and k^- are text embedding for other images from the same batch in the training process (negative key). Both of them are generated by CLIP-T. τ is the temperature hyper-parameter.

4 Experiments and Results

4.1 Experiments Setting

Datasets. We choose MSCOCO (Chen et al., 2015) and Flickr30k (Plummer et al., 2015) with Karpathy (Karpathy and Fei-Fei, 2015) split as our test benchmark. The MSCOCO images are divided into three parts: 113k images for training, 5k images for validation, and the remaining 5k images for testing. The Flickr30k images are divided into three parts: 29k images for training, 1k images for validation, and the remaining 1k images for testing.

Evaluation Metrics. Following standard captioning evaluation protocols (Li et al., 2019), we employ the following five metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). Beyond these traditional metrics, we also incorporate the innovative robust metric CLIP-S (Hessel et al., 2021), which assesses the relevance between the generated caption and the target image independently of reference captions.

Image Captioning Backbones. Our approach is versatile for different image captioning models. To validate its performance, we incorporate our proposed framework with several classic captioners, including: M^2 model (Cornia et al., 2020), CTX model (Kuo and Kira, 2022), DLCT model (Luo et al., 2021), and DIFNet model (Wu et al., 2022).

Comparison Setting. To the best of our knowledge, we are making an early attempt to explore a new image captioning benchmark setting that leverages retrieval-augmented self-supervised learning without annotated labels. There are two comparable settings that we can contrast with our approach: pre-trained models in zero-shot setting and finetuning-based approaches without full supervision. Unlike existing zero-shot methods, our approach uses self-supervised training with generated pseudo sentences, avoiding reliance on large external datasets. Additionally, we compare our method with unsupervised, unpaired, and weakly-supervised finetuning approaches, as both assume the absence of grounded image-text pairs and use pseudo pairs for optimization. Finally, to comprehensively assess the capability of our approach, we extend our test to semi-supervised and cross-domain settings, comparing our model's performance against SOTA models in these scenarios.

Model	Trainable Params	Pre-trained Models	External Dataset	MSCOCO B1	B4	M	C	S	CLIP-S	Flickr30k B1	B4	M	C	S	CLIP-S
SimVLM$_{base}$ (2021)	-	-	1.8B	-	9.5	11.5	24.0	7.5	-	-	-	-	-	-	-
SimVLM$_{huge}$ (2021)	1.4B			-	11.2	14.7	32.2	8.5	-	-	-	-	-	-	-
Re-ViLM$_{base}$ (2023)	158M	CLIP (2021)	762M	-	17.0	-	51.2	-	-	-	-	-	45.2	9.2	-
Re-ViLM$_{large}$ (2023)	806M			-	18.6	-	60.8	-	-	-	-	-	52.1	10.0	-
Flamingo3B (2022)	1.3B	NFNet (2021) and Chinchilla (2022)	312M	-	-	-	73.0	-	-	-	-	-	60.6	-	-
Flamingo80B (2022)	10B			-	-	-	**84.3**	-	-	-	-	-	**67.2**	-	-
MiniGPT4-V1 (2023a)	-	ViT (2020) and Vicuna (2023)	5M	23.6	5.8	20.9	0.0	14.4	34.0	13.2	3.5	15.6	0.0	14.8	32.3
MiniGPT4-V2 (2023)	-		20M	28.6	6.3	24.4	0.0	17.9	35.5	17.5	6.6	22.0	0.0	20.0	32.6
LLaVA1.0 (2023a)	0.14B	CLIP (2021) and Vicuna (2023)	0.59M	38.5	9.1	**26.7**	50.9	**24.2**	34.1	48.0	13.0	**23.4**	52.5	**17.1**	33.7
LLaVA1.5 (2023b)	0.70B		0.66M	30.6	10.1	24.8	41.8	22.6	31.5	35.2	7.7	21.9	34.1	17.0	30.5
Our Pseudo Sents.	0	CLIP (2021), BART (2019), and LLaMA (2023)	**0.45M**	48.1	8.8	18.0	39.3	13.3	**47.6**	43.2	14.5	17.1	21.2	9.3	**45.4**
Ours (w/ CTX)	40M			67.0	18.3	21.2	72.4	14.1	33.6	51.7	17.8	21.0	53.3	10.7	32.6
Ours (w/ M^2)	38M			67.5	18.9	20.9	75.3	14.7	34.3	54.6	17.5	20.7	56.8	11.2	33.8
Ours (w/ DLCT)	63M			69.5	**19.4**	21.1	75.9	14.5	34.5	54.1	18.1	22.6	58.4	11.5	34.1
Ours (w/ DIFNet)	**33M**			**70.5**	19.3	21.4	78.1	14.9	35.8	**55.9**	**18.2**	23.1	59.1	11.8	33.9

Table 1: The comparison of our approach with SOTA zero-shot models on MSCOCO and Flickr30k benchmarks. We denote different captions (i.e., CTX, M^2, DIFNet, and DLCT) inside the brackets. Pseudo Sents. represents the generated pseudo sentences from the RaPSG module. BLIP (Li et al., 2022) and BLIP2 (Li et al., 2023a) are excluded from our comparison due to their use of COCO captions during pre-training process.

4.2 Comparison against Large Pre-Trained Models

We compare our RaPSG approach with the zero-shot models (Wang et al., 2021; Yang et al., 2023; Alayrac et al., 2022; Zhu et al., 2023a; Liu et al., 2023b) on MSCOCO and Flickr30k benchmarks, as they are all built up on LPMs. Table 1 demonstrates that our method surpasses the performance of these models on the MSCOCO benchmark in some metrics (Note that multimodal LLMs like MiniGPT4 and LLaVA are not specifically trained to generate short captions, and their detailed descriptions may not be fully captured by traditional metrics; see Appendix A.1 for more details). Moreover, previous approaches rely on pre-training with a large number of external image-text pairs and demand a considerable number of trainable parameters. For instance, Flamingo3B is pre-trained on 312M external image-text pairs, whereas our model only requires 0.45M (0.14%) generated pseudo sentences, which is more data-efficient. Additionally, we also validate our approach on another popular benchmark Flickr30k. Table 1 shows our method's robustness across datasets, matching SOTA models in performance with fewer trainable parameters (e.g., 6.7% of Flamingo, 4% of Re-ViLM).

4.3 Comparsion against Finetuning-Based Approaches

Next, we compare ours with other models that operate without full supervision, including unsupervised (Zhou et al., 2021; Honda et al., 2021), unpaired (Ben et al., 2021; Liu et al., 2021), and weakly-supervised (Zhang et al., 2022; Zhu et al., 2022) approaches. Unsupervised and weakly-supervised methods retrieve sentences from mismatching corpora, while unpaired methods use the original corpora but each sentence does not pair with the corresponding images. Table 2 indicates that our method surpasses these data-efficient methods by utilizing the generated pseudo sentences instead of fetching complete sentences. It is significant to note that our method even surpasses unpaired setting models that employ real images and real annotations but operate in an unpaired setting. This suggests that generating pseudo sentences may hold greater potential than retrieving complete sentences.

Category	Method	B1	B4	M	R	C	S
Unsupervised	UC-GAN (2019)	41.0	5.6	12.4	28.7	28.6	8.1
	TSGAN (2021)	46.2	6.9	13.0	32.3	28.9	8.3
	RWLSA (2021)	50.2	6.8	14.1	34.8	32.9	8.8
Unpaired	Gra-Align (2023b)	67.1	21.5	20.9	47.2	69.5	15.0
	SCS (2021)	67.1	**22.8**	21.4	47.7	74.7	15.1
	FG-SRE (2021)	67.8	21.8	**22.1**	48.4	75.7	**16.1**
Weakly-supervised	SGCL (2022)	63.6	20.2	20.0	47.9	55.0	13.5
	WS-UIC (2022)	-	21.5	20.1	45.8	65.7	13.6
LPMs + RaPSG	Ours (w/ CTX)	67.0	18.3	21.2	47.9	72.4	14.1
	Ours (w/ M^2)	67.5	18.9	20.9	48.5	75.3	14.7
	Ours (w/ DLCT)	69.5	19.4	21.1	48.6	75.9	14.5
	Ours (w/ DIFNet)	**70.5**	19.3	21.4	**49.0**	**78.1**	14.9

Table 2: The comparison of our method and others without fully supervision on MSCOCO benchmark.

4.4 Extension on Semi-Supervised Image Captioning Benchmarks

Since our approach works well in zero-shot and unsupervised settings, we also test whether it can deal with the data scarcity problem in a semi-supervised setting where only partial images have the corresponding text annotations. Specifically, we fol-

Model	B1	B4	M	R	C	S
Self Distillation (2021)	67.9	25.0	21.7	49.3	73.0	14.5
OSCAR (2020)	67.2	23.3	22.5	49.1	78.4	-
VisualGPT (2022a)	69.5	25.6	22.6	49.6	80.9	-
P^3 (2021)	68.8	27.5	23.4	51.0	84.5	16.1
M^2 (2020)	67.4	22.8	21.4	48.2	70.9	14.8
M^2 + Ours	68.7	23.3	22.1	49.5	83.1	15.8
DLCT (2021)	68.0	24.4	21.3	48.8	74.2	14.3
DLCT + Ours	72.4	27.1	23.1	51.5	90.8	16.5
CTX (2022)	71.6	26.4	23.2	50.8	85.4	16.7
CTX + Ours	72.4	27.7	**23.5**	51.5	90.8	**17.1**
DIFNet (2022)	70.8	24.9	22.2	49.7	81.3	15.3
DIFNet + Ours	**73.5**	**27.7**	23.1	**51.8**	**93.4**	16.7

Table 3: The comparison with SOTA 1% semi-supervised methods on MSCOCO captioning task.

Direction	Method	B4	M	R	C	S
COCO-to-Flickr30k	DeCap (2023b)	16.3	17.9	-	35.7	11.1
	CapDec (2022)	17.3	18.6	42.7	35.7	-
	CgT-GAN (2023)	17.3	19.3	43.9	47.5	**12.9**
	Ours (w/DIFNet)	**17.1**	**20.2**	**44.6**	**51.3**	11.6
Flickr30k-to-COCO	DeCap (2023b)	9.2	16.3	36.7	27.3	-
	CapDec (2022)	12.1	18.0	-	44.4	10.9
	CgT-GAN (2023)	15.2	19.4	40.9	58.7	**13.4**
	Ours (w/DIFNet)	**17.7**	**20.1**	**45.7**	**66.3**	12.2

Table 4: The Comparison with SOTA cross-domain methods on MSCOCO and Flickr30k captioning tasks.

low the existing semi-supervised image captioning benchmark (Chen et al., 2021). The proposed RaPSG is firstly optimized on the 99% images without caption labels. Then, the model is further fine-tuned on the rest of 1% labeled data. We repeat the experiments under 3 different selections of the 1% labeled samples and calculate the average performance as output. As shown in Table 3, compared with current approaches, our approach achieves a performance gain with 93.4 (+8.9) CIDEr score. This indicates that our generated pseudo sentences can alleviate the need for extensive annotations in semi-supervised captioning tasks.

4.5 Extension on Cross-Domain Image Captioning Benchmarks

To further verify the robustness of our model, we evaluate it on a cross-domain image captioning benchmark in comparison with SOTA models (Li et al., 2023b; Nukrai et al., 2022; Yu et al., 2023). Notably, we adhered to the established cross-domain image captioning benchmark protocol (Laina et al., 2019), albeit with the textual corpora replaced by the VG dataset. Table 4 demonstrates a significant improvement of our model, with CIDEr scores of 51.3 (+3.8) and 66.3 (+7.6) compared to competing models in two assessed categories.

Module	B1	B4	M	R	C	S	CLIP-S
RD	10.6	1.5	9.8	19.3	18.3	7.5	**62.7**
PS	48.1	8.8	18.0	33.8	39.3	13.3	47.6
PS+FF	59.4	15.2	20.2	39.6	56.0	13.9	48.2
D+PS	67.9	16.9	20.3	43.9	70.2	13.7	31.5
D+PS+FF	70.3	19.1	21.1	45.9	76.9	14.7	32.1
D+PS+FF+CR (2022)	67.6	17.0	19.9	44.4	69.4	13.3	31.4
D+PS+FF+CG	**70.5**	**19.3**	**21.4**	**46.0**	**78.1**	**14.9**	35.8

Table 5: Ablation study of different proposed modules conducted on the DIFNet. "RD" represents the retrieved region descriptions. "PS" means the generated pseudo sentences. "FF" is the fluency filter. "CG" represents the CLIP guidance. "D" is the DIFNet model. "CR (Cho et al., 2022)" represents training with Cho's CLIP reward instead of our CLIP guidance module.

4.6 Ablation Studies

Contribution of Designed Modules. We investigate the contribution of each designed module, as shown in Table 5. The RaPSG module is crucial for improving the model performance. In addition, the fluency filter is designed to filter out the unnatural sentences among pseudo sentence generation and leave the best one matching the given image. Figure 7 shows one case where the fluency filter picked up the best pseudo sentence based on its CIDEr score. Finally, we introduce CLIP guidance in the retrieval-augmented learning process, which drives the prediction to be semantically consistent with the given image by shrinking the cross-modal distance in the feature embedding space. To demonstrate the efficacy, our experiment, compared against Cho's CLIP reward (Cho et al., 2022), demonstrates that our CLIP guidance approach achieves better results.

Pseudo Sentence Quality. Here, we explore how to regulate the quality of generated sentences and the methods for producing high-quality sentences. Different from the explanation in Section 3.1, due to the absence of a metric to determine the optimal k for region descriptions, we first investigate the parameter m to ascertain the generation of high-quality pseudo sentences. Subsequently, based on the chosen value of m, we explore the selection of top-k. According to the left part of Figure 6, we decide to set $m = 4$ as it yields the best performance within the range of $[1, 6]$. Then, based on the m value, we explore $k \in [4, 24]$. As suggested by the middle segment of Figure 6, the quality of generated pseudo sentences initially improves with increasing k but eventually declines. According to the CIDEr scores, we set $k = 16$ and disregard the subsequent region descriptions. Lastly, we evaluate the efficacy of various summarization models. Based on the result in the right section of Figure 6,

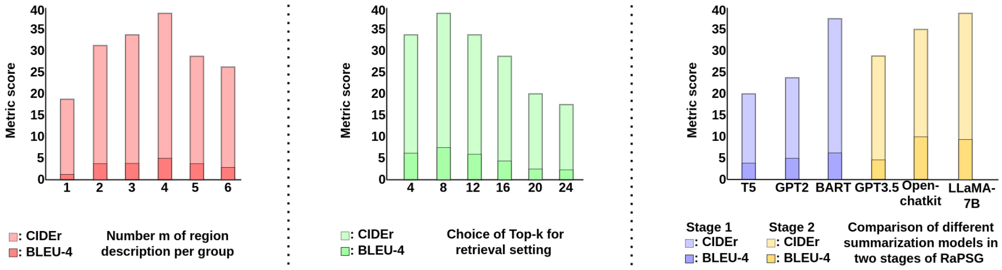

Figure 6: We conduct experiments to compare different settings, aiming to determine the most effective method for generating high-quality pseudo sentences based on region descriptions. These comparisons include, from left to right: the selection of hyperparameter m, the choice of hyperparameter k, and the evaluation of various summarization models.

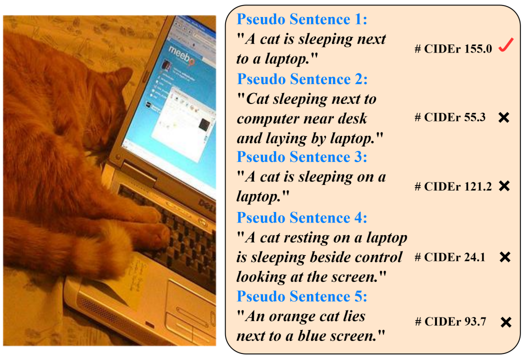

Figure 7: One example of how the fluency filter picks up the best sentence. Best viewed by zooming in.

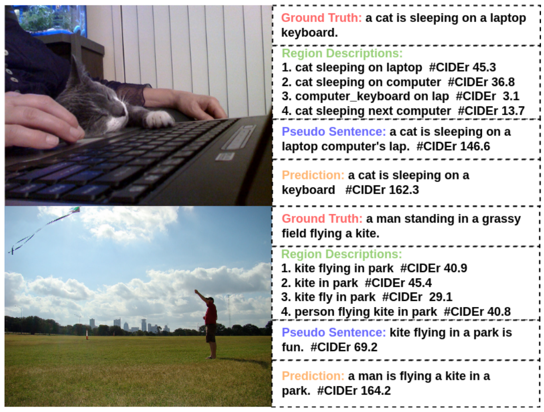

Figure 8: Qualitative results of our approach based on DIFNet model. Best viewed by zooming in.

Source	Method	B1	B4	M	R	C	S
VG	Ours (w/ DIFNet)	**70.5**	**19.3**	**21.4**	**46.0**	**78.1**	**14.9**
GCC	Ours (w/ DIFNet)	56.3	7.5	15.4	39.8	46.4	11.7

Table 6: The comparison between generated sentences and crawled sentences on DIFNet model.

we select the BART for the initial stage and the LLaMA-7B for the subsequent phase of RaPSG.

Generated Sentences VS Crawled Sentences. The previous comparison with unpaired models indicates the potential of the generated pseudo sentences. In this section, we verify whether generated pseudo sentences truly outperform crawled sentences under fair conditions. We pick up another popular corpus, named Google Concept Caption or GCC (Sharma et al., 2018), which is used in most unsupervised image captioning works (Laina et al., 2019; Guo et al., 2020; Honda et al., 2021). For a fair comparison, we also use the pre-trained CLIP (Radford et al., 2021) to fetch the most relevant individual descriptions from GCC dataset and utilize them as supervision for training. According to the results shown in Table 6, it is obvious that VG-based training presents better performance than the GCC-based one on all the metrics.

4.7 Qualitative Results

To highlight our approach's ability, we present qualitative results of our generated pseudo sentences and predictions in Figure 8. The pseudo sentences can avoid the appearance of irrelevant words and keep the diversity, which is attributed to the innovative combination of ranking, grouping, and summarization. However, some examples, while scoring well on CIDEr, may not make sense from a human perspective (e.g., "*kite flying in a park is fun*" should be "*flying a kite in a park is fun*"). This issue may stem from BART's difficulty in handling batches of similar objects, leading to disarranged relationships among region descriptions.

5 Conclusion

In this work, we propose a retrieval-augmented pseudo sentence generation method which leverages the prior knowledge from the frozen LPMs. The generated sentences can avoid the appearance of irrelevant words and keep the diversity of pseudo references, which is attributed to the innovative combination of ranking, grouping, and summarization. In addition, we design a fluency filter to sift the generated sentences and a CLIP guidance module to make the predicted captions semantically consistent with the given image. Our approach outperforms existing state-of-the-art captioning models across various scenarios such as zero-shot, unsupervised, semi-supervised, and cross-domain settings.

6 Limitation

Although our approach surpasses current SOTA captioning models in a range of scenarios, including zero-shot, unsupervised, semi-supervised, and cross-domain settings, it still has two limitations. First, compared to the basic model, it significantly increases time consumption due to the additional processing stages. The retrieval and summarization steps, coupled with the refinement using LPMs, add considerable computational compared with basic models. For instance, using "Ours (w/DLCT)" with a single RTX3090 as an example, each epoch in the training process takes approximately 55 minutes. The entire training process spans 36 epochs, totalling around 35 hours. Table 7 provides a comparison of time consumption against the baseline. Second, the quality of the generated pseudo sentences may be limited by the summarization capabilities of BART and LLaMA-7B. These models sometimes produce sentences where the words are correct but arranged in an unnatural order (Section 4.7). This occurs because BART and LLaMA-7B, while powerful, can struggle with maintaining the natural flow of language when summarizing complex or similar objects, leading to awkward phrasing or disordered relationships among the sentence elements.

Method	GPU	Each Epoch	Total
DLCT	1*RTX3090	35 min	22 hours
Ours (w/DLCT)	1*RTX3090	55 min	35 hours

Table 7: The comparison of time consumption between the baseline and our approach.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 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 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Huixia Ben, Yingwei Pan, Yehao Li, Ting Yao, Richang Hong, Meng Wang, and Tao Mei. 2021. Unpaired image captioning with semantic-constrained self-learning. *IEEE Transactions on Multimedia*, 24:904–916.

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. 2021. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning*, pages 1059–1071. PMLR.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022a. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022b. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Xianyu Chen, Ming Jiang, and Qi Zhao. 2021. Self-distillation for few-shot image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 545–555.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

Together Computer. 2023. OpenChatKit: An Open Toolkit and Base Model for Dialogue-style Applications.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134.

Dan Guo, Yang Wang, Peipei Song, and Meng Wang. 2020. Recurrent relational memory network for unsupervised image captioning. *arXiv preprint arXiv:2006.13611*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. 2021. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. *arXiv preprint arXiv:2104.13872*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379.

Arjit Jain, Pranay Reddy Samala, Preethi Jyothi, Deepak Mittal, and Maneesh Kumar Singh. 2021. Perturb, predict & paraphrase: Semi-supervised learning using noisy student for image captioning. In *IJCAI*, pages 758–764.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Chia-Wen Kuo and Zsolt Kira. 2022. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17979.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8928–8937.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023b. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Fenglin Liu, Meng Gao, Tianhao Zhang, and Yuexian Zou. 2021. Exploring semantic relationships for unpaired image captioning. *arXiv preprint arXiv:2106.10658*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017a. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017b. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195.

Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2286–2293.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2023. Gpt-3.5: Generative pre-trained transformer 3.5. https://www.openai.com/research/gpt-3-5. Accessed: 2023-06-12.

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*, pages 311–318.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. 2022. Difnet: Boosting visual information flow for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18020–18029.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.

Jiarui Yu, Haoran Li, Yanbin Hao, Bin Zhu, Tong Xu, and Xiangnan He. 2023. Cgt-gan: Clip-guided text gan for image captioning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2252–2263.

Chunhui Zhang, Chao Huang, Youhuan Li, Xiangliang Zhang, Yanfang Ye, and Chuxu Zhang. 2022. Look twice as much as you say: Scene graph contrastive learning for self-supervised image caption generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2519–2528.

Yucheng Zhou, Wei Tao, and Wenqiang Zhang. 2021. Triple sequence generative adversarial nets for unsupervised image captioning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7598–7602. IEEE.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Peipei Zhu, Xiao Wang, Yong Luo, Zhenglong Sun, Wei-Shi Zheng, Yaowei Wang, and Changwen Chen. 2022. Unpaired image captioning by image-level weakly-supervised visual concept recognition. *arXiv preprint arXiv:2203.03195*.

Peipei Zhu, Xiao Wang, Lin Zhu, Zhenglong Sun, Wei-Shi Zheng, Yaowei Wang, and Changwen Chen. 2023b. Prompt-based learning for unpaired image captioning. *IEEE Transactions on Multimedia*.

A Appendix

A.1 Comparison against Large Pre-Trained Models

In this section, we provide a detailed comparison of our method against zero-shot models. Unlike existing methods that rely on large external datasets for "mapper" learning, our approach introduces a more efficient learning process through self-supervised training with generated pseudo sentences. Table 8 highlights the effectiveness of our method on the MSCOCO benchmark, where it outperforms SimVLM (Wang et al., 2021), ReViLM (Yang et al., 2023), Flamingo3B (Alayrac et al., 2022), MiniGPT4 (Zhu et al., 2023a; Chen et al., 2023), and LLaVA (Liu et al., 2023b,a). Notably, widely recognized models BLIP (Li et al., 2022) and BLIP2 (Li et al., 2023a) are excluded from our comparison due to their use of COCO captions during pre-training. We also did not compare our method with REVEAL (Hu et al., 2023) due to the absence of official zero-shot results. Some data was sourced directly from the original papers, as many studies lack official GitHub implementations.

Additionally, we acknowledge that recent multimodal large language models like MiniGPT4 and LLaVA are not specifically optimized for generating short captions in the MSCOCO or Flickr style. While comparing these models on the MSCOCO dataset using metrics like BLEU and CIDEr might seem unfair, we specifically use the CLIP-S metric to evaluate the matching level between the target image and generated predictions. According to the CLIP-S performance in Table 8, our method generates captions that more closely match the target image. Figure 9 presents instance comparisons between our method, MiniGPT4, and LLaVA. It is evident that our model excels at generating concise captions, while LLaVA produces medium-length captions with more detail, and MiniGPT4 generates highly detailed descriptions. For example, our model's caption for the first image is "A cat is sitting on a laptop," which is succinct and to the point. In contrast, MiniGPT4 provides a much longer description: "The image shows a cat lying on top of a laptop computer. The cat has blue eyes and is brown and white in color. The laptop appears to be an older model with a black and grey colour scheme. There is a patterned blanket or cloth on the floor in the background." LLaVA offers a middle-size caption that "A cat is lying on a laptop computer, which is placed on a bed."

Figure 9: Two examples of comparing the prediction sentences from our model, MiniGPT4, and LLaVA. Best viewed by zooming in. It appears that our model excels at generating concise image captions, while LLaVA produces medium-length captions with more details, and MiniGPT4 generates highly detailed descriptions.

A.2 Comparison against Finetuning-Based Appraoches

In this section, we provide a comprehensive comparison of our weakly-supervised image captioning models across unsupervised, unpaired, and weakly-supervised scenarios, as shown in Table 9. Our approach and these methods share the assumption of the absence of grounded image-text pairs and propose using pseudo pairs for optimization. This includes benchmarking against unsupervised methods (Laina et al., 2019; Feng et al., 2019; Zhou et al., 2021; Guo et al., 2020; Honda et al., 2021), unpaired methods (Lu et al., 2017a; Ben et al., 2021; Liu et al., 2021; Zhu et al., 2023b), and weakly-supervised approaches (Zhang et al., 2022; Zhu et al., 2022). While unsupervised and weakly-supervised methods retrieve sentences from mismatched corpora and unpaired methods use original corpora without corresponding image-sentence pairs, our experiments reveal that our method, which employs generated pseudo sentences, surpasses these data-efficient techniques. Our method matches or exceeds the performance of unpaired models on most metrics and notably outperforms them in BLEU1 and CIDEr. This suggests that generating high-quality pseudo sentences holds more potential than retrieving complete sentences from corpora, including original ones.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S
				MSCOCO					
SimVLM$_{base}$ (2021)	-	-	-	9.5	11.5	-	24.0	7.5	-
SimVLM$_{large}$ (2021)	-	-	-	10.5	12.0	-	24.9	8.3	-
SimVLM$_{huge}$ (2021)	-	-	-	11.2	14.7	-	32.2	8.5	-
Re-ViLM$_{base}$ (2023)	-	-	-	17.0	-	-	51.2	-	-
Re-ViLM$_{medium}$ (2023)	-	-	-	17.9	-	-	53.6	-	-
Re-ViLM$_{large}$ (2023)	-	-	-	18.6	-	-	60.8	-	-
Flamingo3B (2022)	-	-	-	-	-	-	73.0	-	-
Flamingo9B (2022)	-	-	-	-	-	-	79.4	-	-
Flamingo80B (2022)	-	-	-	-	-	-	**84.3**	-	-
MiniGPT4-V1 (2023a)	23.6	16.2	9.8	5.8	20.9	21.2	0.0	14.4	34.0
MiniGPT4-V2 (2023)	28.6	19.4	12.5	6.3	24.4	27.3	0.0	17.9	35.5
LLaVA1.0 (2023a)	38.5	27.2	17.3	9.1	**26.7**	40.1	50.9	**24.2**	34.1
LLaVA1.5 (2023b)	30.6	21.7	14.8	10.1	24.8	37.4	41.8	22.6	31.5
Our Pseudo Sents.	48.1	27.7	15.7	8.8	18.0	33.8	39.3	13.3	**47.6**
Ours (w/CTX)	67.0	45.3	29.2	18.3	21.2	44.9	72.4	14.1	33.6
Ours (w/M)	67.5	46.5	30.3	18.9	20.9	45.5	75.3	14.7	34.3
Ours (w/DLCT)	69.5	47.5	30.8	**19.4**	21.1	45.6	75.9	14.5	34.5
Ours (w/DIFNet)	**70.5**	**48.1**	**31.0**	19.3	21.4	**46.0**	78.1	14.9	35.8
				Flickr30k					
Re-ViLM$_{base}$ (2023)	-	-	-	-	-	-	45.2	9.2	-
Re-ViLM$_{medium}$ (2023)	-	-	-	-	-	-	52.0	9.8	-
Re-ViLM$_{large}$ (2023)	-	-	-	-	-	-	52.1	10.0	-
Flamingo3B (2022)	-	-	-	-	-	-	60.6	-	-
Flamingo9B (2022)	-	-	-	-	-	-	61.5	-	-
Flamingo80B (2022)	-	-	-	-	-	-	**67.2**	-	-
MiniGPT4-V1 (2023a)	13.2	7.7	5.0	3.5	15.6	16.6	0.0	14.8	32.3
MiniGPT4-V2 (2023)	17.5	11.5	8.5	6.6	22.0	23.9	0.0	20.0	32.6
LLaVA1.0 (2023a)	48.0	31.8	20.5	13.0	**23.4**	43.1	52.5	**17.1**	33.7
LLaVA1.5 (2023b)	35.2	22.2	11.7	7.7	21.9	29.0	34.1	17.0	30.5
Our Pseudo Sents.	43.2	28.0	17.4	14.5	17.1	40.8	21.2	9.3	**45.4**
Ours (w/CTX)	51.7	37.5	24.6	17.8	21.0	46.7	53.3	10.7	32.6
Ours (w/M)	54.6	39.6	25.9	17.5	20.7	47.3	56.8	11.2	33.8
Ours (w/DLCT)	54.1	38.8	25.8	18.1	22.6	47.2	58.4	11.5	34.1
Ours (w/DIFNet)	**55.9**	**39.9**	**26.6**	**18.2**	23.1	**47.5**	59.1	11.8	33.9

Table 8: The detailed comparison of our method and other zero-shot models on MSCOCO and Flickr30k benchmark.

A.3 Generated Sentences VS Crawled Sentences

Section 4.6 provides a brief explanation supported by experimental results on why generated sentences yield better predictions compared to crawled sentences. In this section, we present specific instances for a more detailed explanation. Figure 10 displays examples of two generated pseudo sentences from VG (Krishna et al., 2017) and GCC (Sharma et al., 2018) respectively and real human annotations. We can observe that descriptions from the GCC dataset contain many words or phrases that do not match the given image. This low-relevance information cannot be effectively distinguished from valuable information by the LPMs, which leads to misleading sentence generation. For instance, the GCC-based description "A young man standing, in a red jacket and baseball cap, texting with his cell phone, his shadow behind him" includes irrelevant details that do not correspond to the image, resulting in a CIDEr score of 12.3. This demonstrates how irrelevant information can make sentences excessively long and convoluted, further degrading prediction performance across all metrics (Feng et al., 2019). In contrast, the VG-based description "A man takes a picture of himself standing in a hallway" is more concise and relevant, resulting in a higher CIDEr score of 146.9. This example illustrates how our method of generating high-quality pseudo sentences focuses on relevant content, thereby improving the overall prediction accuracy and performance.

A.4 How does the fluency filter select the optimal pseudo sentence based CIDEr metric

Section 4.6 provides a simple example of how the fluency filter selects the most appropriate sentences. In this section, we offer a clearer explanation. Figure 11 showcases five generated pseudo sentences along with their corresponding CIDEr scores. The

Category	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Unsupervised	SME-GAN (2019)	-	-	-	6.5	12.9	35.1	22.7	7.4
	UC-GAN (2019)	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
	TSGAN (2021)	46.2	26.8	13.5	6.9	13.0	32.3	28.9	8.3
	RM (2020)	51.2	29.5	15.4	8.3	14.0	35.0	29.3	9.6
	RWLSA (2021)	50.2	28.5	13.9	6.8	14.1	34.8	32.9	8.8
Unpaired	Gra-Align (2017b)	67.1	47.8	32.3	21.5	20.9	47.2	69.5	15.0
	SCS (2021)	67.1	47.9	33.4	22.8	21.4	47.7	74.7	15.1
	FG-SRE (2021)	67.8	48.7	33.6	21.8	22.1	48.4	75.7	**16.1**
	PL-UIC (2023b)	-	-	-	**25.0**	**22.6**	**49.4**	77.9	15.2
weakly-supervised	SGCL (2022)	63.6	45.4	30.7	20.2	20.0	47.9	55.0	13.5
	WS-UIC (2022)	-	-	-	21.5	20.1	45.8	65.7	13.6
LPM + RaPSG	Ours (w/CTX)	67.0	45.3	29.2	18.3	21.2	44.9	72.4	14.1
	Ours (w/M)	67.5	46.5	30.3	18.9	20.9	45.5	75.3	14.7
	Ours (w/DLCT)	69.5	47.5	30.8	19.4	21.1	45.6	75.9	14.5
	Ours (w/DIFNet)	**70.5**	**48.1**	**31.0**	19.3	21.4	46.0	**78.1**	14.9

Table 9: The comparison of our method and other models without fully supervision on MSCOCO benchmark.

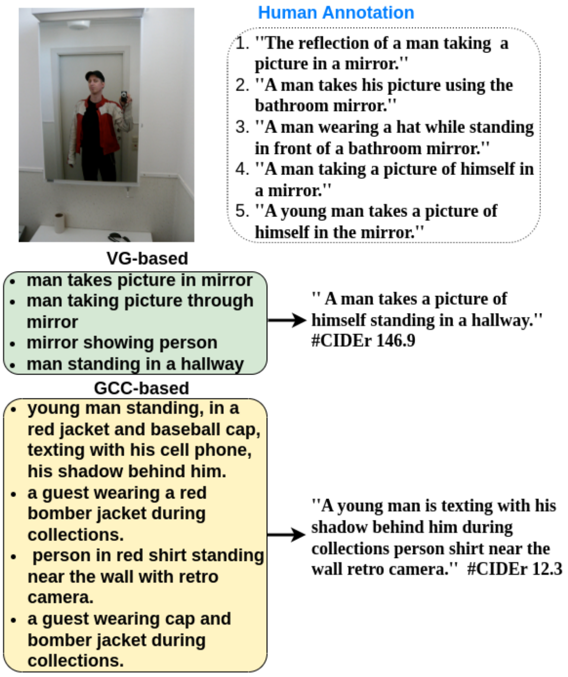

Figure 10: An example of how different corpora affect the fluency of generated sentences. Best viewed by zooming in.

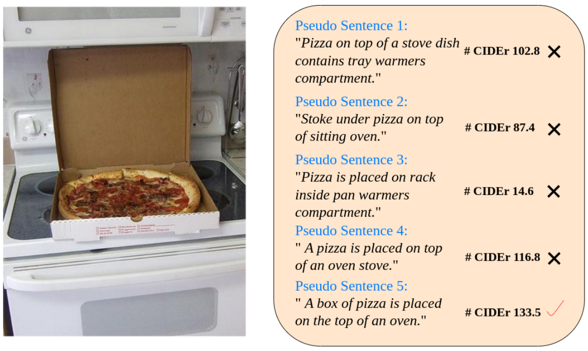

Figure 11: One example of how the fluency filter picks up the best sentence. Best viewed by zooming in.

sentence "A box of pizza is placed on the top of an oven," marked with a red checkmark ($\surd$) and boasting a CIDEr score of 133.5, is highlighted as the best choice. Although the other four sentences are contextually accurate, they are not selected, as indicated by the red crosses ($\times$) and their lower CIDEr scores. This example illustrates how the fluency filter effectively identifies the most relevant and high-scoring sentence from a set of generated options, enhancing the overall quality and accuracy of image captioning.

A.5 How to generate high-quality pseudo sentences with RaPSG

Choice of hyperparameter m In this section, we delve into more details on how to determine the hyperparameter m, which was briefly explained in Section 4.6. Table 10 presents a performance comparison based on varying the number m of region descriptions used to generate pseudo sentences. This comparison evaluates outcomes across several metrics, including BLEU-1 through BLEU-4, METEOR, ROUGE, CIDEr, and SPICE. The results indicate that using four region descriptions ($m = 4$) yields the best performance according to these metrics. This optimal choice of m suggests that incorporating more than four descriptions does not significantly enhance the quality of the generated sentences and may even lead to a degradation in performance. This could be due to the inclusion of redundant or less relevant information, which can dilute the clarity and relevance of the pseudo sentences. Thus, our findings highlight the importance of selecting an appropriate number of region descriptions to balance detail and relevance, ensuring the generation of high-quality pseudo sentences.

Choice of hyperparameter k In this section, we provide a clearer explanation of the choice of the hyperparameter k. Table 11 presents data on how different values of k affect the retrieval of top-k region descriptions and their subsequent perfor-

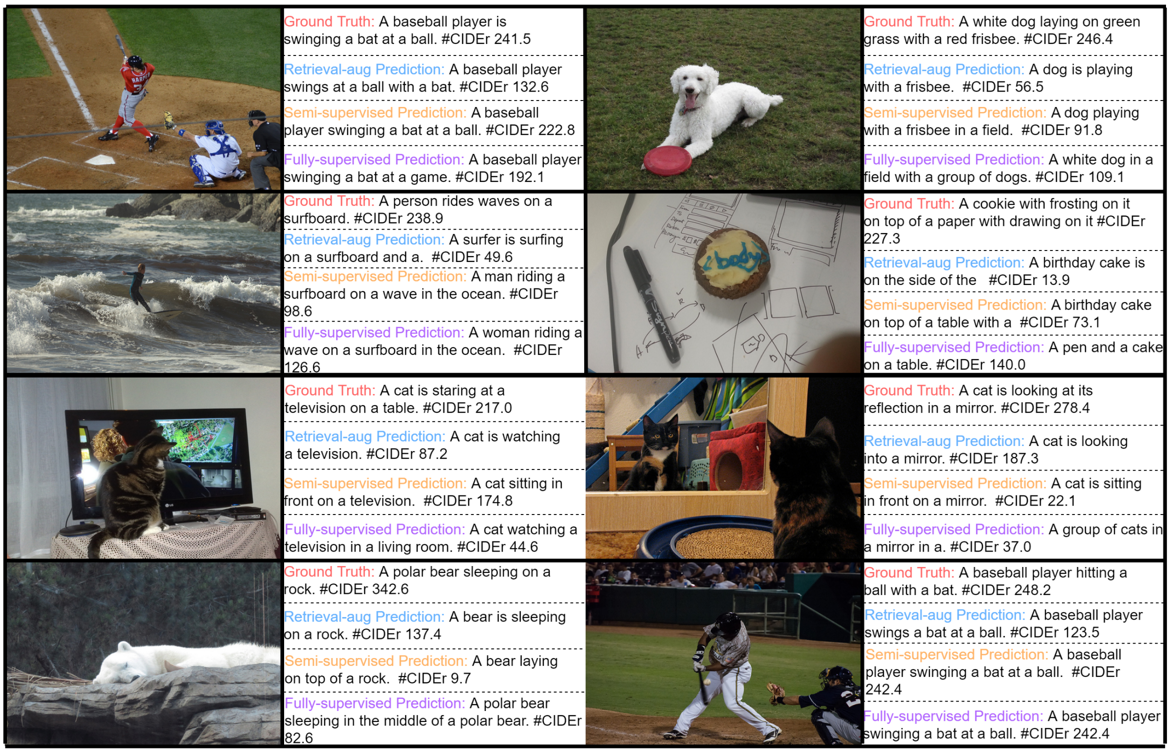

Figure 12: Qualitative results for different supervision levels. Best viewed by zooming in.

Parameter	B1	B2	B3	B4	M	R	C	S
m=1	10.6	4.7	2.5	1.5	9.8	19.3	18.3	7.5
m=2	41.0	21.9	11.4	5.8	14.5	29.5	32.5	9.3
m=3	43.5	22.7	12.0	5.9	15.2	30.1	35.3	10.0
m=4	**45.9**	**24.4**	**12.7**	**6.4**	**15.9**	**30.7**	**37.2**	10.4
m=5	38.5	20.1	10.2	5.0	14.7	27.2	28.1	10.5
m=6	37.5	19.5	9.9	4.8	14.8	27.0	26.6	**10.6**

Table 10: The comparison of different choices of m on assigned region descriptions number.

Parameter	B1	B2	B3	B4	M	R	C	S
k=4	45.1	23.3	11.9	5.9	15.6	30.0	34.7	10.1
k=8	45.9	24.4	12.7	6.4	15.9	30.7	37.2	10.4
k=12	44.6	22.6	11.2	5.5	15.5	29.5	34.2	9.9
k=16	41.6	20.5	9.9	4.7	14.4	27.6	28.9	8.8
k=20	35.3	17.0	7.6	3.6	13.1	26.7	19.5	7.7
k=24	33.4	15.0	6.9	3.3	12.7	24.9	18.8	8.2

Table 11: The comparison of different choices of k on region description retrieval number.

mance across various metrics, including BLEU, METEOR, ROUGE, CIDEr, and SPICE. The results indicate that the quality of pseudo sentences is optimal when $k = 8$, as evidenced by the peak performance in most metrics at this value. Beyond $k = 8$, performance tends to decline, suggesting that retrieving a larger number of top-k region descriptions does not necessarily enhance the quality of pseudo sentences. In fact, including too many descriptions may introduce noise and less relevant information, which can dilute the clarity and coherence of the generated sentences. Therefore, we have chosen $k = 16$ as the cut-off point, where the quality remains good before it starts to significantly decline, as reflected in the metrics. This careful selection of k ensures that we balance the detail and relevance of the region descriptions, leading to the generation of high-quality pseudo sentences. The findings underscore the importance of selecting an appropriate value for k to maximize the effectiveness of our retrieval-augmented pseudo sentence generation process.

Choice of LLMs In this section, we present additional experiments to explain our selection of BART and LLaMA-7B as the large language models (LLMs) for transforming region descriptions into pseudo sentences. Table 12 compares the performance of various methods and models in summarizing region descriptions into pseudo sentences across two stages of processing. The evaluation metrics include BLEU-1, BLEU-4, METEOR, and CIDEr, providing a comprehensive view of each model's effectiveness. The results indicate that BART outperforms other models in the initial

Stage	Type	Method	B1	B4	M	C
One	LM	T5 (2020)	35.3	3.8	13.3	19.5
		GPT2 (2019)	38.7	5.0	12.4	23.5
		BART (2019)	**45.9**	**6.4**	**15.9**	**37.2**
Two	LLM	GPT3.5 (2023)	38.1	4.5	15.8	29.5
		Openchatkit (2023)	44.5	**9.6**	14.1	36.3
		LLaMA-7B (2023)	**48.1**	8.8	**18.0**	**39.3**

Table 12: The comparison of different summarization models on pseudo sentence generation.

stage due to its exceptional summarizing capabilities. BART's ability of distilling concise and relevant information from region descriptions makes it ideal for the first step of the RaPSG process, ensuring that the foundational pseudo sentences are both informative and accurate. In the second stage, LLaMA-7B is chosen based on its high scores across all metrics. LLaMA-7B excels in enhancing the pseudo sentences generated by BART, refining them to be more fluent and contextually appropriate. Its advanced language model capabilities ensure that the final pseudo sentences are not only precise but also exhibit a natural flow, which is crucial for improving image captioning performance. By combining BART's superior summarization skills in the initial stage with LLaMA-7B's advanced language processing in the second stage, our RaPSG process achieves optimal results. This two-stage approach leverages the strengths of both models, resulting in high-quality pseudo sentences that enhance the overall performance of our image captioning system. The experiments underscore the importance of selecting the right models for each stage, highlighting why BART and LLaMA-7B are the best choices for our methodology.

B Qualitative Results

In Section 4.7, we present qualitative results of our generated pseudo sentences to highlight the captioning ability of our approach. We showcase qualitative results for various caption predictions across different supervision levels, comparing them with the ground truth and providing their CIDEr scores, as shown in Figure 12. The examples include images of a baseball player, a surfer, and a cat, among others.

Baseball Player. The ground-truth caption describes a baseball player swinging at a ball. Predictions from retrieval-augmented, semi-supervised, and fully-supervised models offer varying levels of accuracy, with the semi-supervised prediction scoring a CIDEr of 222.8, suggesting a close match to the ground truth.

Surfer. The ground truth involves a person riding waves on a surfboard. Different models interpret this with varying degrees of accuracy. The fully-supervised model scores the highest CIDEr at 126.6, indicating a strong match with the ground truth. Each image and set of predictions illustrate the effectiveness of the models in generating accurate captions, with CIDEr scores providing a quantitative measure of their precision compared to the ground truth.