Mining Fine-Grained Image-Text Alignment for Zero-Shot Captioning via Text-Only Training

Longtian Qiu[1*], **Shan Ning**[1*], **Xuming He**[1,2]

ShanghaiTech University, Shanghai, China[1]
Shanghai Engineering Research Center of Intelligent Vision and Imaging[2]
{qiult, ningshan2022, hexm}@shanghaitech.edu.cn

Abstract

Image captioning aims at generating descriptive and meaningful textual descriptions of images, enabling a broad range of vision-language applications. Prior works have demonstrated that harnessing the power of Contrastive Image Language Pre-training (CLIP) offers a promising approach to achieving zero-shot captioning, eliminating the need for expensive caption annotations. However, the widely observed modality gap in the latent space of CLIP harms the performance of zero-shot captioning by breaking the alignment between paired image-text features. To address this issue, we conduct an analysis on the CLIP latent space which leads to two findings. Firstly, we observe that the CLIP's visual feature of image subregions can achieve closer proximity to the paired caption due to the inherent information loss in text descriptions. In addition, we show that the modality gap between a paired image-text can be empirically modeled as a zero-mean Gaussian distribution. Motivated by the findings, we propose a novel zero-shot image captioning framework with text-only training to reduce the modality gap. In particular, we introduce a subregion feature aggregation to leverage local region information, which produces a compact visual representation for matching text representation. Moreover, we incorporate a noise injection and CLIP reranking strategy to boost captioning performance. We also extend our framework to build a zero-shot VQA pipeline, demonstrating its generality. Through extensive experiments on common captioning and VQA datasets such as MSCOCO, Flickr30k and VQAV2, we show that our method achieves remarkable performance improvements. Code is available at https://github.com/Artanic30/MacCap.

1 Introduction

Image captioning is a fundamental task in vision-language understanding that involves generating natural language descriptions for a given image. It plays a critical role in facilitating more complex vision-language tasks, such as visual question answering (**???**) and visual dialog (**???**). The mainstream image captioning methods (**????**) require expensive human annotation of image-text pairs for training neural network models in an end-to-end manner. Recent developments in Contrastive Image Language Pre-training (CLIP) (**?**) have enabled researchers to explore a new paradigm, zero-shot image captioning, through text-only training. In particular, CLIP learns a multi-modal embedding space where semantically related images and text are encoded into features with close proximity. As such, if a model learns to map the CLIP text features to their corresponding texts, it is feasible to generate image captions from the CLIP image features without needing supervision from caption annotations.

One main advantage of this zero-shot captioning paradigm is that it enables a Large Language Model (LLM) (**??**) with image captioning capabilities using only text data and affordable computational resources. Despite the impressive performance achieved by recent powerful multimodal models (**??**), they typically require large-scale, high-quality human-annotated data and expensive computational resources for fine-tuning an LLM. Zero-shot captioning methods can significantly reduce such costs, which is particularly important in situations of data scarcity and limited resources. Moreover, recent work (**???**) demonstrates that other vision-language tasks, such as VQA, can be addressed by LLMs and image captions. Consequently, the paradigm of zero-shot captioning has the potential to pave the way to solving complex vision-language tasks with LLMs through efficient text-only training.

A critical challenge in zero-shot image captioning through text-only training is to mitigate a widely observed phenomenon known as the *modality gap*. While the features of paired texts and images are close in the CLIP embedding space, there remains a gap between them (**?**). This gap often results in inaccurate mappings from the image embeddings to the text ones. Consequently, without fine-tuning with paired data, it significantly impairs the performance of zero-shot image captioning. Several works have attempted to address the modality gap in zero-shot image captioning, relying mainly on two strategies: (1) The first strategy leverages a memory bank from training text data to project visual embeddings into the text embedding space (**?**). However, this projection prevents it from representing any semantic content outside the distribution of the memory bank features and introduces extra inference costs; (2) The second approach injects noise during training to encourage the visual

*These authors contributed equally. This work was supported by Shanghai Science and Technology Program 21010502700, Shanghai Frontiers Science Center of Human-centered Artificial Intelligence, and the National Nature Science Foundation of China under Grant 62350610269.

embeddings to be included inside the semantic neighborhood of the corresponding text embeddings (**?**). Nonetheless, the noise injection tends to diffuse the distribution of visual inputs at the cost of weakening the semantic correlation between paired images and text embeddings.

To tackle these challenges, we first conduct a thorough analysis of the CLIP feature space, leading to two key observations. First, most text descriptions are unable to fully capture the content of their paired images. However, we empirically find that the visual embedding of certain local regions of an image, named image subregions, have closer proximity to the text embedding of the paired caption. Integrating such image subregions with the global image representation generates a tighter alignment between image and text. Additionally, we analyze the distribution of the gap between the CLIP features of image or subregion-text pairs and find that it closely resembles a zero-mean Gaussian distribution.

Based on our findings, we propose a novel zero-shot image captioning framework, named ***M**ining Fine-Grained Image-Text **A**lignment in **C**LIP for **Cap**tioning* (MacCap), to address the aforementioned challenges. In this framework, we introduce a region-aware cross-modal representation based on CLIP and an effective unimodal training strategy for an LLM-based caption generator. Our cross-modal representation maps an input image into the language space of LLMs and consists of two main components. First, we design a *subregion feature aggregation* module to fuse both global and subregion-level CLIP image features, resulting in a smaller gap between the corresponding CLIP text embedding. Next, we introduce a learnable adaptor-decoder to transform the CLIP representation into the LLM's language space. To train our model with text-only data, we develop a robust procedure to learn a projection from the CLIP embedding space to a language representation, enabling the LLM to reconstruct captions. Specifically, our learning procedure first injects noise into our region-aware CLIP-text representation, mimicking the modality gap between image and text features. This is followed by a multiple sampling and filtering step that leverages the CLIP knowledge to improve the quality of the captioning. In addition to the image captioning task, we further extend our framework to build a zero-shot VQA pipeline, demonstrating the generality of our cross-modal representation for more complex vision-language tasks.

We evaluate our framework on several widely-adopted image captioning benchmarks, such as MSCOCO (**?**) and Flickr30k (**?**), as well as a standard VQA benchmark, VQAV2 (**?**). Our extensive experiments cover multiple vision-language tasks, including zero-shot in-domain image captioning, zero-shot cross-domain image, and zero-shot VQA. The results not only demonstrate the superiority of our methods but also validate our findings on the CLIP embedding space.

2 Related Work

Zero-shot Image Captioning Zero-shot image captioning is an emerging task setting of image captioning, where captions are generated without relying on training with annotated image data. While some approaches (**???**) exploit large noisy image-text datasets, demanding high data and computational resources, an alternative is to leverage pre-trained large models, which is more suitable for low-data scenarios.

The use of pre-trained multi-modality models has enabled progress in text-only training for image captioning, which has demonstrated promising results. CapDec (**?**) utilizes CLIP embeddings and employs a noise injection training strategy for text-only training. Similarly, DeCap (**?**) employs a memory bank to project visual features into the text modality. Furthermore, methods like MAGIC (**?**) and ZeroCap (**?**) achieve zero-shot captioning without a typical training stage, with MAGIC introducing a CLIP-based score to guide language model generation and ZeroCap employing iterative optimization during inference.

Vision-language Models Recent advancements (**?????**) in Vision-Language Models (VLM) have led to significant progress in various downstream tasks (**??????**). Extensive research efforts (**????**) have analyzed the multi-modal embedding space learned through contrastive training. Recently, a geometric phenomenon known as the *modality gap* has been identified in (**?**). This gap arises from misalignment between text and image embeddings of CLIP, impacting their shared representation. The *modality gap* is attributed to the optimization process of contrastive learning and random initialization of different encoders. Addressing this *modality gap* is crucial for enhancing zero-shot capabilities, especially in scenarios with limited fine-tuning opportunities.

3 CLIP Embedding Space Analysis

In this section, we present a detailed analysis of the CLIP embedding space. CLIP offers a pre-trained joint embedding space, enabling zero-shot vision-language tasks. However, there is still a *modality gap* present in the CLIP embedding space (**?**). This gap refers to a geometric property where image and text embeddings occupy distinct regions within the embedding space. We further analyze and demonstrate that the modality gap arises from an inherent ambiguity in matching visual and linguistic embeddings. Empirically, we show that integrating subregion image information with global information can reduce the gap between linguistic and visual representations. Additionally, we explore the disparity between CLIP's text and image representations, conducting an empirical study that reveals the difference follows a Gaussian distribution. These findings inspire our subsequent method design.

3.1 Modality Gap of CLIP Representations

As demonstrated in (**?**), the modality gap phenomenon is primarily caused by the presence of mismatched text-image data during pre-training, further exacerbated by the contrastive learning objective employed in CLIP. However, we observe that even correctly paired image-text may exhibit different semantic contents due to several reasons, including 1) incomplete image description by text; 2) ambiguity in text interpretation; 3) diverse valid descriptions with varying focus on image subregions. An example is illustrated in Figure **??**. Consequently, we observe that certain subregions of

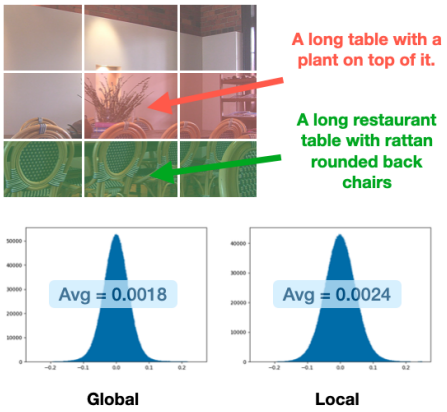

Figure 1: The upper half of this figure is an example of the misalignment in paired image and text description. The lower half of this figure is the distribution of modality gap between text representation and global / local image representation respectively.

	Pair Cosine Similarity		
	Mean	Max	Min
Global representation	0.330	0.446	0.228
Mix representation	0.352	0.422	0.242

Table 1: In this table, we show the mean, max and min value of similarity between text feature and global/mix feature. We find that after adding subregion information, the mean, max and min value all increase. This observation shows that introducing subregion image information benefit the alleviation of modality gap.

an image typically exhibit a smaller modality gap with a specific text description, as illustrated in Figure **??**. To validate this characteristic, we conduct a statistical analysis on the MSCOCO validation set. Specifically, we calculate the cosine similarity between the visual and text features, where the visual features are obtained from the global CLS token or the last layer of the ViT encoder representing the subregions. The results indicate that in **33**% of cases, one of the subregion features has a higher similarity than the global one.

Motivated by the above observation, we propose integrating the global and subregion representations by adding them together. We evaluate the similarity scores between text features and their corresponding subregion-augmented image representations. As illustrated in Table **??**, our augmented image embeddings achieve higher similarity scores when compared with corresponding text embeddings without any fine-tuning, effectively reducing the modality gap.

3.2 Distribution of Modality Gap

We further investigate the distribution of the modality gap through an empirical study on the MSCOCO validation set. Specifically, for each image with its caption, we compute the difference between the image and text embeddings. We calculate these differences separately for the global image feature and the subregion feature. Given these differences, we pool all the feature dimensions together and compute a histogram and mean of the dimension-wise differences between the two modalities. As depicted in the lower half of Figure **??**, we observe that the mean of the gap distribution is close to zero, and the overall distribution resembles a Gaussian distribution for both global and subregion representations. Based on this observation, we adopt a unimodal learning strategy that involves Gaussian noise injection with the text features. This strategy allows us to mimic the image features in the cross-modal inference stage. For a more formal description of our analysis, please refer to the supplementary material.

4 Methodology

4.1 Method Overview

In zero-shot image captioning, the image captioning model is trained with only caption text data. This is possible since CLIP learned a joint space where semantically-related image feature I_c and text feature T_c have closer proximity. By training the model to generate captions conditioned on their CLIP text feature, the model becomes capable of generating captions based on the CLIP image feature without any supervision from paired caption data.

Specifically, we have a caption text corpus $T = \{t^i | i \in \mathbb{N}\}$ and three network modules, contrastive vision language model CLIP with parameter θ_c, pre-trained large language model with parameter θ_l and a learnable adaptor decoder with parameter θ. In text reconstruction training, the adaptor module converts text t_i's CLIP text feature $T_c^i \in R^D$ to a prefix embedding $E^i \in R^{N_q \times D_l}$, where N_q is the length of prefix embedding, D_l is the dimensionality of language model and D is the dimensionality of CLIP feature. The language model generates text t_i based on the prefix embedding E^i. During training, we freeze the parameter of CLIP θ_c and language model θ_l, which makes the adaptor decoder with parameter θ a plug-and-play module to achieve zero-shot captioning. We can formulate the process as follows:

$$\underset{\theta}{\text{maximize}} \quad p(t^i | T_c^i, \theta_c, \theta_l) \tag{1}$$

In inference, CLIP extracts image feature $I_c \in R^D$ for an image. The adaptor decoder converts it to prefix embedding and the language model generates a text describing the image content. We present the overall pipeline in Figure **??** and explain the details of the pipeline in the following sections.

4.2 Text Reconstruction Training

Region Noise Injection The text reconstruction task aims to train our framework to generate text based on the CLIP text features T_c, as illustrated in Figure **??**. Our observation on the CLIP Embedding space demonstrates that the gap between the text embedding and subregion image representation satisfies a Gaussian distribution. To mitigate the gap and maintain a consistent format with visual features in inference, we propose region-aware noise injection. In detail, We first encode t from text corpus T with CLIP text encoder to get text features T_c. The T_c is repeated N_{cr} times and added different

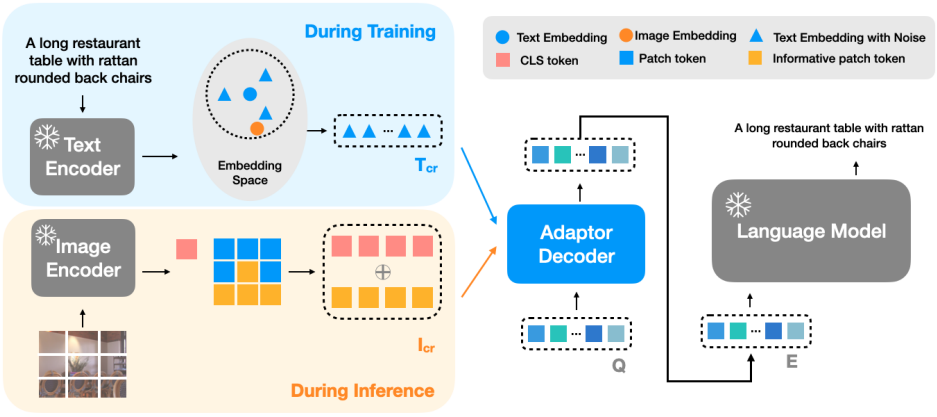

Figure 2: **An overview of MacCap pipeline.** MacCap learns to generate text based on region noise injected CLIP text feature in text reconstruction training. During inference, MacCap can generate caption without paired data in training. The CLIP and language model are kept frozen in both stages.

noise $n_i \in R^D, i \in \{1...N_{cr}\}$ from a uniform distribution with zero means and σ variance. We apply L2 normalization to the resulting text region feature $T_{cr} \in R^{N_{cr} \times D}$. The process is formulated as follows:

$$T_c = \text{CLIP}(t) \in R^D \tag{2}$$

$$T_{cr} = \text{Concat}(T_c, T_c, \dots T_c) \in R^{N_{cr} \times D} \tag{3}$$

$$T^i_{cr} = \text{L2Norm}(T^i_{cr} + n_i) \quad n_i \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

where L2Norm is a $l2$-normalization and Concat is the concatenation operation. The elements in T_{cr} form a cluster of points in the CLIP embedding space centered around T_c, which represent captions semantically similar to caption t.

Adaptor Decoder The adaptor decoder is designed to project the feature in the CLIP embedding space to the language model embedding space, enabling the language model to generate text based on image or text features in CLIP. Specifically, we have a set of learnable queries $Q \in R^{N_q \times D}$, a transformer decoder (**?**), and an MLP module. The learnable queries Q are first updated by self-attention and then fed into a cross-attention module with T_{cr} as the input key and value. The output feature is processed by a feed-forward network to obtain updated learnable queries $Q' \in R^{N_q \times D}$. Finally, Q' is projected by the MLP module to get the prefix embedding $E \in R^{N_q \times D_l}$. The process is formulated as follows:

$$Q = \text{SelfAttn}(Q) \in R^{N_q \times D} \tag{5}$$

$$Q = \text{CrossAttn}(Q, T_{cr}) \in R^{N_q \times D} \tag{6}$$

$$Q' = \text{FFN}(Q) \in R^{N_q \times D} \tag{7}$$

$$E = \text{MLP}(Q') \in R^{N_q \times D_l} \tag{8}$$

Through cross-attention in the adaptor decoder, the learnable queries Q adaptively select informative parts in T_{cr}. At last, the language model generates the input text t based on the prefix embedding E. Our objective can be described as:

$$L_{recons}(\theta) = -\frac{1}{|t|} \sum_{i=1}^{|t|} \log P_\theta(w_i | w_{<i}, E, \theta_c, \theta_l) \tag{9}$$

where w_i is the i^{th} words in t.

4.3 Zero-shot Caption Generation

Sub-region Feature Aggregation Zero-shot caption generation aims to generate captions with a text-only trained adaptor decoder, CLIP, and a language model. Based on our observation that image subregion features exhibit higher similarity to caption features, we propose sub-region feature aggregation to integrate the image global information with subregion information. Specifically, the ViT-based visual encoder processes images by dividing them into patches and incorporates a class token. In the last layer of ViT, we define the image patch features as $I_p \in R^{(N_p+1) \times D_v}$, where N_p is the number of image patches, D_v is the dimensionality of ViT and the first element in I_p is the class token $I_p[0] \in R^D$. The global image feature $I_c \in R^D$ is obtained by a linear projection on class token $I_p[0]$. We select the patches with the top N_{cr} score in the class token's attention weight and denote them as informative patch tokens. The subregion features $I_s \in R^{N_{cr} \times D}$ are obtained by aggregating patch features in informative patch based on corresponding attention weight $A \in R^{N_{cr} \times (N_p+1)}$. Finally, the subregion-enhanced image feature $I_{cr} \in R^{N_{cr} \times D}$ is acquired by taking the average of I_c and I_s. The process can be formulated as follows:

$$I'_p = \text{Linear}(I_p) \in R^{(N_p+1) \times D} \tag{10}$$

$$I_c = I'_p[0] \in R^D \tag{11}$$

$$I_s = A I'_p \in R^{N_{cr} \times D} \tag{12}$$

$$I_{cr} = \text{Concat}(I^1_s + I_c, \dots, I^{N_{cr}}_s + I_c) \in R^{N_q \times D} \tag{13}$$

where I'_p is the projected image patch features, and Linear is the linear projection that projects visual features to CLIP multimodal embedding space. The I_{cr} represents the image feature in CLIP space and is used to generate text in the same way as text region feature T_{cr}.

Multiple Sampling and Filtering The *modality gap* is alleviated by the region noise injection in text reconstruc-

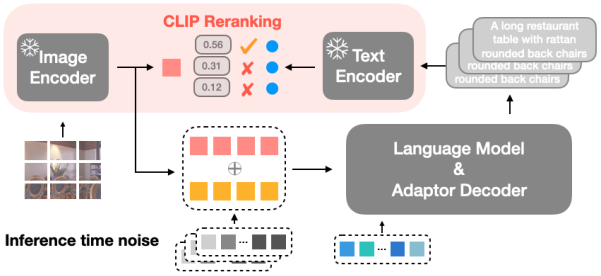

Figure 3: **Multiple sampling and filtering pipeline** During inference, each image uses noise to generate several different captions, which are reranked by CLIP to output the best.

tion training, however, the noise introduces additional uncertainty. We propose a *multiple sampling and filtering* strategy that incorporates inference-time noise and CLIP reranking to address the issue and boost performance, which is illustrated in Figure **??**. Specifically, we introduce inference time noise sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$ into the subregion enhanced image feature $I_{cr} \in R^{N_{cr} \times D}$. We generate text based on the perturbed subregion enhanced image feature I_{cr}, and repeat this process S times to generate S diverse texts. CLIP is utilized to evaluate the cosine similarity between the generated texts and the image. We select the text with the highest similarity as the predicted image caption. This strategy leverages CLIP knowledge to improve the generation quality of our model.

4.4 Zero-shot Visual Question Answering with Captioning

In this part, we illustrate the potential extensibility by showing the pipeline for zero-shot VQA with text-only trained MacCap. As there is no supervision from VQA data, we convert the image into a caption by MacCap. To answer the question about the image, we only use the LLM in MacCap to do open-end text generation based on a VQA prompt. The prompt contains the question and image caption, which is *"[caption] Question: [question] Answer:"*. Given the inherent difficulty of the zero-shot VQA task, we transform the VQA task into an image-text retrieval task. Specifically, the generated answer is embedded by the CLIP text encoder and computed cosine similarity with CLIP text embedding from answer candidates.

5 Experiments

5.1 Experimental setting

Datasets and Evaluation Our experimental evaluations are performed on two common benchmark datasets for image captioning: MSCOCO (**?**) and Flickr30k (**?**). The MSCOCO dataset contains over 11,000 images, each associated with five captions. We follow previous works (**?**) using the widely-used Karpathy et al. split, which partitions the dataset into 5000 images for validation and 5,000 for testing. The Flickr30k dataset consists of 29,000 images for training and 1,000 images for testing. For training, we use the texts from MSCOCO, Flickr30K, and CC3M (**?**) datasets. Following previous works (**?**), we remove the sentences with more than fifteen words. The resulting training text corpus has 566,747 sentences for MSCOCO, 144,998 sentences for Flickr30K, and 3,302,783 sentences for CC3M. To evaluate the performance of our models, we report the results on four standard evaluation metrics for captioning: BLEU (**?**), METEOR (**?**), CIDEr (**?**), and SPICE (**?**). Additionally, we use the popular visual question answering benchmark VQAV2 (**?**) to evaluate the model's ability on complex visual tasks. The number of answer candidates for VQAV2 is 3,128. We randomly chose 20,000 samples from VQAV2 validation set for testing.

Implementation Details For a fair comparison with previous works (**????**), we employ a frozen Vit-B/32 CLIP model. The adaptor decoder consists of one layer Transformer Decoder (**?**) with 8 attention heads. For text reconstruction training, we set the noise variance σ to 0.016 as suggested in (**?**), and the region concept feature length N_{cr} is set to 10. In caption generation, the sampling number S in inference is set to 20. The text generation strategy is beam search with 4 beams. For the language model, we adopt a frozen pre-trained OPT (**?**) 1.3b model. Our model and the reproduced baseline are trained with a batch size of 128 and a learning rate of 4e-4.

Baselines The following zero-shot captioning methods are compared in this study. **ZeroCap** leverages CLIP and GPT-2 to solve an optimization problem iteratively during inference. **DeCap** (**?**) utilizes a memory bank to project image embeddings into the text embedding space. We sample 50M texts in CC3M (**?**) datasets to generate the memory bank for DeCap. Additionally, we define **Baseline** as the model trained in the same way as DeCap but using image embedding directly in inference. **MAGIC** (**?**) incorporates a CLIP-induced score during inference to influence the language model's caption generation process. We show the performance of MAGIC with a fine-tuned language model as reported in (**?**) and the performance of MAGIC with a frozen pre-trained language model. **CLIPRe** is a retrieval-based baseline mentioned in (**?**). **CapDec** (**?**) applies noise injection strategy in text reconstruction training, enabling the direct use of visual embedding in inference. We use the MLP variant of CapDec model. **SM** (**?**) is a modular framework in which multiple pre-trained models may be composed zero-shot. It uses the GPT-3 (**?**) API from OpenAI and achieves favorable performance.

5.2 Zero-Shot Cross Domain Captioning

In this section, we present a comprehensive evaluation of our method in the zero-shot cross-domain captioning setting. The zero-shot means our model is trained with only text data and the cross-domain means the training caption texts are different from captions in the test dataset. The CC3M is web-scale noisy captions data while MSCOCO and Flickr30K are human-annotated high-quality caption data. We compare our approach with previous methods on three different cross-domain settings to assess its performance and generalizability. We show our cross-domain image captioning results in Table **??**. We have observed a positive correlation between

Method	CC3M to MS-COCO						MS-COCO to Flickr30k						Flickr30k to MS-COCO					
	B@1	B@4	M	R_L	C	S	B@1	B@4	M	R_L	C	S	B@1	B@4	M	R_L	C	S
CLIPRe	-	0.046	0.133	-	0.256	0.092	0.387	0.044	0.096	0.272	0.059	0.042	0.414	0.052	0.125	0.307	0.183	0.057
ZeroCap	-	0.026	0.115	-	0.146	0.055	-	-	-	-	-	-	-	-	-	-	-	-
MAGIC	-	-	-	-	-	-	0.464	0.062	0.122	0.313	0.175	-	0.414	0.052	0.125	0.307	0.183	-
CapDec	-	-	-	-	-	-	0.602	0.173	0.186	0.427	0.357	-	0.433	0.092	0.163	0.367	0.273	-
DeCap	-	0.088	0.160	-	0.421	0.109	-	0.163	0.179	-	0.357	0.111	-	0.121	0.180	-	0.444	0.109
Frozen Language Model																		
Baseline	0.318	0.034	0.094	0.221	0.101	0.046	0.429	0.072	0.126	0.305	0.140	0.067	0.375	0.049	0.118	0.296	0.112	0.055
MAGIC†	0.188	0.004	0.054	0.142	0.021	0.011	0.188	0.006	0.051	0.134	0.021	0.013	0.188	0.004	0.054	0.142	0.021	0.011
CapDec†	0.372	0.046	0.116	0.288	0.093	0.052	0.490	0.105	0.153	0.373	0.183	0.090	0.453	0.087	0.151	0.353	0.178	0.064
DeCap†	0.462	0.089	0.152	0.341	0.292	0.093	0.511	0.099	0.153	0.362	0.247	0.087	0.455	0.088	0.162	0.360	0.273	**0.101**
MacCap	**0.591**	**0.176**	**0.200**	**0.443**	**0.525**	**0.120**	**0.595**	**0.154**	**0.179**	**0.413**	**0.303**	**0.114**	**0.473**	**0.092**	**0.166**	**0.362**	**0.278**	0.092

Table 2: **Zero-shot Cross Domain Captioning**: We conduct experiments on cross-domain image captioning tasks. X to Y means source to target domain. We reproduce Magic, CapDec, and DeCap under the frozen LLM setting and mark them with †.

Method	Data			MSCOCO						Flickr30k					
	P.	I.	T.	B@1	B@4	M	R_L	C	S	B@1	B@4	M	R_L	C	S
CLIPCap (**?**)	✓			-	0.335	0.275	-	1.131	0.211	-	-	-	-	-	-
CLIP-VL (**?**)	✓			-	0.375	0.281	-	1.231	0.219	-	-	-	-	-	-
UVC-VI (**?**)	✓			-	0.220	0.214	-	0.723		-	-	-	-	-	-
Barraco et al. (**?**)	✓			-	0.360	0.278	-	1.149	0.208	-	-	-	-	-	-
ESPER-Style (**?**)		✓	✓	-	0.219	0.219	-	0.782		-	-	-	-	-	-
ESPER-Free (**?**)		✓		-	0.063	0.133	-	0.291		-	-	-	-	-	-
ZeroCap⋆(**?**)			✓	0.498	0.007	0.154	0.318	0.345	0.092	0.447	0.054	0.118	0.273	0.168	0.062
CLIPRe (**?**)			✓	0.395	0.049	0.114	0.290	0.136	0.053	0.385	0.052	0.116	0.276	0.100	0.057
MAGIC (**?**)			✓	0.568	0.129	0.174	0.399	0.493	0.113	0.445	0.064	0.131	0.316	0.204	0.071
CapDec (**?**)			✓	0.692	0.264	0.251	0.518	0.918	-	0.555	0.177	0.200	0.439	0.391	-
DeCap (**?**)			✓	-	0.247	0.250	-	0.912	0.187	-	0.212	0.218	-	0.567	0.152
Frozen Language Model															
Baseline			✓	0.414	0.069	0.141	0.317	0.221	0.079	0.418	0.069	0.127	0.308	0.136	0.070
MAGIC†(**?**)			✓	0.188	0.004	0.054	0.042	0.021	0.011	0.188	0.006	0.051	0.134	0.021	0.013
CapDec†(**?**)			✓	0.537	0.156	0.206	0.435	0.465	0.134	0.429	0.072	0.136	0.336	0.127	0.054
DeCap†(**?**)			✓	0.531	0.125	0.188	0.403	0.427	0.126	0.485	0.096	0.143	0.351	0.213	0.079
MacCap			✓	**0.614**	**0.174**	**0.223**	**0.459**	**0.697**	**0.157**	**0.564**	**0.153**	**0.189**	**0.414**	**0.358**	**0.124**

Table 3: **Zero-shot In Domain Captioning**: The notation "P.", "I.", and "T." are used to represent paired data, unpaired image data, and unpaired text data, respectively. We reproduce Magic, CapDec, and DeCap under frozen language model setting and mark them with †. Results tagged ⋆ are from (**?**)

Method	VQAV2 Val (%)			MSCOCO		
	Top1	Top5	Top10	R_L	C	S
Finetuned Language Model						
CapDec†	0.86	3.13	3.71	**0.518**	**0.918**	-
DeCap†	4.09	10.89	14.33	-	0.912	**0.187**
Frozen Language Model						
Baseline	3.26	7.24	11.21	-	-	-
CapDec	6.53	11.06	15.00	0.435	0.465	0.134
DeCap	6.00	11.81	15.57	0.403	0.427	0.126
MacCap	**7.96**	**14.00**	**18.72**	0.459	0.697	0.157

Table 4: **Zero-shot VQA results** on VQAV2 validation set. † means the models come from the official release. Our method achieves superior performance under a frozen language model setting.

Method	Flickr30K					
	$B@1$	$B@4$	M	R_L	C	S
Text Reconstruction Training						
single token w/o noise	0.396	0.055	0.123	0.277	0.147	0.070
single token w/ noise	0.445	0.082	0.110	0.282	0.143	0.065
multiple token w/o noise	0.388	0.049	0.122	0.277	0.135	0.060
multiple token w/ noise	0.520	0.107	0.153	0.369	0.197	0.084
Zero-shot Caption Generation						
CLS token	0.493	0.096	0.134	0.350	0.140	0.064
subregion aggregation	0.520	0.107	0.153	0.369	0.197	0.084
sampling and filtering	**0.542**	**0.130**	**0.152**	**0.378**	**0.199**	**0.078**

Table 5: **Ablation Results** on Flickr30K datasets. We evaluate the effectiveness of our training and inference paradigms.

the size of the training text corpus and the performance gain of MacCap. Our method achieves superiority in both domains in most metrics under our frozen language model setting.

5.3 Zero-Shot In Domain Image Captioning

Setting In this section, we conduct zero-shot in-domain image captioning experiments, where the models are trained and tested on the same domain. We compare our method with other supervised methods, unpaired image captioning methods, and text-only training methods.

Results We show our results on MSCOCO (**?**) and Flickr30K (**?**) in Table **??**. Our method outperforms other methods on both domains in all metrics under our frozen language model setting and shows a gain of +27 in CIDEr compared with DeCap (**?**). We also achieve higher performance compared with fully supervised methods such as Laina et al. (**?**) and Feng et al (**?**) in terms of CIDEr on MSCOCO.

5.4 Zero-Shot Visual Question Answering

In this section, we conduct zero-shot visual question-answering experiments. Due to the redundancy of the large language model generation results, we do not use the VQAV2 traditional evaluation metric where the predicted answer should be the same as the ground truth answer but instead use an image-text retrieval task where the answer candidates are provided. We report the top 1,5 and 10 accuracies in % for 20,000 VQAV2 validation set samples. Compared with previous zero-shot captioning methods (**??**), including a baseline where the language model generates answers solely based on the question without a caption, our results are presented in Table **??**. Our method outperforms other methods with and without a frozen language model. We observe a performance degradation between the frozen language model and the finetuned language model, which indicates that the finetune language model on zero-shot captioning model harms model performance on other tasks.

5.5 Ablation Study

In this section, we conduct an ablation study to provide a comprehensive interpretation of our proposed methods. The experiments are conducted under zero-shot cross-domain image captioning settings where the model is trained on CC3M text corpus and evaluated on Flickr30K dataset.

Text Reconstruction Training To validate the effectiveness of the region noise design in our framework, we conducted experiments to determine whether the observed improvements were due to the noise injection or the sequential representation of the text. we modify the input text feature of the adaptor decoder module, where we define two modes: *single token* and *multiple token*. In the *single token* mode, a single text embedding T_c is provided as input to the adaptor decoder. In contrast, in the *Multiple tokens* mode, multiple text embeddings are used as input for the adaptor decoder. The *w/ noise* or *w/o noise* indicate whether adding noise to the embedding. We present the results in Table **??**. Based on the results, we can conclude that the strong performance of our approach can be attributed to the effective combination of noise injection and sequential representation.

Zero-shot Caption Generation We conduct an ablation study based on the model trained with region noise. We modified the input image feature of the adaptor decoder module. We defined two modes: *CLS token* indicate I_{cr} doesn't contain subregion features I_s and *subregion aggregation* indicate I_{cr} is the sum of I_s and I_c. Furthermore, we incorporated the *sampling and filtering* strategy. The obtained outcomes are illustrated in Table **??**. Our observations reveal that the adoption of the *sampling and filtering* approach led to a noteworthy improvement in the BLEU metric, signifying the rectification of erroneous instances within the generated captions. However, its impact on semantic comprehension and contextual coherence was relatively modest in comparison.

6 Conclusion

We present an in-depth analysis of the *modality gap* phenomenon in the CLIP latent space, uncovering two key phenomena: tighter proximity of CLIP visual features within image subregions to paired captions and a modality gap adhering to a zero-mean Gaussian distribution. In response to these insights, we introduced a novel approach—region noise-injected text reconstruction training. Leveraging subregion feature aggregation in zero-shot caption generation,

we harnessed subregion information to create a closer visual representation with paired caption representation. Additionally, we propose an inference-time noise and CLIP reranking to further boost performance. Experimental results demonstrate that MacCap outperforms state-of-the-art methods.

1 Detail of Baselines

In this section, we provide a detailed explanation for the setting difference between our paper and previous methods DeCap(**?**), CapDec(**?**). The key difference is whether the language model or text decoder is trainable. DeCap and CapDec train the language model and achieve better performance in zero-shot in-domain captioning. However, we propose a more practical setting where the language model is frozen. The reason is that the large language model undergoes training through an auto-regression objective, thereby acquiring the capacity to proficiently execute a multitude of tasks, such as questing answering, translation, and automatic summarization. Training language model for a specific task leads to a degradation in other tasks, which is demonstrated in our zero-shot VQA experiments. Besides, the impressive performance of large language models comes with scale which makes finetuning language models computationally expensive. To provide a comprehensive picture, we collect the model parameters information in Table 1. We observe that the better performance of CapDec and DeCap in zero-shot in-domain captioning is achieved with more parameters, however, MacCap outperforms CapDec and DeCap in zero-shot cross-domain captioning with fewer parameters.

2 Ablation Study on Hyperparameters

Noise Variance The noise injection is utilized to reduce the widely observed *modality gap* phenomenon in CLIP embedding space. We show the impact of noise variance when generating the text-region feature mentioned in Section 4.2 of the main paper. The results are shown in Figure **??**. We observed the best performance is achieved at noise variance 0.016 which is the same as suggested in (**?**). Noise variances smaller than 0.016 cause performance drop due to the *modality gap* while noise variances larger than 0.016 introduce extensive semantic ambiguity in text reconstruction training.

Number of Image Patches The sub-region feature aggregation in Section 4.3 of the main paper aims to extract the features of image subregions. We select image patches based on their attention scores as the patches with higher scores tend to contain more semantic information. We empirically study the number of selected patches, which indicates how much background information is introduced when generating captions. The results are shown in Figure **??**. We observe the best performance is achieved when the number of selected patches is the same as the number of noise-injected text-region features.

3 Visualization of Image Sub-region

In this section, we visualize the distribution of the global image embeddings, local image embeddings (i.e. the embeddings of image sub-region), and the text embeddings. We show the relation between image subregions and corresponding captions in Section 3.1 of the main paper, which reveals that subregion embeddings may have a higher similarity with the caption text embedding as they can be the specific image regions described by the accompanying caption. We visualize 5000 samples from the MSCOCO dataset with the dimensionality reduction method UMAP following (**?**). The results

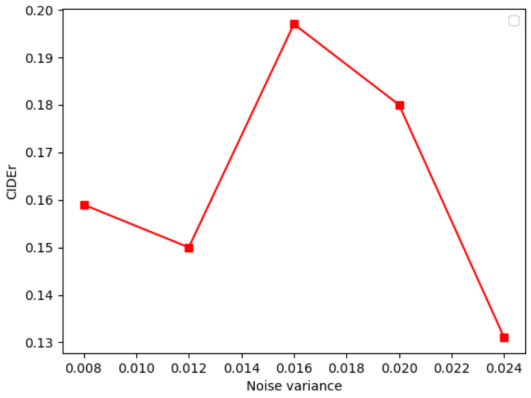

Figure 4: **Performance of MacCap with different training noise.** The MacCap is trained in CC3M dataset and tested on Flickr30K datasets.

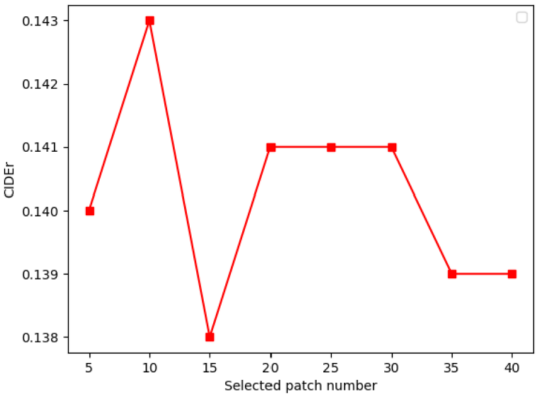

Figure 5: **Performance of MacCap with different patch numbers in inference.** The MacCap is trained in CC3M dataset and tested on Flickr30K datasets. The length of the text region feature in text reconstruction training is set to 10.

are shown in Figure **??**. We observe the distances between text embeddings are relatively small and all of the text embeddings are distributed near the (0, 0). Furthermore, there exists a modality gap between global image embedding and text embedding. The local image embeddings are distributed in the surrounding region of global image embeddings and are relatively closer to text embeddings than the global ones.

4 Details of Modality Gap Distribution Analysis

In this section, we provide implementation details of modality gap distribution experiments.

We have text modality representation $T^i \in \mathbb{R}^D$ and vision modality representation. To be noticed, for vision modality, we have two sets of representations: the global embedding representing the overall information and the patch embeddings representing the image subregions information. The global embedding is $I_c^i \in \mathbb{R}^D$, and the patch embedding is $I_p^i \in \mathbb{R}^{N \times D}$. Both global embedding and local embedding are obtained by CLIP. D is the dimension of CLIP

	# Trainable Parameter	# Frozen Parameter	CC3M to MSCOCO(CIDEr)	MSCOCO(CIDEr)	Support Memory
CapDec	181M	0	-	0.918	0
DeCap	68M	0	0.421	0.912	50,000
Reproduced CapDec	1.5M	1.3B	0.093	0.465	0
Reproduced DeCap	1.5M	1.3B	0.292	0.427	50,000
MacCap	5.7M	1.3B	0.525	0.687	0

Table 6: The parameter size and captioning performance of baseline methods. We display zero-shot cross-domain and zero-shot in-domain captioning results from the main paper. The support memory is only used by DeCap, where 50,000 texts are used to construct a memory bank for inference.

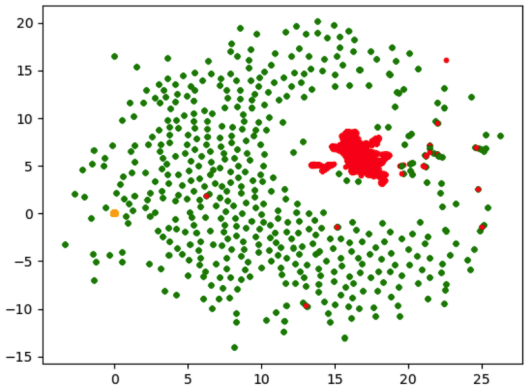

Figure 6: **Visualization of Embedding Distributions on MSCOCO.** The red points stand the global image embedding, each green point stand a local image embedding, and the yellow points near (0, 0) stand for the text embedding.

embedding and N is the sequence length of CLIP. $i \in \{1, 2, ..., num\_samples\}$

For paired images and text descriptions, we evaluate the gap between text representation and both global and patch image representation. For each image-text pair, we compute $G_i^c = T^i - I_c^i$ and $G_i^p = repeat(T^i, N) - I_p^i$. $G_i^c \in \mathbb{R}^D$, $G_i^p \in \mathbb{R}^{N \times D}$.

We can check the overall distribution over all D dimensions. In this case, we treat data in all dimensions equally. We compute the mean over all image-text pairs and draw the histogram of the data distribution. Thus we compute the mean of the gap distribution for global vision and language representation as $Avg_{(i,d)}(I_c^i[d])$, and the gap for local vision and language representation as $Avg_{(i,s,d)}(I_p^i[s][d])$. $s \in [N], d \in [D]$.