Bridging Semantic and Modality Gaps in Zero-Shot Captioning via Retrieval from Synthetic Data

Zhiyue Liu[1,2]*, **Wenkai Zhou**[1]
[1]School of Computer, Electronics and Information, Guangxi University, Nanning, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology
liuzhy@gxu.edu.cn, 2413394045@st.gxu.edu.cn

Abstract

Zero-shot image captioning, which aims to generate image descriptions without relying on annotated data, has recently attracted increasing research interest. Pre-trained text-to-image generation models enable the creation of synthetic pairs solely from text data, while existing methods fall short in mitigating the discrepancy caused by the inability of synthetic images to fully capture the semantics of the textual input, resulting in unreliable cross-modal correspondences. To address this, we propose a retrieval-based framework that leverages only existing synthetic image-text pairs as its search corpus to systematically bridge the gap when using synthetic data for captioning. For the semantic gap between a synthetic image and its input text, our framework retrieves supplementary visual features from similar synthetic examples and integrates them to refine the image embedding. Then, it extracts image-related textual descriptions to mitigate the modality gap during decoding. Moreover, we introduce a plug-and-play visual semantic module that detects visual entities, further facilitating the construction of semantic correspondences between images and text. Experimental results on benchmark datasets demonstrate that our method obtains state-of-the-art results.

1 Introduction

Supervised image captioning approaches rely on large-scale labeled image-text pairs, which are often costly to acquire (Zhu et al., 2023). To address this data dependency limitation, zero-shot image captioning approaches have attracted growing research interest. Previous methods (Li et al., 2023) adopt text-only training and subsequently perform inference using images. However, the mismatch between training and inference leads to a modality gap (Liang et al., 2022), which hinders the model's performance. Most existing works use pre-trained multi-modal models (Radford et al., 2021) to map texts and images into the same space, which mitigates the modality gap (Fei et al., 2023; Lee et al., 2024). Since using text as a proxy for visual information cannot fully capture visual semantics, text-only methods, which cannot utilize images during training, still maintain a gap between modalities.

To address the above problem, many recent studies have utilized image generative models to construct synthetic image-text pairs (Ma et al., 2024; Liu et al., 2024), using synthetic visual semantics for training. Although these pre-trained generative models could produce rich visual information, they suffer from information loss during both the text-to-vector (Kamath et al., 2023) encoding process and the image generation stage (Ray et al., 2023; Sariyildiz et al., 2023) due to inherent model limitations. This results in a semantic mismatch between the synthetic image and its corresponding text, commonly referred to as the semantic gap (Luo et al., 2024). Some straightforward strategies (e.g., replacing images based on similarity) could alleviate the semantic gap (Liu et al., 2024, 2025), but existing works heavily rely on the capabilities of image generative models and directly utilize synthetic image-text pairs with semantic mismatches, as illustrated by Figure 1(a). As a result, they fail to handle the discrepancies arising from the inability of synthetic images to fully capture the semantics of textual inputs, resulting in unreliable cross-modal correspondences.

This paper proposes a method, called ROSCap, which mitigates both the semantic and modality gaps in synthetic image-text pairs using existing information and reduces reliance on image generative models. Since synthetic images are generated based on text, the corresponding original text preserves complete semantic information. We retrieve embeddings of synthetic images using the original text and construct a support domain to mitigate gaps in image-text pairs, thereby improving the se-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14010–14023

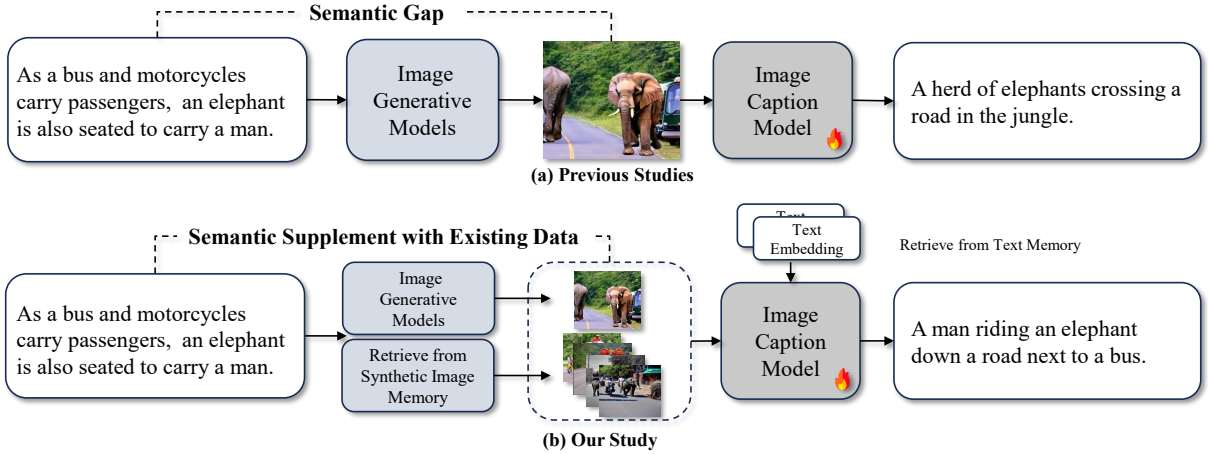


Figure 1: Comparison between our method and previous studies. (a) shows that previous work trains directly on synthetic data while neglecting the gap between the synthetic image and text. (b) shows that our method utilizes existing visual and textual information to mitigate this semantic gap. The reconstructed outputs on the right side demonstrate that our method enables the model to learn image-text correspondence patterns more consistent with the semantics of the conditional text.

mantic consistency within synthetic pairs, as shown by Figure 1(b). Leveraging semantically enhanced optimization for synthetic image embeddings not only fully exploits the existing information, but also supplements missing semantic content through retrieval, effectively addressing the semantic gap in synthetic image-text pairs. Besides, considering the issue of modality gap in images during decoding, we utilize optimized synthetic image embeddings to retrieve semantically similar text for modality fusion. The cross-modal fusion of synthetic image and text embeddings reduces the modality gap and further bridges the semantic gap by leveraging semantically aligned textual information. Accurate visual prompts could effectively guide the model in establishing associations between image and text. To this end, we construct a visual semantic module that jointly retrieves text and synthetic images, applies a filtering mechanism, and extracts entity information across multiple modalities, thereby enabling the construction of precise visual prompts.

In summary, our contributions are as follows:

- We propose ROSCap that optimizes synthetic image embeddings and retrieves textual information for fusion, aiming to bridge the semantic gap and construct generalizable image-text correspondences from synthetic data.
- A plug-and-play visual semantic module is designed to extract accurate visual entities from both modalities for prompt construction.
- Extensive experiments on several benchmarks demonstrate that ROSCap achieves state-of-the-art performance for zero-shot captioning.

2 Related work

2.1 Supervised Image Captioning

Supervised image captioning relies on training encoders and decoders using annotated image-text pairs. Traditional research employed convolutional neural networks as image encoders and recurrent neural networks as decoders to implement captioning (Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Gu et al., 2017). The subsequent emergence of transformer architectures and attention mechanisms (Vaswani et al., 2017) has further advanced the field, offering greater potential for improving image captioning performance (Cornia et al., 2020; Pan et al., 2020; Yang et al., 2021; Barraco et al., 2022). Recent text retrieval-based approaches have attracted increasing attention (Ramos et al., 2023; Kim et al., 2025), with images serving as queries to retrieve semantically relevant texts, thereby improving performance and narrowing the modality gap between image and text during decoding. While these methods exhibit competitive performance, they are constrained by the dependence on manually annotated image-text pairs, whose collection is both time-consuming and costly.

2.2 Zero-Shot Image Captioning

Zero-shot image captioning could be generally categorized into training-free approaches and text-only ones. Training-free methods (Tewel et al., 2022;

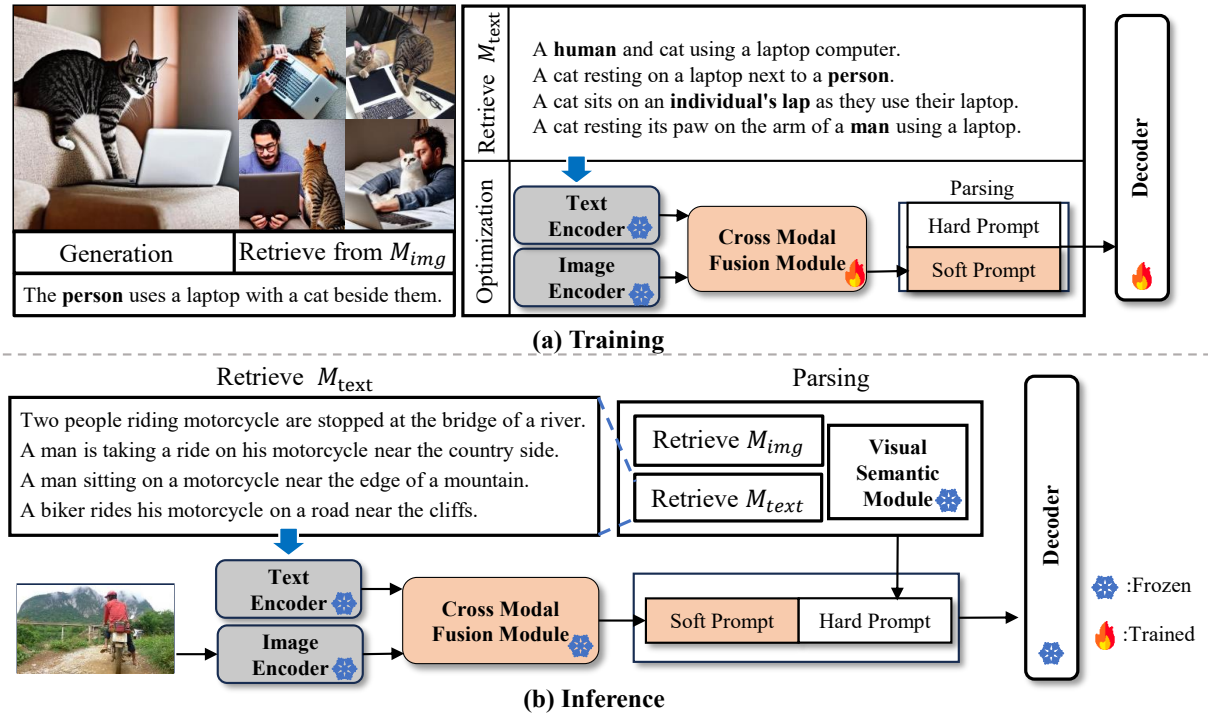


Figure 2: The overview of ROSCap. During training, text embeddings are first used to retrieve the most relevant synthetic images, which are then employed to optimize the corresponding synthetic image embeddings. The optimized embeddings are subsequently utilized to retrieve relevant texts, thereby contributing to bridging the modality gap. During inference, the retrieved texts and synthetic images are processed by the visual semantic module to extract salient entities and construct a hard prompt, which is finally fed into the decoder.

Zeng et al., 2024) rely on CLIP (Radford et al., 2021) zero-shot classification capabilities, limiting their ability to capture fine-grained visual details. Recent text-only methods demonstrate potential for zero-shot learning (Fei et al., 2023; Lee et al., 2024; Yang et al., 2023, 2024), which map images and texts into a shared space. However, utilizing textual data for training while only visual inputs for inference would induce a modality gap. This originates from the geometric constraints in the image-text mapping network, where the embedding space inherently assumes a narrow conical structure. Thus, a separation between modalities exists from the outset (Liang et al., 2022).

Recent advances attempt to use synthetic images instead of textual data to mitigate the modality gap. SynTIC (Liu et al., 2024) explored training models with synthetic images and used contrast learning to reduce the feature distance between synthetic and real images. Due to the limitations of image generative models (Kamath et al., 2023; Sariyildiz et al., 2023), there may exist semantic inconsistencies between synthetic images and conditional texts. PCM-Net (Luo et al., 2024) proposed enhancing semantic alignment by selectively fusing key visual concepts from synthetic images with text embeddings. SaCap (Liu et al., 2025) reconstructed original texts and re-generated multiple synthetic images for matching. However, these methods rely on the capacity of image generation and directly utilize synthetic pairs that exhibit the semantic gap. In summary, the discrepancy between image and text embeddings brings the modality gap, whereas the semantic gap emerges when synthetic images fail to faithfully capture the semantics of associated texts. To address these, our paper introduces a retrieval-based framework that reduces reliance on image generative models and effectively exploits existing data to bridge both the semantic and modality gaps.

3 Methodology

Our proposed zero-shot image captioning method, ROSCap, is based on the retrieval of semantic information. The overall framework is illustrated in Figure 2. During the training process, we leverage only text and text-based synthetic images. First, in the selective projection optimization method (Section 3.1), a support domain is constructed by retrieving synthetic image embeddings. The original synthetic image embeddings are then projected into

this domain to achieve feature optimization. This process not only preserves the original information of the synthetic images but also introduces additional semantic information to help alleviate the semantic gap between the synthetic image-text pairs. These optimized synthetic image embeddings are used to retrieve the most relevant text embeddings from the text memory, which are then fed into the cross-modal fusion module (Section 3.2) for feature fusion. It addresses the modality gap between image-text pairs and further bridges the semantic gap. During inference, to capture more accurate entity information, we employ the visual semantic module (Section 3.3) to retrieve data simultaneously from both the synthetic image memory and the text memory, enabling the extraction of entity information from multiple modalities.

3.1 Selective Projection Optimization

The textual semantics could be translated into a corresponding image using stable diffusion (Rombach et al., 2022), thereby obtaining the synthetic image-text pairs (I, T). At the same time, the synthetic images and texts are encoded into embeddings using an encoder, forming the synthetic image memory M_{img} and text memory M_{text}. This process could be represented as follows:

$$M_{\text{img}} = \text{encoder}_{\text{img}}(I_1, I_2, \ldots, I_N), \quad (1)$$

$$M_{\text{text}} = \text{encoder}_{\text{text}}(T_1, T_2, \ldots, T_N). \quad (2)$$

Considering the semantic gap in image-text pairs, we construct an approach that leverages the powerful cross-modal capabilities of CLIP. Specifically, we calculate the cosine similarity between the original text embedding T_t and all synthetic image embeddings in M_{img}, and select the top-k synthetic image embeddings with the highest similarity. These synthetic image embeddings would constitute the synthetic image support domain, denoted as support set $(I_{t1}, I_{t2}, \ldots, I_{tk})$ which is used to optimize the embedding of the original synthetic image.

To prevent semantic redundancy and preserve the fine-grained characteristics embedded in the original synthetic images, we optimize only those image–text pairs exhibiting a semantic gap, namely when the text fails to retrieve its corresponding synthetic image within the support set, an indication of low consistency between the synthetic image and the associated text. We need to ascertain whether the synthetic image domain support set S_u, constructed through the retrieval of T_t, includes the image I_t. When the embedding of a synthetic image is absent from the support domain retrieved using its associated text embedding, we classify the image-text pair as exhibiting a semantic gap. To mitigate this discrepancy, we perform projection-based optimization by aligning the image embedding with the support domain, thereby improving its semantic consistency with the corresponding text. This process could be expressed as follows:

$$S_u = \text{argtop}_k\left(\cos\left(T_t, M_{\text{img}}\right)\right), \quad (3)$$

$$I'_t = \begin{cases} \left(\frac{\exp\left((I_t S_u^\top)/\tau\right)}{\sum_j \exp\left((I_t S_{u_j}^\top)/\tau\right)}\right) S_u & \text{if } I_t \notin S_u, \\ I_t & \text{else.} \end{cases} \quad (4)$$

By constructing an in-domain synthetic image set through the retrieval of synthetic images most semantically similar to the original text from existing data, we could strengthen the alignment between synthetic images and their corresponding captions, thereby enhancing the model's capacity to learn accurate cross-modal mappings.

3.2 Cross-Modal Fusion Module

The use of synthetic image embeddings inevitably introduces semantic information loss during the embedding compression process. Also, the modality gap and semantic gap occur during the decoding of synthetic images into text. In order to alleviate these issues, we propose a cross-modal fusion module designed to bridge both the modality and semantic gaps while enhancing the richness of information representation.

Specifically, we first use the optimized synthetic image embedding to retrieve the l most similar text embeddings from the constructed M_{text}, based on the cosine similarity. The retrieved text embeddings are fused with the optimized synthetic image embedding to supplement semantic information and further alleviate both the semantic and modality gaps. A cross-attention mechanism is then employed to obtain T_{att}, where the optimized image embedding serves as the query and the retrieved text embeddings serve as the keys and values. The overall process could be formulated as follows:

$$\begin{aligned} \text{Text}_t &= \{T_{t1}, T_{t2}, \ldots, T_{tL}\} \\ &= \text{argtop}_l\big(\cos(I'_t, M_{\text{text}})\big), \end{aligned} \quad (5)$$

$$T_{att} = \text{crossattn}(I'_t, \text{Text}_t). \quad (6)$$

The fused representation is subsequently concatenated with the learnable vector θ and fed into a

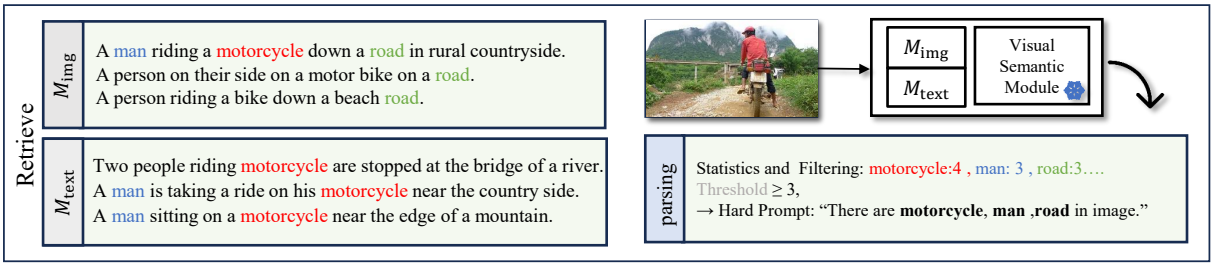


Figure 3: Our visual semantic module. This module facilitates the extraction of enriched semantic information across multiple modalities, enabling more precise identification of visual objects. By incorporating hard prompts, it establishes accurate and robust associations between images and their corresponding textual descriptions.

trainable 8-layer transformer to learn the image-text mapping. Our cross-modal fusion module integrates a mapping network composed of a cross-attention mechanism and a transformer:

$$\text{prompt}_{\text{soft}} = \text{transformer}(\text{concat}(T_{att}, \theta)), \tag{7}$$

where $\text{prompt}_{\text{soft}}$ would combine optimized embeddings with textual information, bridging both the modality and semantic gaps for decoding.

3.3 Visual Semantic Module

Previous studies have primarily focused on retrieving text to extract visual entities, but reliance solely on textual information can lead to incomplete or potentially misleading retrieval results. To overcome this limitation, we propose a visual semantic module to capture more comprehensive and fine-grained visual information, thereby improving the image-to-text mapping.

Our visual semantic module facilitates the construction of multi-view associations between images and text by integrating both visual and textual information to extract representative entities. Importantly, the module is applied exclusively during inference, following a plug-and-play paradigm that allows seamless integration into existing image captioning frameworks. As shown in Table 9, incorporating this module yields substantial performance gains across multiple baseline models.

The module encodes a real image using the CLIP image encoder. The resulting real image embedding I_{rt} is then used to retrieve the top-l text embeddings and synthetic image embeddings from the pre-constructed sets M_{text} and M_{img}:

$$\begin{aligned}\text{Text}_r = &\ \text{argtop}_l\left(\cos(I_{rt}, M_{\text{text}})\right) \\ &+ \text{argtop}_l\left(\cos(I_{rt}, M_{\text{img}})\right),\end{aligned} \tag{8}$$

where retrieved embeddings Text_r consist of both text and image embeddings, from which text labels are extracted. Entities are subsequently extracted from the retrieved text labels using a syntactic parsing tool, and those with frequencies exceeding a pre-defined threshold d are selected to construct entity-based hard prompts, employed to guide the caption generation:

$$\text{prompt}_{\text{hard}} = \text{parsing}(\text{Text}_r). \tag{9}$$

Figure 3 illustrates an example processed by our visual semantic module. Relevant semantics are retrieved from M_{text} to identify entities such as "people" and "motorcycles" while M_{img} provides additional visual cues, including contextual elements like "roads" and enhances the separation between target entities and irrelevant information. This multi-modal retrieval strategy facilitates the extraction of richer semantic information, thereby improving the accuracy of visual object recognition. By leveraging hard prompts, the model establishes precise image–text correspondences. Entities are extracted from the retrieved texts and their frequency statistics are computed using the syntactic parser tool NLTK (Bird and Loper, 2004).

The constructed hard prompts are concatenated with soft prompts and fed into the decoder to generate the final captions. The cross-modal fusion module and the decoder are trained jointly under an auto-regressive loss, formally defined as follows:

$$L = -\sum_{t=1}^{T} M_t \log p_\theta(y_t \mid \text{prompt}, y_{<t}), \tag{10}$$

where prompt denotes the concatenation of two prompts $(\text{prompt}_{\text{hard}}, \text{prompt}_{\text{soft}})$, and y denotes the ground truth text.

4 Experiments

Datasets and Metrics. The experiments are conducted on three benchmark datasets, consisting of

Method	Image Encoder	Text Decoder	MSCOCO				Flickr			
			B@4	M	C	S	B@4	M	C	S
CapDec (2022)	RN50x4	GPT-2_{large}	26.4	25.1	91.8	11.9	17.7	20.0	39.1	9.9
DeCap (2023)	ViT-B/32	Transformer$_{L=4,\ H=4}$	24.7	25.0	91.2	18.7	21.2	21.8	56.7	15.2
CLOSE (2022)	ViT-L/14	T5$_{base}$	–	–	95.3	–	–	–	–	–
ViECap (2023)	ViT-B/32	GPT-2_{base}	27.2	24.8	92.9	18.2	21.4	20.1	47.9	13.6
MeaCap$_{InvLM}$ (2024)	ViT-B/32	GPT-2_{base}	27.2	25.3	95.4	19.0	22.3	22.3	59.4	15.6
IFCap (2024)	ViT-B/32	GPT-2_{base}	30.8	26.7	108.0	20.3	23.5	23.0	64.4	17.0
SynTIC (2023)	ViT-B/32	Transformer$_{L=4,\ H=4}$	29.9	25.8	101.1	19.3	22.3	22.4	56.6	16.6
PCM-Net (2024)	ViT-B/32	GPT-2_{base}	**31.5**	25.9	103.8	19.7	26.9	23.0	61.3	16.8
SaCap (2025)	ViT-B/32	GPT-2_{base}	31.0	25.9	104.2	19.7	**27.1**	22.4	64.3	15.9
ROSCap (Ours)	ViT-B/32	GPT-2_{base}	31.3	**26.9**	**112.0**	**20.9**	23.4	**23.1**	**65.6**	**17.6**

Table 1: In-domain captioning results on the MSCOCO and Flickr test splits. All results of baselines are directly cited from their original papers. The upper block of the table presents the performance of text-only approaches, while the lower block reports the results using synthetic images. The best results are highlighted in bold.

Method	MSCOCO ⇒ Flickr				Flickr ⇒ MSCOCO			
	B@4	M	C	S	B@4	M	C	S
DeCap (2023)	16.3	17.9	35.7	11.1	12.1	18.0	44.4	10.9
ViECap (2023)	17.4	18.0	38.4	11.2	12.6	19.3	54.2	12.5
SynTIC (2023)	17.9	18.6	38.4	11.9	14.6	19.4	47.7	12.5
SynTIC-TT (2023)	19.4	20.2	43.2	13.9	20.6	21.3	64.4	14.3
IFCap (2024)	17.8	19.4	47.5	12.7	14.7	20.4	60.7	13.6
IFCap-TT (2024)	21.2	21.8	59.2	15.6	19.0	23.0	76.3	17.3
PCM-Net (2024)	20.8	19.2	45.5	12.9	17.1	19.6	54.9	12.8
SaCap (2025)	21.2	20.2	50.9	13.2	17.5	19.7	62.5	13.5
ROSCap	**21.3**	21.1	59.9	14.8	15.7	19.5	62.5	13.6
ROSCap-TT	20.7	**21.8**	**61.4**	**16.0**	**23.5**	**23.7**	**88.6**	**18.7**

Table 2: Cross-domain captioning results. X⇒Y means source domain ⇒ target domain. TT denotes that the model could access the target domain's corpus during inference. Since SaCap uses additional out-of-domain data for cross-domain experiments, we conduct experiments under the TT setting.

Method	In		Near		Out		Entire	
	C	S	C	S	C	S	C	S
ClipCap (2021)	79.7	12.2	67.6	11.2	49.3	9.7	65.7	11.1
SmallCap (2023)	88.3	–	77.1	–	65.0	–	75.8	–
DeCap (2023)	65.2	–	47.8	–	25.8	–	45.9	–
CapDec (2022)	60.1	10.2	50.2	9.3	28.7	6.0	45.9	8.3
ViECap (2023)	66.0	10.4	64.3	9.9	65.0	8.6	66.2	9.5
PCM-ViECap (2023)	61.1	10.7	72.7	10.6	**75.7**	9.7	74.7	10.3
IFCap (2024)	70.1	11.2	72.5	10.9	72.1	9.6	74.0	10.5
IFCap+VSM (Ours)	73.2	11.8	74.3	11.2	72.2	9.7	75.5	10.8
ROSCap (MSCOCO)	**76.2**	**12.4**	**76.5**	**11.7**	68.3	9.7	75.8	**11.1**
ROSCap (CC3M)	73.6	11.5	76.3	11.5	74.1	**10.2**	**76.6**	**11.1**

Table 3: NoCaps validation segmentation results. ClipCap and SmallCap are supervised methods, while the others are zero-shot methods. In, Near, Out, and Entire represent the in-domain, near-domain, out-domain, and overall experimental results, respectively.

MSCOCO (Chen et al., 2015), Flickr (Young et al., 2014), and NoCaps (Agrawal et al., 2019). We follow Karpathy (Karpathy and Fei-Fei, 2015) to split MSCOCO and Flickr into training, validation and testing sets. We train the captioning model using only text annotations from the training set and synthetic images generated by image generative models. For evaluation, we adopt standard image captioning metrics, with CIDEr (**C**) (Vedantam et al., 2015), SPICE (**S**) (Anderson et al., 2016) as the primary evaluation metrics, and BLEU-4 (**B@4**) (Papineni et al., 2002) and METEOR (**M**) (Banerjee and Lavie, 2005) as supplementary metrics.

Implementation Details. We use stable diffusion v1-5 (Rombach et al., 2022) as the text-to-image model. We leverage the pre-trained multi-modal model CLIP (ViT-B/32) (Radford et al., 2021) as our encoder and GPT2$_{base}$ (Radford et al., 2019) as the decoder. During the training process, the encoder is frozen and only the cross-modal fusion module and the image decoder need to be trained. We train the captioning model with a learning rate of 2×10^{-5}, using the linear schedule as our scheduler with 5000 warmup steps. Adam (Kingma and Ba, 2014) is used as the optimizer, and the batch size is set to 80. Our source code is available at https://github.com/wkwinking/ROSCap.

Baselines. We compare the performance of our ROSCap with state-of-the-art works for zero-shot captioning, including text-only methods and the synthetic image methods. The text-only ones contain CapDec (Nukrai et al., 2022), DeCap (Li et al., 2023), CLOSE (Gu et al., 2023), ViECap (Fei et al., 2023), MeaCap (Zeng et al., 2024), and IFCap (Lee et al., 2024). The methods using synthetic images include SynTIC (Liu et al., 2024), PCM-Net (Luo

Method	B@4	M	C	S
Baseline	26.6	24.2	91.6	17.4
+selective projection optimization (SPO)	26.0	24.3	92.5	18.2
+visual semantic module (VSM)	28.5	25.7	102.1	19.2
+cross-modal fusion module (CMF)	28.4	25.3	104.5	19.5
+SPO & VSM w/o CMF	28.7	26.0	104.7	19.9
+CMF & VSM w/o SPO	29.3	26.2	108.2	20.5
+CMF & SPO w/o VSM	31.1	26.5	109.6	20.1
ROSCap	**31.3**	**26.9**	**112.0**	**20.9**

Table 4: Ablation study results on MSCOCO under the in-domain captioning setting.

k	B@4	M	C	S
3	30.2	26.6	109.7	20.6
5	**31.3**	**26.9**	**112.0**	**20.9**
7	30.4	26.6	110.3	20.6
9	30.2	26.6	109.9	20.6

Table 5: The performance of ROSCap with different synthetic image numbers k to build the support set.

et al., 2024), and SaCap (Liu et al., 2025).

4.1 In-Domain Captioning Results

We first evaluate the proposed ROSCap on both the MSCOCO and Flickr datasets under the in-domain captioning setting. Table 1 illustrates the performance comparison between our method and other state-of-the-art image captioning models. Our ROSCap consistently outperforms state-of-the-art methods across most evaluation metrics on both the MSCOCO and Flickr datasets. In particular, it achieves a CIDEr score of 112.0, representing an absolute improvement of 4.0 over IFCap on the same benchmark, thereby demonstrating that training with synthetic images can effectively mitigate the modality gap during both training and inference. Compared with PCM-Net, ROSCap further improves the CIDEr score by 8.2, highlighting the effectiveness of first optimizing synthetic images to reduce their discrepancy with the corresponding texts and subsequently employing a cross-modal fusion module to narrow both the modality and semantic gaps. Overall, our method achieves the best performance on the primary image captioning metrics, CIDEr and SPICE, underscoring its effectiveness for in-domain captioning.

4.2 Cross-Domain Captioning Results

We further evaluate the generalization ability of ROSCap under the cross-domain setting. Given that ROSCap is a retrieval-based image captioning method, we additionally consider an experimental setting in which the model has access to the target domain. The experimental results, which cover cross-domain evaluations between MSCOCO and Flickr in Table 2 as well as evaluations on the challenging NoCaps validation set in Table 3, demonstrate that ROSCap consistently achieves state-of-the-art results across the majority of evaluation metrics, while maintaining second-best performance on the remaining ones. Notably, in the MSCOCO⇒Flickr experiment, ROSCap achieves a CIDEr score of 59.9, which surpasses the previous state-of-the-art method with a score of 47.5 by 12.4 points. Moreover, it even outperforms the in-domain SynTIC method, which obtains a score of 56.6. For the experiments on NoCaps, ROSCap achieves state-of-the-art results on most metrics, and integrating our visual semantic module into IFCap leads to consistent improvements across all metrics. We also observe that performance varies depending on the type and size of memory. Taking CC3M (Sharma et al., 2018) as memory yields overall performance that even surpasses the retrieval-based supervised method SmallCap, highlighting the strong generalization capability of our method.

When encountering out-of-domain data in practical scenarios, retraining the model is unnecessary, since only the in-domain memory needs to be updated. Under this setting, ROSCap-TT achieves a CIDEr score of 88.6, representing an improvement of 12.3 points over the strongest result of baselines. In the MSCOCO⇒Flickr experiment, updating the memory with ROSCap-TT yields a CIDEr score of 61.4, surpassing the in-domain PCM-Net method. Both configurations, utilizing only the training set (i.e., ROSCap) and augmenting the memory with external data (i.e., ROSCap-TT), demonstrate the effectiveness and flexibility of our method.

4.3 Ablation Study

Effect of Components. An ablation study is performed on MSCOCO under the in-domain setting, with results in Table 4. We build a *Baseline* model without our proposed selective projection optimization (SPO), visual semantic module (VSM), and cross-modal fusion module (CMF) to validate their effectiveness. The modest improvement could be achieved by adding SPO alone, as it helps address the modality gap during the decoding process. To further narrow the modality gap and mitigate the

l	B@4	M	C	S
3	30.5	26.7	110.3	20.6
5	**31.3**	**26.9**	111.8	20.8
7	31.0	26.7	111.4	20.7
9	**31.3**	**26.9**	**112.0**	**20.9**
11	**31.3**	**26.9**	111.4	20.8

Table 6: Performance of ROSCap with different text embedding numbers l for cross-modal fusion.

d	B@4	M	C	S
3	23.0	27.0	84.9	21.7
4	28.0	27.3	103.8	21.7
5	30.1	27.1	109.8	21.1
6	**31.3**	**26.9**	**112.0**	20.9
7	**31.3**	26.6	111.0	20.2

Table 7: Performance of ROSCap with different thresholds d from our visual semantic module.

semantic gap, we integrate text and image embeddings using CMF, which leads to a performance boost. Additionally, our VSM enhances the extraction of visual entities from multiple perspectives, thereby improving results and supporting the establishment of a more robust mapping between images and captions. In the setting (+CMF & VSM w/o SPO), SPO is disabled and synthetic images lacking semantic information are still used to retrieve texts. This approach does not resolve the semantic gap. Mismatched semantic information may exacerbate it, highlighting the necessity of embedding optimization. The model could handle both gaps by first optimizing synthetic image embeddings to reduce the semantic gap with their corresponding captions and then using the optimized embeddings to retrieve additional textual information as semantic supplements. The performance under the setting (+CMF & SPO w/o VSM) also demonstrates the ability to mitigate the modality and semantic gaps.

Effect of Hyperparameters. We evaluate the effect of hyperparameters on MSCOCO under the in-domain setting. For SPO, as shown in Table 5, the synthetic image support set is constructed with k ranging from 3 to 9, achieving optimal performance when five synthetic images are retrieved as the support domain. For CMF, as presented in Table 6, the parameter l is varied from 3 to 11 to determine the optimal setting. For VSM, as shown in Table 7, the number of retrieved sentences followed the CMF configuration, with the parsing and filtering threshold d ranging from 3 to 7. ROSCap achieves the best performance when integrating all three components, which demonstrates the effectiveness of the overall configuration.

4.4 Visualization Analysis

To further assess the effectiveness of the SPO strategy, we conduct t-SNE visualization on 1,000 randomly sampled instances from MSCOCO. In Figure 4, red points represent real image embeddings,

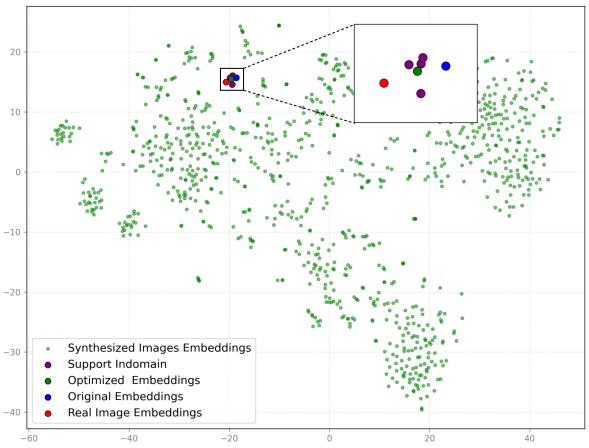


Figure 4: Projection-optimized image embeddings vs. original synthetic image embeddings.

blue points denote the original synthetic image embeddings, green points correspond to the optimized synthetic image embeddings, and purple points indicate the synthetic image support domain. Using the real image embeddings as a reference, the original synthetic embeddings are projected into the support domain. The resulting projected embeddings are closer to the real image embeddings than the original synthetic ones, demonstrating that SPO effectively reduces the semantic gap between synthetic images and their corresponding text.

4.5 Quantitative Analysis

According to CLIPScore (Hessel et al., 2021) and PCM-Net (Luo et al., 2024), higher CLIP similarity reflects stronger alignment between images and their corresponding texts. We conduct a quantitative analysis on two synthetic datasets of differing quality. In the MSCOCO dataset, approximately 75% of the training synthetic images are optimized, whereas in the Flickr dataset, about 46% of the synthetic images are optimized. This indicates that a larger proportion of high-quality samples can be retrieved from the MSCOCO dataset. As shown in Table 8, despite differences in data quality, our method consistently enhances semantic alignment

Training Data	MSCOCO	Flickr
Original Synthetic Pairs	0.3092	0.3043
Projected Synthetic Pairs	**0.3413**	**0.3181**

Table 8: Similarity comparison of original and projected synthetic image-text pairs.

Method	B@4	M	C	S
ViECap (2023)	27.2	24.8	92.9	18.2
ViECap+Mea (2024)	27.2(0)	25.3(0.5 ↑)	95.4(2.8 ↑)	19.0(0.8 ↑)
ViECap+ViP (2025)	27.3(0.1 ↑)	25.1(0.3 ↑)	93.6(0.7 ↑)	18.4(0.2 ↑)
ViECap+VSM (Ours)	29.2(2.0 ↑)	26.0 (1.2 ↑)	103.3 (10.4 ↑)	19.8(1.6 ↑)
SynTIC (2024)	29.9	25.8	101.0	19.3
SynTIC+VSM (Ours)	29.7(0.2 ↓)	26.4(0.6 ↑)	104.0 (3.0 ↑)	20.0(0.7 ↑)
IFCap (2024)	30.8	26.7	108.0	20.1
IFCap+VSM (Ours)	30.4(0.4 ↓)	26.9(0.2 ↑)	109.2(1.2 ↑)	21.0(0.9 ↑)

Table 9: Performance of various baselines integrated with our VSM. +Mea indicates the adoption of the strategy from MeaCap, while +ViP denotes the use of the strategy from ViPCap.

between images and texts, achieving state-of-the-art performance on both datasets.

4.6 Extensibility

To evaluate the extensibility of our proposed VSM, we integrate it into various baseline methods, consisting of ViECap, IFCap, and SynTIC. To further validate its effectiveness, we perform comparative analyses with other representative plug-and-play approaches, such as the zero-shot strategy from MeaCap and the supervised one from ViPCap (Kim et al., 2025). Notably, the visual semantic module operates exclusively during the inference phase and can be seamlessly integrated into any method that maps images to text via entities. As shown in Table 9, incorporating our module consistently improves performance across nearly all metrics, with CIDEr exhibiting a notable increase of 10.4 points over the ViECap method. These results highlight the broad applicability of our visual semantic module, which effectively constructs entity-based hard prompts from multi-modal visual content and guides the model to generate captions that are semantically consistent with the images.

5 Conclusion

We propose ROSCap, a retrieval-based zero-shot image captioning framework designed to bridge both the semantic and modality gaps between synthetic images and their corresponding texts while reducing reliance on image generative models. ROSCap employs a post-processing strategy that optimizes synthetic image embeddings and leverages cross-modal fusion to enhance image-text alignment. Additionally, the visual semantic module enables the extraction of representative entities from multiple modalities, improving the generation of semantically consistent captions. Extensive experiments on benchmark datasets demonstrate that ROSCap achieves state-of-the-art performance and exhibits a strong generalization capability.

Limitations

Although ROSCap outperforms other zero-shot image captioning models trained solely on text, it relies on retrieval and requires a memory containing sufficient information to reduce the semantic gap, making its performance dependent on the availability of relevant semantics. Recent studies have explored reconstructing data to generate semantically similar content. In future work, we plan to investigate the use of reconstructed data as a memory to supplement missing semantic information.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62406081), and the Guangxi Natural Science Foundation (No. 2025GXNSFBA069232).

References

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

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision (ECCV)*, pages 382–398.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The unreasonable effectiveness of clip features for image captioning: An experimental analysis. In *2022*

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 4661–4669.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

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*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10575–10584, Los Alamitos, CA, USA. IEEE Computer Society.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3113–3123.

Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language cnn for image captioning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1231–1240.

Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. 2023. I can't believe there's no images! : Learning Visual Tasks Using Only Language Supervision . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2672–2683, Los Alamitos, CA, USA. IEEE Computer Society.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, Singapore. Association for Computational Linguistics.

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 (CVPR)*.

Taewhan Kim, Soeun Lee, Si-Woo Kim, and Dong-Jin Kim. 2025. Vipcap: retrieval text-based visual prompts for lightweight image captioning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2024. IFCap: Image-like retrieval and frequency-based entity filtering for zero-shot captioning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20715–20727, Miami, Florida, USA. Association for Computational Linguistics.

Weijie Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023. Decap: Decoding clip latents for zero-shot captioning via text-only training. In *International Conference on Learning Representations (ICLR)*.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Zhiyue Liu, Jinyuan Liu, Xin Ling, Qingbao Huang, and Jiahai Wang. 2025. Synthesize then align: Modality alignment augmentation for zero-shot image captioning with synthetic data. *Knowledge-Based Systems*, 315:113274.

Zhiyue Liu, Jinyuan Liu, and Fanrong Ma. 2024. Improving cross-modal alignment with synthetic pairs for text-only image captioning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

Jianjie Luo, Jingwen Chen, Yehao Li, Yingwei Pan, Jianlin Feng, Hongyang Chao, and Ting Yao. 2024. Unleashing text-to-image diffusion prior for zero-shot image captioning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LVII*, page 237–254, Berlin, Heidelberg. Springer-Verlag.

Feipeng Ma, Yizhou Zhou, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. 2024. Image captioning with multi-context synthetic data. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*

and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI Press.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4055–4063, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10968–10977.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2840–2849.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A. Plummer, Ranjay Krishna, and Kate Saenko. 2023. Cola: How to adapt vision-language models to compose objects localized with attributes? *CoRR*, abs/2305.03689.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA. IEEE Computer Society.

Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. 2023. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8011–8021.

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, Melbourne, Australia. Association for Computational Linguistics.

Yoad Tewel, Yossi Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17918–17928.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023. MultiCapCLIP: Auto-encoding prompts for zero-shot multilingual visual captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11908–11922, Toronto, Canada. Association for Computational Linguistics.

Bang Yang, Fenglin Liu, Yuexian Zou, Xian Wu, Yaowei Wang, and David A. Clifton. 2024. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5712–5724.

Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. 2021. Auto-parsing network for image captioning and visual question answering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2177–2187.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14100–14110.

Peipei Zhu, Xiao Wang, Yong Luo, Zhenglong Sun, Wei-Shi Zheng, Yaowei Wang, and Changwen Chen. 2023. Unpaired image captioning by image-level weakly-supervised visual concept recognition. *IEEE Transactions on Multimedia*, 25:6702–6716.

Hyperparameters	MSCOCO	Flickr	NoCaps
Training epoch	5	30	6
Batch size	80	80	80
Optimizer	Adam	Adam	Adam
Learning rate	2e-5	1e-5	2e-5
Temperature τ	1/100	1/100	1/100

Table 10: Hyperparameters.

Method	Image Retrieval	Text Retrieval	Image Decoding	Total
SaCap	–	0.44ms	282.80ms	283.24ms
IFCap	–	0.62ms	140.52ms	141.14ms
ROSCap	0.51ms	0.51ms	139.94ms	140.96ms

Table 11: Comparison of inference overhead.

A Hyperparameter Settings

Our detailed hyperparameter settings, except for k, l, and d, to train our captioning model on three datasets are listed in Table 10. All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs.

B Computational Overhead

We compare the overhead of our proposed ROSCap with that of baselines on MSCOCO. For training, both ROSCap and IFCap take about 2 hours using one NVIDIA GeForce RTX 4090 GPU, while SaCap takes about 4 hours due to the use of additional transformer layers for feature reconstruction. For inference, we list the per-image inference overhead in Table 11, indicating that our method does not introduce additional overhead compared to the baselines using synthetic data, since they also employ some auxiliary strategies like feature reconstruction and object detection. This indicates that the additional retrieval does not increase inference time, demonstrating the efficiency of our method.

C Theoretical Analysis

We provide a simple theoretical derivation showing that our SPO leads to better image-text alignment of embeddings. Prior work indicates that higher CLIP similarity correlates with stronger correspondence between images and their associated texts (Hessel et al., 2021; Luo et al., 2024), and our embedding optimization could improve this similarity. Specifically, the cosine similarity between the synthetic image I and its corresponding text T is defined as $s_0 = \cos(f(I), f(T))$, where $f(\cdot)$ represents the CLIP encoding function. If $f(I)$ is optimized, each image I_j in the support set S_u exhibits higher similarity with the text T compared to I as follows:

$$s_j = \cos(f(I_j), f(T)) \geq s_0.$$

The optimized image embedding Project$(f(I))$ is obtained via a weighted combination of the support set. The similarity between the optimized embedding and the text T is calculated by:

$$\begin{aligned} s &= \cos\big(\text{Project}(f(I)), f(T)\big) \\ &= \sum_{j=1}^{k} a_j s_j \geq \sum_{j=1}^{k} a_j s_0 = s_0, \end{aligned}$$

where a_j denotes the weight for s_j, and $a_1 + a_2 + \cdots + a_k = 1$. Thus, the above derivation shows that the similarity between the optimized image-text pairs is no less than that of the original pairs.

Method	Training data	B@4	M	C	S
ROSCap	Synthetic pairs	31.3	26.9	112.0	20.9
ROSCap	Real pairs	32.7	27.3	115.1	20.8

Table 12: In-domain captioning results on MSCOCO.

D Synthetic vs. Real Images for Training

We compare the results of training the model with synthetic data vs. real data. As shown in Table 12, ROSCap achieves strong performance by applying semantic supplementation to bridge the semantic gap in synthetic image-text pairs using existing information. Notably, several metrics approach or even exceed those based on real image-text pairs, highlighting the effectiveness of our method.

E More Details for Memory

We provide additional details on the memory construction for our selective projection optimization. Both the text memory and image memory are constructed from the training data. Specifically, we employ Stable Diffusion v1-5 (Rombach et al., 2022) to generate a synthetic image for each training text. Following the Karpathy split, the MSCOCO training set contains 566,747 text annotations, which are used to build the MSCOCO text memory, while an equal number of synthetic images are generated to form the corresponding image memory. For Flickr, the Karpathy split provides 145,000 training texts, which are included in the text memory

Figure 5: Examples for retrieving synthetic images to supplement semantics. The training texts and their corresponding synthetic images are shown on the left side, while synthetic images from memory are shown on the right.

alongside 145,000 synthetic images in the image memory. For CC3M, we generate synthetic images for 3,318,333 training texts. In brief, M_{text} stores the encoded features of the training texts, and M_{img} stores the encoded features of their corresponding synthetic images.

F Qualitative Evaluation

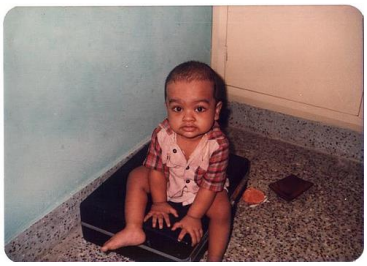
Semantic Supplementation via Memory. As exhibited in Figure 5, we illustrate how ROSCap optimizes features through three representative cases. In (a), even for relatively simple sentences, the image generative model fails to capture fine-grained semantics such as "frisbee". By performing retrieval from the image memory, we successfully supplement the missing semantics, thereby producing a caption of "children playing with a frisbee". In (b), the model omits multiple key elements in this more complex description. By retaining the synthetic image and retrieving semantically adjacent images, we enhanced its expressiveness, restoring missing semantics such as "people riding elephants". In (c), although fine-grained attributes like "dark blue and white sweater" are preserved, the main entity "woman's son" is missing. This gap is filled by retrieving information from a synthetic image that is semantically close to the text.

Generated Captions. We compare the captions generated by ROSCap with those from IFCap under the in-domain setting on the MSCOCO dataset, as shown in Figure 6. The detected objects are listed, with those supplementing semantics and overlooked by IFCap highlighted in blue. By incorporating semantic supplementation and our visual semantic module, ROSCap produces captions that are more accurate and contextually appropriate.

In (a), our method correctly identifies a "duck", thereby rectifying IFCap's misclassification of birds. Similarly, in (b) and (c), ROSCap accurately recognizes "golden hair" and avoids confusing objects such as a hair dryer or sink with a toothbrush, demonstrating strong fine-grained attribute recognition. In (d), our method extracts diverse objects (e.g., toilet), contributing to more complete image descriptions. In (e), IFCap misinterprets the relationship between a man and a horse. In (f), the absence of a detected basketball in the hard prompt leads IFCap to erroneously generate "tennis". In (g), our method accurately identifies the spatial relationships between objects, whereas IFCap misinterprets the scene as a woman next to a dog; our approach correctly recognizes the presence of a man. In (h), ROSCap demonstrates strong perceptual reasoning, successfully inferring a dusk environment and the presence of a sunset even when the hard prompt does not explicitly specify the object. Finally, in (i), although IFCap recognizes a woman sitting, it misidentifies the context by associating her with a bench, whereas our method leverages visual semantics to generate a more accurate caption.

GT: Two ducks floating together on a body of water.
IFCap: Two birds swimming in a body of water.
ROSCap: Two ducks swimming in a body of water.
Object : [water, duck]
(a)

GT: A little girl holding a blow dryer next to her head.
IFCap: A young girl is brushing her hair with a toothbrush.
ROSCap: A little girl with blonde hair holding a hair dryer.
Object : [hair, dryer, girl]
(b)



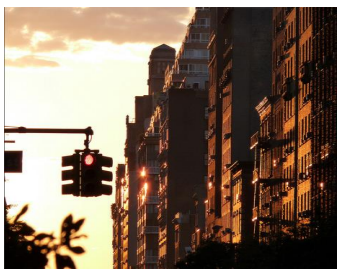
GT: A young baby sits on top of a briefcase.
IFCap: A baby sitting in a sink holding a toothbrush.
ROSCap: A baby sitting in a kitchen next to a sink.
Object : [baby]
(c)

GT: Bathroom with a shower, sink, and toilet in it.
IFCap:A bathroom with a brown tiled floor and a white tiled shower.
ROSCap: A bathroom with a toilet and a shower.
Object : [bathroom, shower, toilet]
(d)

GT: A man riding a brown horse down a city street.
IFCap: A man standing next to a horse on a street.
ROSCap: A man is riding a horse down the street.
Object : [horse, man, street]
(e)

GT: A young man in a green jersey is holding a ball.
IFCap: A young man holding a tennis racquet in front of him.
ROSCap: A young man holding a basketball in his right hand and a ball in his left hand.
Object : [man, basketball, ball]
(f)

GT: A large white dog is sitting on a bench beside an elderly man.
IFCap:A woman sitting on a bench with a dog.
ROSCap:A man sitting on a bench next to a white dog.
Object : [dog, bench]
(g)



GT: A sun setting over a large city and buildings.
IFCap: A view of a city street with tall buildings in the background.
ROSCap: Sunset on a city street with buildings in the background.
Object : [street, city, building]
(h)

GT: A woman is sitting with a suitcase on some train tracks.
IFCap: A woman that is sitting on a bench.
ROSCap: A woman sitting on top of a suitcase.
Object : [woman]
(i)

Figure 6: Visualization of the captioning results from our method and IFCap on the MSCOCO dataset. We show a comparison with the ground truth (GT).