Zero-Shot Image Captioning with Multi-type Entity Representations

Delong Zeng[1], Ying Shen[1,2,3]*, Man Lin[1], Zihao Yi[1], Jiarui Ouyang[1]

[1]Sun Yat-Sen University
[2]Pazhou Lab, Guangzhou 510005, China
[3]Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology, Guangzhou 510006, China
sheny76@mail.sysu.edu.cn

Abstract

As data and computational resources continue to expand, incorporating a variety of knowledge during the pre-training phase enhances large models, providing them with strong zero-shot capabilities. Due to the alignment of modal features by visual language models, zero-shot image captioning no longer necessitates pre-training on paired image-text labeled data, enabling accurate text description generation for images not encountered before. While recent researches focus on methods utilizing entity retrieval as anchors to bridge the gap between different modalities, these approaches often fall short of thoroughly analyzing the impact of entity retrieval recall on the zero-shot generation capabilities. To address this issue, we propose **MERCap**, a zero-shot image captioning method employing **M**ulti-type **E**ntity representation **R**etrieval. More specifically, we first approximate image representation using the CLIP representation of text and Gaussian noise to address the modality gap. Then, we train a GPT-2 to reconstruct text using entities as hard prompts and CLIP representations as soft prompts. Additionally, we construct a domain-specific entity set, assigning multiple representations to each entity and refining their representation vectors through contrastive learning. During inference, we retrieve entities and input them into the decoder to generate corresponding captions. Extensive experiments validate that our approach is efficient, achieving a new state-of-the-art level in cross-domain captioning and in-domain captioning compared to existing methods.

Introduction

With the rapid development of the internet, a large number of unlabeled corpora emerge, driving the advancements in unsupervised learning. In the multimodal domain, various large-scale image-text alignment models proliferate, such as Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021), which achieves alignment between different modalities through pretraining on a substantial number of image-text pairs. These pre-trained models demonstrate impressive zero-shot transferability in various discriminative tasks, such as classification, segmentation, and detection (Siddique et al. 2021; Ding et al. 2022). However, the effective transfer of the perceptual capabilities of these models to zero-shot generative tasks, such as captioning, remains a subject worthy of exploration (Tewel et al. 2022). Image captioning aims to convert image content into textual descriptions (Cheng et al. 2017; Smeaton and Quigley 1996). Conventional approaches typically require a large amount of image-text pair data for training, posing challenges in data acquisition (Farhadi, Hejrati, and Sadeghi 2010; Hodosh, Young, and Hockenmaier 2013; Sun, Gan, and Nevatia 2015). Recently, zero-shot image caption has gained attention. These methods extensively explore the cross-modal alignment information of pre-trained large models, utilizing the representation of text for alignment by introducing noise instead of image representation (Li et al. 2023b). As seen in Figure 1, zero-shot methods do not use image-text pairs for training but instead match through retrieval of shared entities, concepts, or similar sentences as anchors (Tewel et al. 2022; Yang, Liu, and Wu 2023). Through training the text generation model on the task of reconstructing text using textual representations to simulate image representations, these methods aim to achieve image-text generation without relying on paired data.

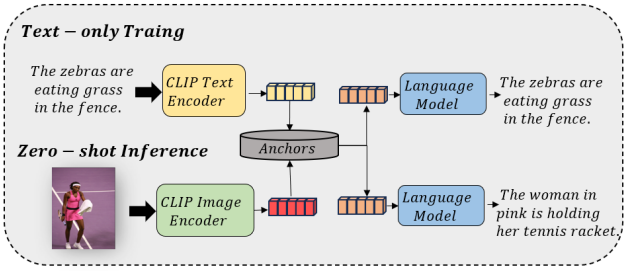


Figure 1: The paradigm of zero-shot image captioning. Only textual data participates in training, and no image-text pairs are utilized. Anchors can be entities or sentences, used to further align the gap between different modalities.

However, recent zero-shot image captioning methods primarily focus on integrating detected concepts such as entities into the input, with less emphasis on researching how to train the matching relationship between images and entities through techniques like contrastive learning in such a zero-shot setting. This leads to the generation of text by the model containing more irrelevant content and lower accu-

*corresponding author

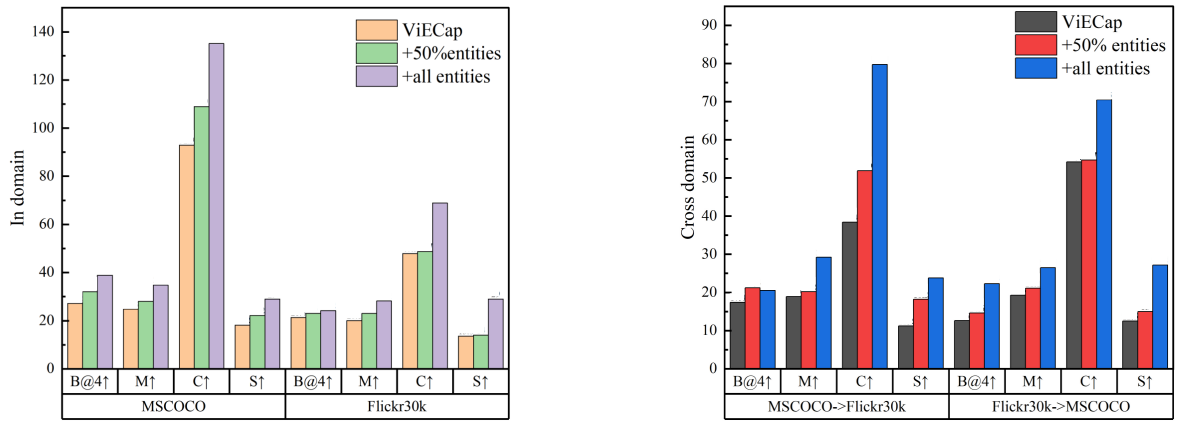


Figure 2: Comparison of In-domain captioning (the image on the left) and Out-of-domain captioning (the image on the right) results with varying degrees of entity recall. The introduced entities are directly extracted from the text using a grammar parser. "50% entities" refers to mask 50% of the corresponding entities.

racy. We observe that more informative entities lead to the generation of higher-quality captions. Figure 2 illustrates the captioning scenario in VieCap(Fei et al. 2023), a method trained with GPT-2 (Radford et al. 2019)using entity hard embeddings and CLIP soft embeddings. In this case, we introduce more entities included in the golden caption as hard embeddings during captioning. The added entities are directly extracted from entities corresponding to the images through a syntax decoder, although achieving this in practical inference is idealistic and represents an upper bound. From Figure 2, it is evident that the inclusion of entities significantly enhances the quality of generated captions, both in In-Domain captioning on Flickr30k(Young et al. 2014) and Cross-Domain captioning from Flickr30k(Young et al. 2014) to MSCOCO(Chen et al. 2015). Even with the recall of only 50% of entities, there is a substantial performance improvement. This inspires us to design an entity retrieval method, such as using contrastive learning to minimize the distance between text and corresponding entity representations, in order to improve retrieval recall.

On the other hand, it is equally important to explore how to represent entities of multiple types with different semantics in various scenarios for entity retrieval. The representation of entities can typically be obtained by feeding them into language models using certain templates. Previous research makes a somewhat limiting assumption that an entity only has one representation (Obeidat et al. 2019; Chatterjee and Dietz 2022). In reality, an entity should have different representations in different contexts. For example, "worker" can have various corresponding images, such as a sanitation worker associated with a street or a construction worker linked to a construction site. Overall, there is still limited research on how to represent entities with multiple types.

To overcome these issues, we propose **MERCap**, a zero-shot image captioning method utilizing multi-type entity representation retrieval. Initially, we use the CLIP representation of text to approximate image representation, addressing the modality gap with the addition of Gaussian noise. Subsequently, we train a GPT-2 decoder capable of reconstructing text through entities as hard prompts and CLIP representations as soft prompts. On the other hand, by constructing a domain-specific entity set, we assign multiple representations to each entity and train the entity's representation vector through contrastive learning. During the inference process, we retrieve entities and input them into the decoder to generate corresponding captions. Our contributions are as follows:

- We design a contrastive learning-based entity representation retrieval algorithm, enabling entities contained in images to be used for caption expression through hard prompts.
- We propose the idea of multiple representations for entities. Assigning different representations to each entity in various scenarios improves retrieval accuracy.
- Experiments show that our method has good performance in zero-shot image generation tasks compared to other baselines. It performs well in both in-domain and out-of-domain scenarios.

Related Work

Supervised Image Captioning. Conventional Supervised methods often employ encoder-decoder architectures. For instance, the Show and Tell model (Xu, Ba, and Kiros 2015) integrates Convolutional Neural Networks with Long Short-Term Memory Networks. And VST(Zhang, Xie, and Liu 2023) introduces attention mechanisms, enhancing the encoding and decoding capability. Recently, various large multimodal models have been preconditioned on extensive text-image paired data, which demonstrate significant effectiveness in various tasks. CLIP (Radford et al. 2021) employs common text and image encoders for feature extraction from multimodal data and trains on a large dataset. BLIP (Li et al. 2022) adopts a Multimodal mixture of Encoder-Decoder (MED) structures for multitask pretraining and transfer learning. Then BLIP-2 (Li et al. 2023a)

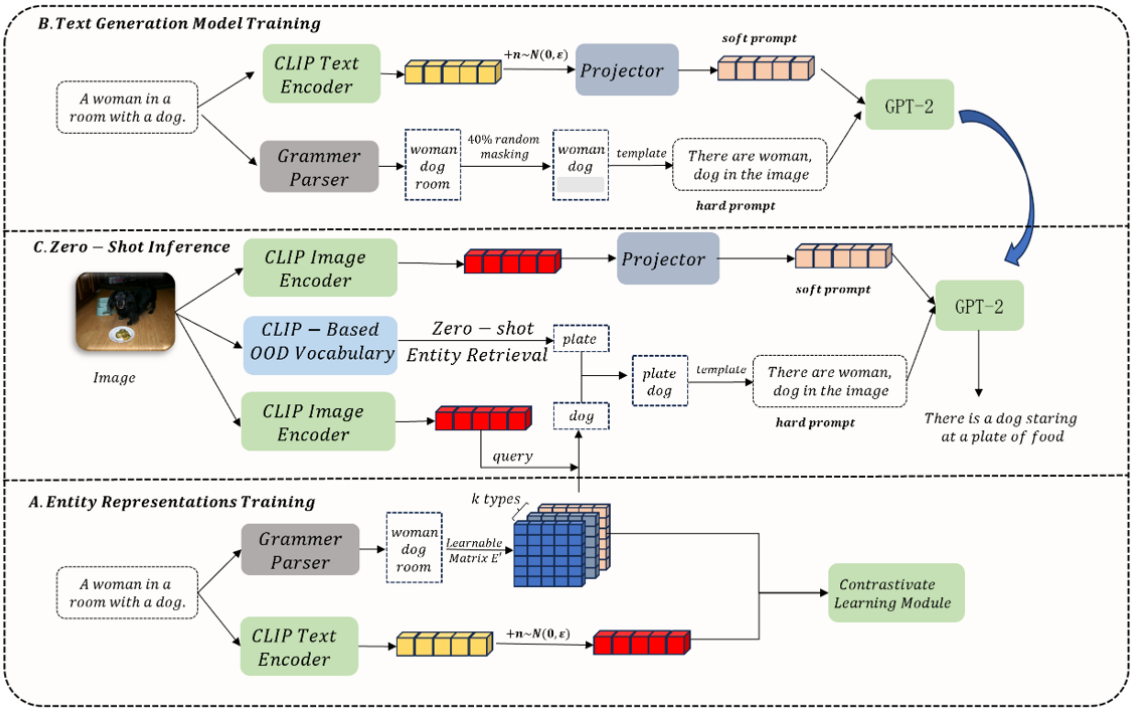


Figure 3: The pipeline of MERCap. The multiple-type representations of entities are trained through contrastive learning by extracting positive and negative entity pairs (Part A). The training of the decoder involves using CLIP representations as soft embeddings and entities as hard embeddings (Part B). During inference, entity retrieval is performed separately for Out-of-Domain (OOD) and In-Domain (ID), and captioning is carried out using CLIP-aligned soft embeddings (Part C).

enhances multimodal effects and reduces training costs by leveraging Vision-Language Pre-Training (VLP) models. The VLP model provides high-quality visual representations and robust language generation capabilities. Many recent researches fine-tune VLP models to generate image captions. For instance, K-Replay(Cheng, Song, and Ma 2023) employs a series of knowledge-enhanced techniques to mitigate the issue of hallucinated information. To bridge the modality gap in the caption generation, some methods use contrastive learning(Yang et al. 2022; Deng, Zhong, and Wang 2023) for multi-granularity(Cho et al. 2022) and multi-angle(Tu et al. 2024, 2023) alignment. CCR(Tu et al. 2025) further reduces the modality gap by maximizing the contrastive alignment between relevant features and generated words.

Zero-shot Image Captioning. With the development of large scale models, an increasing number of zero-shot fine-tuning methods are proposed. The objective of zero-shot image text generation is to generate image captions without relying on annotated data (Li et al. 2023b).

Some methods in this field (Changpinyo et al. 2021; Wang et al. 2022) pre-train models on a large-scale dataset of weak image-text pairs and evaluate the models on a target benchmark without further fine-tuning. Another set of methods (Tewel et al. 2022; Su et al. 2022; Yang, Liu, and Wu 2023; Fei et al. 2023) achieves zero-shot image captioning by combining large Visual-Linguistic Models (VLMs) and large Language Models (LLMs). Specifically, VLMs provide visually guided language cues, directing LLMs to generate captions related to the images. ZeroCap (Tewel et al. 2022) aligns images and text in the same space using CLIP. It introduces Gaussian noise to perturb textual representations, simulating image representations. However, while CLIP aligns the two modalities, simple noise addition is insufficient to bridge the semantic gap between the two modalities. Some methods utilize entity concepts as anchors, performing secondary alignment of vectors from both modalities. For instance, MultiCapCLIP (Yang, Liu, and Wu 2023) uses concept words retrieved from a corpus, i.e., descriptive phrases, as soft embeddings to complement representations. K-night (Wang et al. 2023) collected k-nearest neighbor representations as soft prompts, extracting more semantic information. ViECap (Fei et al. 2023), on the other hand, masks training by extracting nouns and enriched semantic representations of generated content with hard prompts. MeaCap(Zeng et al. 2024) further obtains the key concepts from memory sentences. However, these methods do not explore the more reasonable role of entity representation in aiding captioning.

Method

Problem Definition

In this paper, our objective is to generate a textual caption, complying with the image description, for each image I_i. Given the entity set $E' = \{e_1, e_2, \ldots, e_M\}$, for each image I_i, we initially employ the retrieval algorithm R to retrieve

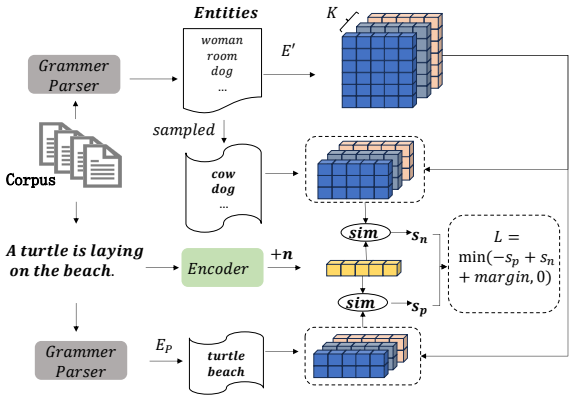


Figure 4: The training process of entity representation. The text undergoes encoding by CLIP's text encoder and cannot approximate the image representation. The grammar parser is employed to extract positive entities and negative entities are randomly sampled for contrastive learning training.

its implied entities $\{e_{i1}, e_{i2}, \dots\}$. Subsequently, through the generation model G, we produce the corresponding text t_i. Our scenario operates under zero-shot conditions, meaning that both the retrieval algorithm R and the generation model G are trained on a corpus of pure text $T = \{t_1, t_2, \dots, t_N\}$, without utilizing corresponding image information.

As seen in Figure 3, we begin by training representations of entities using contrastive learning. Then our textual data, combined with the information of the entity through CLIP encoding, is fed into GPT-2 for decoding training. Finally, zero-shot image captioning can be achieved through the utilization of entity representations and the decoder.

Entity Representations Training

First, we assign a representation $I'_i \in \mathbb{R}^d$ to each image. However, in the text-only setting, we cannot directly obtain images. As mentioned earlier, using CLIP allows for a preliminary alignment of images and their corresponding text. This alignment enables the approximation of image representations using text representations:

$$I'_i = \text{CLIP}(I_i) \approx \text{CLIP}(t_i). \tag{1}$$

Inspired by the previous work(Li et al. 2023b; Fei et al. 2023), we can use the sum of noise within a spherical space to substitute the representation of images with the representation of text:

$$I'_i = \text{CLIP}(t_i) + \boldsymbol{n}, \boldsymbol{n} \sim N(0, \epsilon), \tag{2}$$

where ϵ is the given variance. It can be seen that our approach falls within the **zero-shot** category as we solely utilize textual representations to approximate images, without incorporating matched image-text pairs.

The training of entity representations is a focal point in this research. As seen in Figure 4, our objective is to assign a representation to each entity, allowing us to capture entities relevant to a given image I_i based on a specified similarity measurement function. Firstly, we extract all nouns from the corpus that appear more than N_{thres} times to construct the entity set. Then, we initialize a learnable matrix $E' \in \mathbb{R}^{M \times K \times d}$, where M is the number of entities. It is noteworthy that we aim to provide multiple types of representations for each entity. Therefore, the vector dimension corresponding to each entity is Kd, where d is the dimensionality similar to that of the image, and K represents the specified number of representation types for each entity.

We employ **contrastive learning** to train entity representations. Initially, we obtain positive instances of entities. Since there is no image information available, we utilize the NLTK[1] tool to extract entities. The specific approach involves extracting all noun phrases for each text, forming the corresponding positive entity set E_p. Additionally, we randomly select N' negative entities from the total entity set E', excluding those in the positive entity set E_p, to construct the negative entity set E_n. Our goal is to maximize the similarity between the representation I'_i of the image and positive entities while minimizing the similarity with negative entities. In our experiments, we find that representing positive and negative similarities using the lowest similarity with positive entities and the highest similarity with negative entities is more effective than using average values. Therefore, the contrastive learning loss function can be expressed as:

$$L = \max\left(-s_p + s_n + margin, 0\right), \tag{3}$$

$$s_p = \max_{e \in E_p} \boldsymbol{similarity}(I'_i, E'_e), \tag{4}$$

$$s_n = \max_{e \in E_n} \boldsymbol{similarity}(I'_i, E'_e), \tag{5}$$

where $\boldsymbol{margin}$ is the tolerated gap, $\boldsymbol{similarity}$ is the designed similarity measurement function. Since each entity has K representations, we use the cosine similarity between the image representation and the maximum cosine similarity with these multiple representations to measure similarity, as follows:

$$\boldsymbol{similarity}(I'_i, E'_e) = \max_{i=1,2,\dots,K} \left(\frac{I'_i \cdot E'_e}{|I'_i| \cdot |E'_{ei}|} \right). \tag{6}$$

Text Generation Model Training

Due to the availability of only textual data, our objective is to train a language decoder capable of text reconstruction through the perception of entities. We extract two types of information from the ground truth text corresponding to images, which are helpful for modeling examples:

1) Entity Information: To prevent the model from simply copying all nouns into the final generated text, we follow the setting of VieCap and masked 40% of the entities. Thus, for each image I_i corresponding to a text t_i, we obtain the entity set $\{e_{i1}, e_{i2}, \dots\}$. We use the following template to create a hard embedding that allows the decoding model to understand entity information:

$$\boldsymbol{hard}(\{e_{i1}, \dots\}) = \text{“}There\ are\ e_{i1},\ e_{i2},\ \dots,\ in\ the\ Image.\text{”} \tag{7}$$

2) Embedding Information: Similar to the retrieval module, we use the noised text representation to simulate the

[1]https://www.nltk.org/

image representation I_i'. Some studies concatenate the features of several training texts with the highest cosine similarity to the input features to form a soft prompt, thereby enhancing the in-domain representation capability. For comparability, we also use L_s nearest neighbor texts according to K-night's(Wang et al. 2023) setup and project them into a soft prompt through a Multilayer Perceptron(MLP).

$$\boldsymbol{soft}(I_i') = MLP(I_i, T_1, ..., T_{L_s-1}) \tag{8}$$

$$T_j \in topk(\text{similarity}(I_i, T)) \tag{9}$$

We fine-tune the GPT-2 model for decoding both types of information, using autoregressive target functions to train the model G with parameters θ, defined as follows:

$$L_{\text{G}} = -\frac{1}{|w|}\sum_{i=1}^{|w|}\log p(w_i|\boldsymbol{soft}(I_i'); \boldsymbol{hard}(\{e_{i1}, \ldots\}); w_{\leq i}; \theta). \tag{10}$$

Zero-Shot Inference

During the inference process, for a given image I_i, we first need to retrieve relevant entities. Since the caption is not available during the inference stage, we cannot utilize nltk for extraction. Instead, we obtain entities by calculating the cosine similarity between entity representations and image representations.

$$p_e = \frac{\exp\left(\frac{\text{similarity}(I_i', E_e')}{\tau}\right)}{\sum_{j=1}^{M}\exp\left(\frac{\text{similarity}(I_i', E_j')}{\tau}\right)}, \tag{11}$$

where τ is the temperature, used to adjust the coherence of the probability distribution. We select the top k class names with probabilities p_e greater than the threshold p_{thres} as the retrieved entities. Based on this trained retrieval module, primarily constructed from the internal dataset dictionary, we obtain the in-domain entity set:

$$E_{\text{in}} = \{e \,|\, p_e > p_{\text{thres}}, e \in \text{top}_k(E')\}. \tag{12}$$

Additionally, to enhance the out-of-distribution (OOD) capability of entity retrieval, we follow the approach of VieCap and introduce an external word table to retrieve the external entity set E_{out}. Therefore, the relevant entities retrieved for I_i during the latest inference stage are:

$$\{e_{i1}, e_{i2}, \ldots\} = E_{\text{in}} \cup E_{\text{out}}. \tag{13}$$

Thus, the final caption can be generated as:

$$T' = G(\boldsymbol{soft}(I_i'), \boldsymbol{hard}(e_{i1}, e_{i2}, \ldots) \,|\, \theta). \tag{14}$$

Experiment

In this section, we first introduce the datasets, evaluation metrics and baselines, and implementation details. Then we make a thorough examination to answer the following research questions:

- **RQ1**: Can training entity representations through contrastive learning enhance the quality of captions?
- **RQ2**: How does entity multi-type representation influence captioning capability?

Datasets and Evaluation Metrics. Experiments are conducted on three widely used image captioning benchmarks: NoCaps (Agrawal et al. 2019), COCO (Chen et al. 2015), and Flikcr30k (Young et al. 2014). For COCO and Flickr30k, we use the commonly adopted Karpathy split. For NoCaps, the model is trained on the COCO training set, and results are reported on the validation set following OSCAR's recommendations. Common captioning evaluation metrics BLEU@4 (B@4) (Papineni et al. 2002), METEOR (M) (Denkowski and Lavie 2014), CIDEr (C) (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (S) (Anderson et al. 2016) are used to demonstrate our results in comparison with baseline methods. In addition, we use R@5 to evaluate the recall rate of different settings.

Methods. We include several captioning methods as baselines. On the one hand, we leverage several supervised methods as references: 1) OSCAR (Li, Yin, and Li 2020) as a classical supervised approach, 2) I-Tuning(Luo et al. 2023) and SmallCap as(Ramos et al. 2023) lightweight paired captioning methods, using GPT-2 for CLIP-based caption generation. On the other hand, we compare MERCap with several zero-shot image captioning methods trained solely on text: 3) MAGIC employs late-stage bootstrapped decoding(Su et al. 2022), 4) CapDec(Nukrai, Mokady, and Globerson 2022) leverages Gaussian noise to mitigate the modality gap, 5) DeCap (Li et al. 2023b) and Knight (Wang et al. 2023) incorporate a straightforward support memory containing embeddings of the text corpus in the pre-training stage. Because Knight used a larger language model, and in order to align it, we modified it to the GPT-2_{Base} (Wolf et al. 2020) and reran the code according to the paper. 6) VieCap (Fei et al. 2023) utilizes specific templates to fill entities for entity retrieval. 7) MeaCap (Zeng et al. 2024) extracts entities from memory corpora as hard embeddings.

Implementation Details. We utilize CLIP-ViT-B/32[2] as the backbone. The language model adopted is GPT-2_{Base} (Wolf et al. 2020). All our experiments are executed on a single Nvidia GPU RTX 3090 Ti with 24GB of physical memory. During the training of the decoder,we maintain a batch size of 80 and a learning rate of 2e-5 for both Flickr30k and MSCOCO datasets, conducting training for 30 epochs. Additionally, we set the length of soft prompts $L_s = 7$. In the training of entity representations, we set ϵ to 0.016, N' to 50, batch size to 256, and a learning rate of 0.01. Specifically, we configure $K = 5, N_{thres} = 15$ for Flickr30k, and $K = 3, N_{thres} = 15$ for MSCOCO, training each for 10 epochs. In the inference phase, the temperature is set to 0.01, while the threshold is set to 0.4 and 0.8 for Flickr30k and MSCOCO, respectively. To ensure a fair comparison, we align with VieCap in terms of the external dictionary. For Flickr30k testing, we use the VGOI vocabulary (Zhang et al. 2021), and for MSCOCO and NoCap testing, we employ the COCO vocabulary (Fei et al. 2023).

[2]https://huggingface.co/sentence-transformers/clip-ViT-B-32

		MSCOCO⇒Flickr30k				Flickr30k⇒MSCOCO			
	Venue	BlEU@4↑	METEOR↑	CIDEr↑	SPICE↑	BlEU@4↑	METEOR↑	CIDEr↑	SPICE↑
MAGIC	ArXiv'22	6.2	12.2	17.5	5.9	5.2	12.5	18.3	5.7
CapDec	EMNLP'22	17.3	18.6	35.7	-	9.2	16.3	27.3	-
DeCapc	ICLR'23	16.3	17.9	35.7	11.1	12.1	18.0	44.4	10.9
Knight*	IJCAI'23	16.2	17.3	33.6	10.4	12.3	17.3	40.1	9.7
ViECap	ICCV'23	17.4	18.0	38.4	11.2	12.6	19.3	54.2	12.5
MeaCap	CVPR'24	**18.5**	19.5	43.8	12.8	13.1	19.7	56.4	13.2
MERCap	-	17.9	**19.6**	**45.6**	**13.0**	**14.4**	**20.2**	**59.8**	**13.8**

Table 1: Experimental results of cross-domain captioning. Training is performed on the source domain corpus, followed by inference on the target domain test set images. * are our reproduced results.

COCO ⇒ NoCapsval								
Methods	ID		ND		OD		Overall	
	C↑	S↑	C↑	S↑	C↑	S↑	C↑	S↑
Supervised learning, zero-shot inference								
OSCAR	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
I-Tuning	83.9	12.4	70.3	11.7	48.1	9.5	67.8	11.4
SmallCap	83.3	-	77.1	-	65.0	-	75.8	-
Text-only training, zero-shot inference								
CapDec	60.1	10.2	50.2	9.3	28.7	6.0	45.9	8.3
DeCap	65.2	-	47.8	-	25.8	-	45.9	-
ViECap	61.1	10.4	64.3	9.9	**65.0**	8.6	66.2	9.5
MERCap	**67.0**	**11.4**	**69.4**	**11.5**	64.0	**9.9**	**69.4**	**11.0**

Table 2: Results of in-Domain (ID),Near-Domain (ND) and Out-Domain (OD) captioning evaluated on the NoCaps validation set are presented. Results from CapDec are sourced from ViECap. The supervised learning methods are used for reference only and are not included in the comparison.

	MSCOCO				Flickr30k			
	B↑	M↑	C↑	S↑	B↑	M↑	C↑	S↑
ZeroCap	7	15.4	34.5	9.2	5.4	11.8	16.8	6.2
MAGIC	12.9	17.4	49.3	11.3	6.4	13.1	20.4	7.1
CapDec	26.4	25.1	91.8	-	17.7	20	39.1	-
Decap	24.7	25	91.2	18.7	21.2	21.8	56.7	15.2
Knight*	25.9	24.5	90.4	18.0	21.9	20.8	50.7	14.0
ViECap	27.2	24.8	92.9	18.2	21.4	20.1	47.9	13.6
MeaCap	27.2	25.3	95.4	19.0	22.3	**22.3**	**59.4**	15.6
MERCap	**27.3**	**25.5**	**96.0**	**19.5**	**23.2**	**22.3**	57.2	**15.9**

Table 3: In-domain captioning results on the COCO test set and Flickr30k test set. It should be noted that the results of ZeroCap are copied from MAGIC. Bold represents the best results, and underlining indicates the second-best results.

Cross Domain Captioning

Table 1 shows the cross-domain validation results with COCO and Flickr30k as the source and target domains, respectively, while Table 2 presents the results of training on the COCO dataset and evaluating on the NoCaps dataset. From these tables, the following conclusions can be drawn:

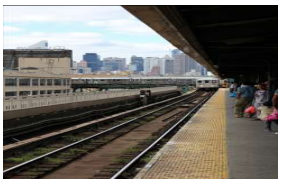
- Table 1 illustrates that our method outperforms others in the mutual validation between the MSCOCO and Flickr30k datasets, indicating good transferability. The entity representations trained on one dataset effectively capture the relationships between entities and images.
- Compared to VieCap and MeaCap, which also utilizes entities as hard embeddings, in Table 1, we achieve superior performance in METEOR, CIDEr, and SPICE metrics. Particularly noteworthy is the substantial improvement in CIDEr scores, rising from 43.8 to 45.6 and from 56.4 to 59.8, respectively. This enhancement indicates that training entity representations through contrastive learning makes positive entities more retrievable, thereby increasing the relevance of generated content, which also answers **RQ1**.
- As seen in Table 2, among the baselines, both DeCap and CapDec perform well in the in-domain setting but poorly in the out-of-domain setting. VieCap(Fei et al. 2023) performs well in the out-of-domain setting due to the introduction of an external dictionary, but, conversely, experiences a slight decline in in-domain performance. Our approach constructs two dictionaries for in-domain and out-of-domain, simultaneously training the in-domain dictionary by contrastive learning. Consequently, our method performs well in both in-domain and out-of-domain settings. In the near-domain comparison, our method shows a significant improvement in both metrics compared to other unsupervised baselines, achieving state-of-the-art performance in the overall setting. Compared to some supervised methods, our approach still demonstrates competitive CIDEr and SPICE scores in near-domain, out-of-domain, and overall settings. This validates that enhancing coverage through entity retrieval improves the quality of generation.

In Domain Captioning

The experimental results of In-Domain Captioning can be seen in Table 3, and we make the following observations:

- On both the MSCOCO and Flickr30k datasets, our model exhibits the best or second-best performance across various metrics. Specifically, on MSCOCO, we achieve the highest scores in three metrics compared to other models, demonstrating the competitiveness of our approach in the In-Domain Captioning task.
- In comparison to Knight, which also utilizes a GPT-2 trained decoder and employs CLIP representations of images and neighbor texts' representations as soft embed-

Our predicted entities: plate, dog
Our predicted caption:
A small brown dog sitting on top of a green plate.
VieCap predicted entities: None
VieCap predicted caption:
A brown dog with a green collar eating a hot dog.
Grouth captions:
A small black dog standing over a plate of food.



Our predicted entities: train, platform
Our predicted caption:
A large long train on a steel track next to a platform.
VieCap predicted entities: train
VieCap predicted caption:
A large long train on a steel track.
Grouth captions:
some people on a platform and a silver train.

Figure 5: Case Study. We conducted a comparative analysis between our method and the baseline, focusing on the retrieved entities and the generated sentences from images.

dings, our method outperforms across all metrics. This suggests that incorporating retrieved entities as hard embeddings contributes to more accurate generation.

Ablation Study

We conduct ablation experiments on the Flickr30k dataset to investigate caption quality under varying entity types by controlling the parameter K while keeping other parameters constant . As displayed in Table 4, when $K = 1$, the absence of multiple entity representations results in suboptimal performance across all metrics. By introducing multiple types ($K > 1$), we observe peak performance, particularly in R@5 and other metrics, with CIDEr (C) showing a 1.5 increase at $K = 5$ compared to $K = 1$. This suggests that diverse entity representations enhance recall and improve generated content quality, which also provides confirmation for **RQ2**. Moreover, comparative results indicate that higher K values do not guarantee improved metrics; $K = 5$ yields the best performance, with subsequent increases in K leading to metric declines. This finding suggests that five representations per entity are optimal for Flickr30k. Additionally, we conducted further ablation studies by removing neighboring text, followed by the elimination of the soft prompt. To compare with the results obtained with contrastive learning, we directly encoded each entity using CLIP, averaging the text encodings that included the entity along with a cluster center of three for multi-type representation. The results presented in Table 4 indicate that all proposed modules make a significant contribution to the overall performance.

Case Study

We selected three images from the MSCOCO test set and performed zero-shot inference using our method and VieCap as a baseline. The comparison of the entities retrieved by each method and the corresponding generated caption results can be seen in Table 5.

	B@4↑	M↑	C↑	S ↑	R@5 ↑
w/o entities	7.6	9.1	13.8	5.2	-
w/o multi-type	22.4	21.9	55.7	15.6	37.52
$K = 3$	22.7	22.1	57.0	15.6	40.48
$K = 5$	**23.2**	**22.3**	**57.2**	**15.9**	**40.77**
$K = 7$	23.0	22.0	56.4	15.7	39.29
w/o soft prompts	4.7	10.1	18.3	6.1	-
w/o neighbor	21.3	20.7	49.6	14.6	-
w/o contrastive learning					
CLIP embedding	22.2	21.5	55.0	15.2	28.8
Caption mean	22.7	21.6	55.2	15.4	31.8
Gaussian clustering	22.7	21.7	56.0	15.4	35.7

Table 4: Results of ablation experiments on Flickr30k. w/o means without. When $K = 1$, no multi-entity type representation is introduced; when $K > 1$, each entity has a corresponding number of representations. The same decoder is used under different configurations.

Compared to the baseline, our method retrieves a more comprehensive set of entities. In the first image, our method identifies both ”plate” and ”dog” entities, while the baseline fails to detect any entities. Similarly, in the other images, our method retrieves ”platform” representing locations. As a result, our method provides more comprehensive descriptions of the images in the generated captions, approaching the ground truth captions. On the other hand, in comparison to the baseline, when the retrieved entities are too few, the decoder itself exhibits some entity prediction capability due to training with partially masked entities. However, this predictive ability is evidently less accurate than the entities obtained through vector retrieval. For instance, in the first image, the baseline produces a subpar caption by predicting ”hot dog”, a content not mentioned in the image.

Conclusion

In this paper, we first observe that improving entity retrieval accuracy and assigning multiple representations to entities are beneficial for captioning. Consequently, we propose a contrastive learning-based entity multi-representation training method. Experimental results in both In-domain and Cross-domain scenarios demonstrate the competitiveness of our entity representation algorithm compared to other baseline methods. Additionally, our ablation experiments on varying the number of entity types validate the effectiveness of multi-type representations. On the efficiency front, despite the increased inference steps with entity retrieval, accurate entities assist the decoder in reducing irrelevant content generation, thereby enhancing captioning efficiency.

In future work, we aim to investigate the nuanced relationship between entity recall rates and captioning. Exploring common retrieval metrics revealed that the correlation is not always positive, possibly due to varying weights of entities influencing caption quality. We plan to leverage entity co-occurrence and frequency information for a more refined understanding of this association.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3558–3568.

Chatterjee, S.; and Dietz, L. 2022. BERT-ER: query-specific BERT entity representations for entity ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1466–1477.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cheng, K.; Song, W.; and Ma, Z. e. a. 2023. Beyond generic: Enhancing image captioning with real-world knowledge using vision-language pre-training model. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Cheng, Y.; Huang, F.; Zhou, L.; Jin, C.; Zhang, Y.; and Zhang, T. 2017. A hierarchical multimodal attention-based neural network for image captioning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 889–892.

Cho, J.; Yoon, S.; Kale, A.; Dernoncourt, F.; Bui, T.; and Bansal, M. 2022. Fine-grained Image Captioning with CLIP Reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 517–527.

Deng, L.; Zhong, Y.; and Wang, M. e. a. 2023. CONICA: A Contrastive Image Captioning Framework with Robust Similarity Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11583–11592.

Farhadi, A.; Hejrati, M.; and Sadeghi, M. A. e. a. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, 15–29. Springer.

Fei, J.; Wang, T.; Zhang, J.; He, Z.; Wang, C.; and Zheng, F. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3136–3146.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, W.; Zhu, L.; Wen, L.; and Yang, Y. 2023b. DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training. In *The Eleventh International Conference on Learning Representations*.

Li, X.; Yin, X.; and Li, C. e. a. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer.

Luo, Z.; Hu, Z.; Xi, Y.; Zhang, R.; and Ma, J. 2023. I-Tuning: Tuning Frozen Language Models with Image for Lightweight Image Captioning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Nukrai, D.; Mokady, R.; and Globerson, A. 2022. Text-Only Training for Image Captioning using Noise-Injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4055–4063.

Obeidat, R.; Fern, X.; Shahbazi, H.; and Tadepalli, P. 2019. Description-based zero-shot fine-grained entity typing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ramos, R.; Martins, B.; Elliott, D.; and Kementchedjhieva, Y. 2023. SmallCap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2840–2849.

Siddique, A.; Jamour, F.; Xu, L.; and Hristidis, V. 2021. Generalized zero-shot intent detection via commonsense knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1925–1929.

Smeaton, A. F.; and Quigley, I. 1996. Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 174–180.

Su, Y.; Lan, T.; Liu, Y.; Liu, F.; Yogatama, D.; Wang, Y.; Kong, L.; and Collier, N. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Sun, C.; Gan, C.; and Nevatia, R. 2015. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE international conference on computer vision*, 2596–2604.

Tewel, Y.; Shalev, Y.; Schwartz, I.; and Wolf, L. 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*, 17918–17928.

Tu, Y.; Li, L.; Su, L.; Yan, C.; and Huang, Q. 2025. Distractors-Immune Representation Learning with Cross-modal Contrastive Regularization for Change Captioning. In *European Conference on Computer Vision*, 311–328. Springer.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2805–2815.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, J.; Yan, M.; Zhang, Y.; and Sang, J. 2023. From association to generation: text-only captioning by unsupervised cross-modal mapping. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4326–4334.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *International Conference on Learning Representations*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

Xu, K.; Ba, J.; and Kiros, R. e. a. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Yang, B.; Liu, F.; and Wu, X. e. a. 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)*, 11908–11922.

Yang, C.-F.; Tsai, Y.-H. H.; Fan, W.-C.; Salakhutdinov, R. R.; Morency, L.-P.; and Wang, F. 2022. Paraphrasing is all you need for novel object captioning. *Advances in Neural Information Processing Systems*, 35: 6492–6504.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 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.

Zeng, Z.; Xie, Y.; Zhang, H.; Chen, C.; Chen, B.; and Wang, Z. 2024. MeaCap: Memory-Augmented Zero-shot Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14100–14110.

Zhang, J.; Xie, Y.; and Liu, X. 2023. Improving Image Captioning through Visual and Semantic Mutual Promotion. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5579–5588.