IFCap: Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning

Soeun Lee[*] **Si-Woo Kim**[*] **Taewhan Kim** **Dong-Jin Kim**[†]
Hanyang University, South Korea.
{soeun, boreng0817, taewhan, djdkim}@hanyang.ac.kr

Abstract

Recent advancements in image captioning have explored text-only training methods to overcome the limitations of paired image-text data. However, existing text-only training methods often overlook the modality gap between using text data during training and employing images during inference. To address this issue, we propose a novel approach called Image-like Retrieval, which aligns text features with visually relevant features to mitigate the modality gap. Our method further enhances the accuracy of generated captions by designing a fusion module that integrates retrieved captions with input features. Additionally, we introduce a Frequency-based Entity Filtering technique that significantly improves caption quality. We integrate these methods into a unified framework, which we refer to as IFCap (**I**mage-like Retrieval and **F**requency-based Entity Filtering for Zero-shot **Cap**tioning). Through extensive experimentation, our straightforward yet powerful approach has demonstrated its efficacy, outperforming the state-of-the-art methods by a significant margin in both image captioning and video captioning compared to zero-shot captioning based on text-only training.[1]

1 Introduction

The task of image captioning generates appropriate textual descriptions for images by combining computer vision (CV) and natural language processing (NLP). With the emergence of Large Language Models (LLMs) and Vision and Language Models (VLMs), various works have studied efficient training methods for image captioning methods [15, 17, 22]. These approaches develop effective captioning by using pre-trained models with few parameters or lightweight networks. However, they rely on paired image-text data, which is costly. To overcome this, recent studies have explored text-only training methods for image captioning, aiming to solve the problem using only textual data [7, 12, 14, 16, 18, 30, 35].

Text-only training introduces a new direction in which models are trained solely using text data. Recent existing works have studied what to use as extra cues, such as extracted nouns [7], generated synthetic images [14, 16] for training, and extracted tags from object detectors [14]. However, existing methods that rely on object information are sensitive to incorrect data, and utilizing large external models (e.g., stable diffusion [23] or object detectors [5]) incur additional costs. Thus, we aim to address the problem by acquiring diverse information cost-effectively without additional models.

The retrieval task involves finding relevant information in a database for a given query. Initially rooted in NLP [11], the field has expanded into CV and into multi-modal retrieval. Depending on

[*] Equal contribution. [†] Corresponding author.

[1] Code: `https://github.com/boreng0817/IFCap`

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

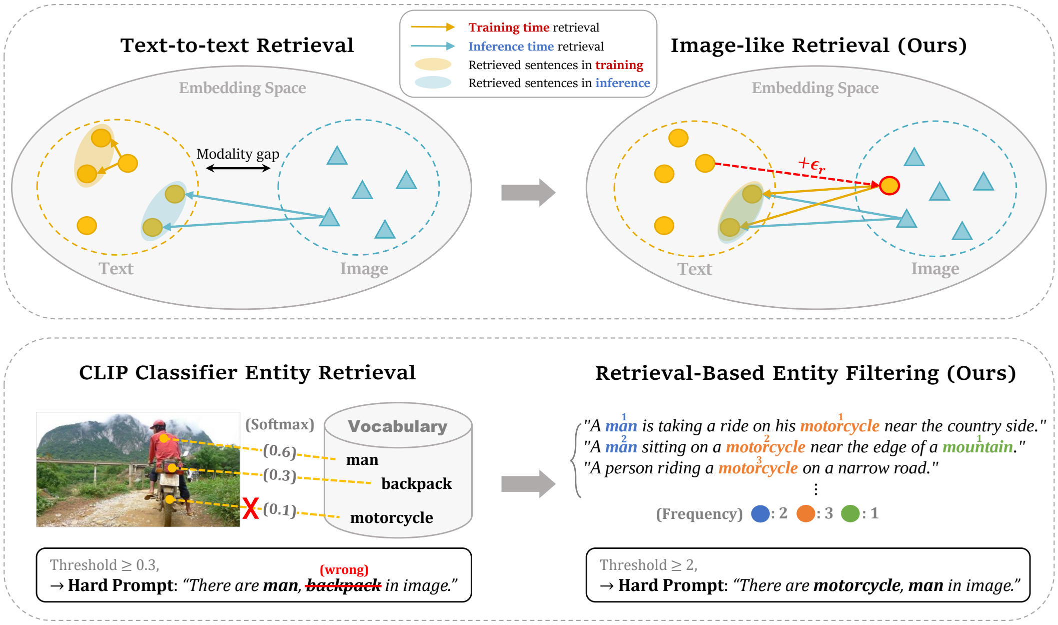

Text-to-text Retrieval
Training time retrieval
Inference time retrieval
Retrieved sentences in training
Retrieved sentences in inference
Image-like Retrieval (Ours)
Embedding Space
Modality gap
Text
Image
Embedding Space
$+\epsilon_r$
Text
Image
CLIP Classifier Entity Retrieval
Retrieval-Based Entity Filtering (Ours)
(Softmax)
Vocabulary
(0.6)
man
(0.3)
backpack
(0.1)
motorcycle
"A man is taking a ride on his motorcycle near the country side."
"A man sitting on a motorcycle near the edge of a mountain."
"A person riding a motorcycle on a narrow road."
(Frequency) ●: 2 ●: 3 ●: 1
Threshold ≥ 0.3,
(wrong)
→ Hard Prompt: "There are man, ~~backpack~~ in image."
Threshold ≥ 2,
→ Hard Prompt: "There are motorcycle, man in image."

Figure 1: (Top) The previous text-to-text retrieval approach overlooks the modality gap, leading to different information use between training and inference. Our approach addresses this by aligning text features with the image embedding space during retrieval. (Bottom) The traditional CLIP classifier-based entity retrieval method struggles with entity detection as vocabulary size grows. Our approach detects frequently occurring words in retrieved captions, extracting entities more accurately without relying on a limited vocabulary.

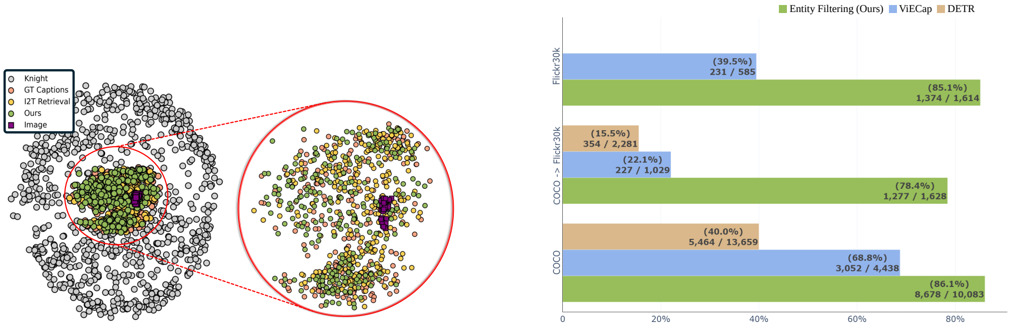
the input data and database, various retrieval methods are possible, such as image-to-text [22] and text-to-text retrieval [30]. In the existing text-only training study, there have been attempts to use the text-to-text retrieval method [30]. However, existing works can't address the modality gap inherent in text-only training settings, where training is performed with text and inference with images. In addition, such works rely too much on retrieved captions without considering visual information. This modality gap and the use of a narrow scope of information may lead to performance degradation.

To verify this, we visualize the analysis result of the CLIP embedding feature of retrieved captions that the model uses in training via t-SNE in Fig. 2a. The analysis is done on the COCO [6] validation split, and the CLIP similarity-based KNN algorithm is used for retrieval. In the figure, there is a large difference between the distribution of features used after image-to-text retrieval and text-to-text retrieval, which shows that a modality gap exists between image and text.

To tackle this issue, we propose a novel approach called "image-like retrieval," that addresses the modality gap between image and text data. We inject a noise into CLIP text feature to act as a query in image feature distribution. Visualization results for this approach are shown in Fig. 2a right, demonstrating that our method exhibits a distribution highly similar to that of image-to-retrieval results and ground truth captions, unlike traditional text-to-text retrieval methods. Indeed, when our method is applied to the existing research [30], performance improvements are observed, as shown in the supplementary (Table 1).

Prior research [30] relies solely on retrieved captions, which may include wrong information to the input caption, potentially leading to inaccurate outputs. To address this, we design a *Fusion Module* that effectively integrates both the original input and additional representations. Additionally, as shown by numerous studies [7], prompts can clarify the information provided to the language model. We extract keywords from the input caption to construct a hard prompt, which is fed to the LLM, offering explicit guidance. This approach maximizes the utility of text data, guiding the model to generate accurate and relevant captions.

Guiding caption decoder with extracted entities from an image helps the model generate an accurate description of the image. However, we find that the previous works [7, 14] show low entity detection



(a) (b)

Figure 2: (a) The distribution of CLIP embedding features corresponding to images ■, paired captions ●, retrieved captions ● for a specific image, and result of text-to-text retrieval ● and our Image-like Retrieval ● (b) Precision of extracted entities in COCO test set, total 5,000 images. If an extracted entity exists in the ground-truth caption, it counts as correct or else wrong. Three methods (Ours, ViECap[7], DETR[5]) are compared with 3 different settings. Our method is illustrated in 3.3, and ViECap uses CLIP based classifier with the source domain's vocabulary list. We follow the way SynTIC [14] uses DETR and employ the COCO vocabulary list. Due to the inaccessible vocabulary list of Flickr30k, DETR can't be compared, and ViECap uses the VGOI [36] vocabulary list in Flickr30k. Our method dominates the precision score and quantity of entities in every setting.

precision, especially when the vocabulary is large as shown in Fig. 2b. Therefore, we propose a retrieval-based entity filtering technique precisely utilizing entity information without relying on the vocabulary. During inference, we utilize retrieved sentences from images, parsing them into nouns and calculating their frequency. Then, we filter nouns with pre-defined thresholds and curate hard prompts for the text decoder. This simple method yields remarkable performance improvements.

In summary, our contributions are as follows:

- We propose a novel approach, *Image-like Retrieval*, which achieves effects similar to image-to-text retrieval in text-only training. Then, we introduce a fusion module for interaction between existing and additional representations.
- We propose an entity filtering technique in inference, *Frequency-based Entity Filtering*, enhancing the language model by filtering frequently appearing entities in retrieved captions.
- Extensive evaluations show IFCap achieves state-of-the-art performance in various benchmarks, including video captioning.

2 Related work

2.1 Text-only Captioning

The advantage of CLIP [21] has been utilized in a variety of tasks, such as image captioning, image generation, and object detection. In the realm of image captioning, text-only training research is emerging that uses only text data for learning without image data, taking advantage of the CLIP characteristic that image embeddings and text embeddings are learned to be close. DeCap [12] trains a text decoder using only textual data and introduces a support memory mechanism to project input images into the text embedding space during inference, facilitating the generation of captions. ViECap [7] recognizes the main entity of text data that comes as input and configures it as a prompt, allowing LLM to perform object-agnostic learning based on open vocabulary retrieval using CLIP.

2.2 Modality Gap

Vision language models such as CLIP aim to embed images and text closely in a shared space. However, it has been shown that these embeddings are located in two separate regions, with a significant gap between the modalities [13]. This modality gap hinders the interaction between vision and text modalities and limits the quality of generated captions. Among the notable approaches addressing this issue, CapDec [18] assumes that the image embeddings paired with text embeddings

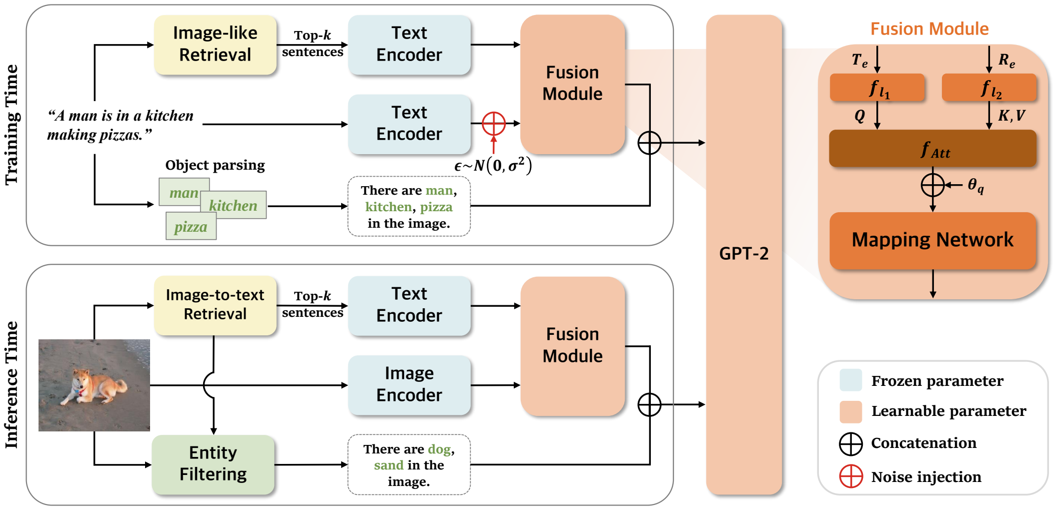


Figure 3: The overview of IFCap. During training, we extract nouns from the input text and retrieve k similar sentences using our image-like retrieval method. Extracted nouns are incorporated into a prompt template to form a hard prompt. Both the input text and retrieved sentences are encoded using the text encoder. These embeddings interact and combine through our fusion module before being fed into the LLM for sentence generation. During inference, we retrieve k sentences similar to the input image and construct a hard prompt by extracting entities via frequency-based filtering from the retrieved sentences. The sentences are encoded using a text encoder, and the input image is encoded using an image encoder, followed by input into the fusion module. The subsequent process follows a procedure similar to the training phase.

are located within a small radius around the text embeddings and mitigates the gap with noise injection. CLOSE [8] highlights the low cosine similarity between images and their paired texts and uses a hyper-parameter-scaled noise injection technique to bridge the gap.

We focus on the modality gap for retrieval from a new perspective. Our goal is to perform text retrieval similar to image-to-text retrieval, considering the modality gap. The distinction from existing methods can be observed in Fig. 2a left.

2.3 Retrieval Augmented Generation

Retrieval has been used in diverse ways in NLP. Image captioning also benefits from retrieval modules by incorporating novel objects and new information into captions, allowing access to new domains without additional training. Retrieval is applied in various ways in image captioning models. For instance, Smallcap [22] retrieves captions relevant to the input image and uses them as instructions for the text decoder. In text-only image captioning, ViECap [7] retrieves novel objects from the input image and uses them as prompts, while Knight [30] uses retrieved captions as text features.

Most retrieval methods are based on image-to-text retrieval, but text-only captioning performs text-to-text retrieval. However, during inference, the modality gap caused by the input image leads to poor performance. Our method carefully addresses this issue to improve performance by considering the gap between image and text.

3 Methods

We propose a new text-only image captioning model, IFCap, which is illustrated in Fig. 3. During training, the model only utilizes text data, as is standard for text-only training models. First, we embed the input text using a text encoder. The text embeddings are then fed into a mapping network to close the gap between different modalities. Finally, the processed embeddings go through a caption decoder to generate the output caption.

Our IFCap utilizes a simple yet powerful retrieval mechanism and addresses the modality gap between image and text with "image-like retrieval" (Section 3.1). After performing image-like retrieval, we employ a fusion layer (Section 3.2) to merge input embeddings with the retrieved features. During inference, we use the retrieved captions from the image to find accurate and detailed entities (Section 3.3).

3.1 Image-like Retrieval (ILR)

While text-to-text retrieval can be effectively performed during training, it is likely to suffer from performance degradation during inference when an image is provided as input due to the modality gap. Therefore, Image-like Retrieval (ILR) aims to perform text-to-text retrieval in a manner that resembles image-to-text retrieval outcomes, given text input. For this, we propose an approach that inserts noise into the feature space of the input text, bringing it closer to the image feature space. The augmentation process is as follows:

First, we utilize the CLIP to embed the input text t_i and the text corpus $\mathcal{T} = \{t_i\}_{i=1}^{N_c}$ with a text encoder $\mathcal{E}_T$. Then, we introduce noise $\epsilon_r \sim N(0, \sigma_r^2)$ into the embedding of input text T_i, aiming to adjust the text features to align more closely with the image feature space:

$$T_i = \mathcal{E}_T(t_i), \quad T_i^\epsilon = T_i + \epsilon_r. \tag{1}$$

Next, the retrieval step is performed using the noise-injected input text T_i^ϵ. To identify the descriptions most relevant to T_i^ϵ, the top-k descriptions are retrieved by calculating the cosine similarity between T_i^ϵ and all sentence embeddings in the text corpus. This process closely follows previous methods in image-to-text retrieval [22], with the distinction that we perform retrieval based on T_i^ϵ instead of images.

By utilizing this approach during training, we can enhance the ability of a model to provide image-like information even in a text-only training setting, thereby narrowing the modality gap and improving performance.

3.2 Fusion Module (FM)

In text-only image captioning, choosing which additional information to inject into the model and dealing with new representations with given data appropriately are important issues. To handle this problem, we use attention mechanism [27] to fuse input text features and retrieved captions features for extracting their meaningful interaction. The attention mechanism emphasizes certain important features, and due to its effectiveness, it has been widely utilized in the field of captioning [33].

We first encode input text and retrieved captions using CLIP [21] text encoder, then inject a Gaussian noise $\epsilon \sim N(0, \sigma^2)$ to input text feature for relieving the modality gap between image and text. Then we adjust the dimension of input text feature and retrieved captions feature to caption decoder's embedding space with linear layer f_{l_1} and f_{l_2} respectively, and apply cross-attention $\boldsymbol{f}_{Att}$ with T_e as query and R_e as key, then create fusion representation F_e containing input text and retrieved captions. Finally, F_e is fed into a trainable Projector, which encodes the overall contents of the given input. We can summarize this process with equations.

$$T_e = T_i + \epsilon, \quad R_e = \mathcal{E}_T(R(T_i)), \tag{2}$$

$$F_e = \boldsymbol{f}_{Att}(f_{l_1}(T_e), f_{l_2}(R_e)), \tag{3}$$

$$\boldsymbol{F} = \mathrm{Map}(F_e; \theta_q). \tag{4}$$

The noun implies intuitive and explicit information about objects in the image. For employing property of noun, we extract entities in each training text corpus and input images. We build a hard prompt h with Extracted entities $E = \{e_1, e_2, ..., e_n\}$ to make the model aware of existing entities in the image. With retrieved captions and hard prompts with entities, the model can learn the ability to generate proper captions without images. We use auto-regressive loss for tuning our projector and caption decoder. (Details about the fusion module are in Sec. 4.1).

$$L_\theta = -\frac{1}{N} \sum_{i=1}^{N} \log(y_i | \boldsymbol{F}; \boldsymbol{h}; y_{<i}; \theta). \tag{5}$$

Method	**Image Encoder**	**Text Decoder**	**COCO**				**Flickr30k**			
			B@4	M	C	S	B@4	M	C	S
CapDec [2022]	RN50x4	GPT-2$_{\text{Large}}$	26.4	25.1	91.8	11.9	17.7	20.0	39.1	9.9
DeCap [2023]	ViT-B/32	Transformer$_{\text{Base}}$	24.7	25.0	91.2	18.7	21.2	21.8	56.7	15.2
CLOSE [2022]	ViT-L/14	T5$_{\text{base}}$	-	-	95.3	-	-	-	-	-
ViECap [2023]	ViT-B/32	GPT-2$_{\text{Base}}$	27.2	24.8	92.9	18.2	21.4	20.1	47.9	13.6
MeaCap$_{\text{InvLM}}$ [2024]	ViT-B/32	GPT-2$_{\text{Base}}$	27.2	25.3	95.4	19.0	22.3	22.3	$\underline{59.4}$	15.6
Knight [2023]	RN50x64	GPT-2$_{\text{Large}}$	27.8	$\underline{26.4}$	98.9	$\underline{19.6}$	22.6	**24.0**	56.3	16.3
ICSD♠ [2023]	ViT-B/32	BERT$_{\text{Base}}$	$\underline{29.9}$	25.4	96.6	-	**25.2**	20.6	54.3	-
SynTIC♠† [2023]	ViT-B/32	Transformer$^{\text{L=4}}_{\text{H=4}}$	$\underline{29.9}$	25.8	$\underline{101.1}$	19.3	22.3	22.4	56.6	$\underline{16.6}$
IFCap	ViT-B/32	GPT-2$_{\text{Base}}$	**30.8**	**26.7**	**108.0**	**20.3**	$\underline{23.5}$	$\underline{23.0}$	**64.4**	**17.0**

Table 1: Result on the In-domain captioning including COCO test split and Flickr30k test split. Every result is copied from the original papers. ♠: Utilizes Text-to-Image generation model in the training time, †: Utilizes object detector during the training and inference time. IFCap achieves state-of-the-art in most metrics. The best number overall is in **bold** and second best in underline.

3.3 Frequency-based Entity Filtering (EF)

After retrieving l captions from an image, we use grammar parser tools (e.g., NLTK [4]) to extract nouns from the retrieved sentences and calculate the frequency of these extracted nouns as $F = [f_1, f_2, ..., f_n]$. We then select nouns that have a frequency larger than a predefined threshold and place them into a hard prompt.

Heuristic threshold: Since frequency is discrete, we can manually find the best threshold by conducting experiments with every possible threshold. This allows us to determine the global optimal threshold.

Adaptive threshold: We can use a heuristic threshold, but these thresholds are often unsuitable for different environments, and performing extensive experiments incurs unnecessary costs. Instead, we can estimate the common distribution of noun frequencies as certain probability distributions. We can assume frequencies follow $N(\mu_F, \sigma_F^2)$.

$$\tau_{\text{adap}} = \mu_F + \sigma_F. \tag{6}$$

Any nouns with a frequency larger than τ_{adap}, which places them in the upper 15%, can be considered outliers. Using this adaptive threshold, we can implement a flexible threshold that fits various settings. However, it does not guarantee global optima, leading to a trade-off relationship between heuristic thresholds and adaptive thresholds.

4 Experiments

4.1 Implementation Details

During verifying the state-of-the-art performance of our model, we use CLIP(ViT-B/32) as the image encoder and GPT2$_{\text{base}}$ [20] as the text decoder. Parameters in the image encoder are frozen during training, and the text decoder and Fusion Module are trained. We train a total of 5 epochs, learning rate as 2×10^{-5}, use scheduler for learning rate scheduler, AdamW optimizer [10], and set batch size 80. We use a single NVIDIA RTX4090 with 24GB VRAM; it takes about an hour and uses 12GB of VRAM during training.

Image-like Retrieval: We first discover adequate σ_r for Image-like Retrieval. Based on our experiment, we choose σ_r as 0.02 based on Fig. 4. We retrieve k sentences with noise-injected input text feature T_e.

Fusion Module: We project $T_e \in \mathbb{R}^d$ and $R_e \in \mathbb{R}^{d \times k}$ with f_{l_1}, f_{l_2} into $\mathbb{R}^{d_{gpt}}$, $\mathbb{R}^{d_{gpt} \times k}$ respectively where d is CLIP dimension and d_{gpt} is dimension of GPT-2 embedding space. We use projected T_e as query and R_e as key in $\boldsymbol{f}_{Att}$ layer. Finally, F_e and θ_q are concatenated and fed into the Mapping layer, which is consisted of 8 layered transformers [27].

Method	**COCO ⟹ Flickr**				**Flickr ⟹ COCO**			
	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
Knight [2023]	21.1	**22.0**	48.9	14.2	19.0	22.8	64.4	15.1
SynTIC [2023]	17.9	18.6	38.4	11.9	14.6	19.4	47.0	11.9
SynTIC-TT	19.4	20.2	43.2	13.9	**20.6**	21.3	64.4	14.3
IFCap*	17.8	19.4	47.5	12.7	14.7	20.4	60.7	13.6
IFCap-TT	**21.2**	21.8	**59.2**	**15.6**	19.0	**23.0**	**76.3**	**17.3**

Table 2: Results on the Cross-domain captioning. $-TT$: model can access to target domain's corpus during inference time. ⋆: without Entity Filtering module in the inference time. IFCap achieves state-of-the-art in most metrics.

Method	**COCO ⟹ NoCaps Val**							
	In		Near		Out		Entire	
	C	S	C	S	C	S	C	S
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]	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5
IFCap*	**70.1**	**11.2**	**72.5**	**10.9**	**72.1**	**9.6**	**74.0**	**10.5**

Table 3: Results on the NoCaps validation split. ⋆: without Entity Filtering module in the inference time. IFCap achieves state of the art in every metrics.

Method	**MSR-VTT**				**MSVD**			
	B@4	M	C	S	B@4	M	C	S
ZeroCap [2022]	2.3	12.9	5.8	-	2.9	16.3	9.6	-
MAGIC [2022]	5.5	13.3	7.4	4.2	6.6	16.1	14.0	2.9
CLMs [2022]	6.2	17.8	10.1	6.5	7.0	16.4	20.0	3.1
CapDec [2022]	8.9	23.7	11.5	5.9	7.9	23.3	34.5	3.2
EPT [2022]	3.0	14.6	11.3	-	3.0	17.8	17.4	-
Knight [2023]	25.4	**28.0**	31.9	**8.5**	37.7	**36.1**	63.8	5.0
IFCap	**27.1**	25.9	**38.9**	6.7	**40.6**	34.2	**83.9**	**6.3**

Table 4: Results on the Video captioning including MSR-VTT and MSVD. IFCap achieves state-of-the-art in most metrics.

Image-like Retrieval	**Fusion Module**	**Entity Filtering**	**COCO**			
			B@4	M	C	S
			27.2	24.8	92.9	18.2
✓			27.7	25.6	99.0	19.4
		✓	27.2	24.7	97.3	18.5
✓	✓		28.5	26.0	102.0	20.0
✓		✓	29.2	26.0	104.0	19.9
✓	✓	✓	**30.8**	**26.7**	**108.0**	**20.3**

Table 5: Ablation studies of the key components of IFCap.

Frequency-based Entity Filtering: From input image we retrieve l sentences and extracted nouns to obtain frequency F. With a predefined threshold, we filter entities and build hard prompt $\boldsymbol{h}$, providing more accurate and diverse entities to the caption decoder.

Datasets, metrics We evaluate our model in human annotated datasets. For in-domain generalization, we test our model on MS-COCO [6], Flickr30k [34] and utilize Karpathy split [9]. Also, to check the model's performance in the unseen scenarios, we use the NoCaps [1] validation set. For metrics, we use common image captioning metric CIDEr [28], SPICE [2], BLEU@n [19], and METEOR [3]. More details about datasets and metrics are included in the in the supplementary (Section D).

4.2 Text-only Captioning

We compare our model with other state-of-the-art text-only image captioning models. CapDec [18] and ViECap [7] are based on Clipcap [17]. They use predefined Gaussian noise for aligning text and image features. Similarly, CLOSE [8] uses various noise settings, and DeCap [12] uses a memory bank. And a recent approach to text-only image captioning, Knight [30] only utilizes text features with a retrieval mechanism, also MeaCap [35] processes retrieved sentences into Subject-Predicate-Object triplets and employs them as additional information. ICSD [16] and SynTIC [14] utilize text-to-image generation models like Stable Diffusion [23] for closing the gap.

4.3 In-domain Captioning

We benchmark our IFCap on in-domain setting in Table 1 including COCO and Flickr30k. We compare our methods with previous state-of-the-art in text-only image captioning. Our IFCap dominates every metric in the COCO dataset compared to models that utilize larger model [8, 30] and have complex training time [14, 16]. Also, in Flickr30k, IFCap shows decent performance in B@4 and METEOR and achieves the best scores in CIDEr and SPICE.

Transformer # Layers	Cross-Attention # Layers	COCO B@4	 M	 C	 S
1	1	23.9	24.6	86.9	17.8
	4	26.2	24.4	92.8	18.0
2	1	27.4	24.9	95.0	18.5
	4	26.4	24.9	95.5	18.4
4	1	27.4	25.5	99.7	19.1
	4	27.9	25.8	99.1	19.4
8	**1**	28.3	**26.0**	**102.0**	**20.0**
	4	**28.4**	25.7	100.6	19.5

Table 6: Ablation studies of the number of transformer layer and cross-attention layer of **Fusion Module**.

Design Choice Reference	Pre-ϵ	Post-ϵ	Retrieval	COCO B@4	 M	 C	 S
ViECap				27.2	24.8	92.9	18.2
Smallcap			✓	23.5	24.2	88.5	18.2
Knight		✓	✓	26.0	24.6	92.9	18.3
Knight + ILR	✓	✓	✓	27.2	25.0	93.9	18.3
IFCap	✓		✓	**28.5**	**26.0**	**102.0**	**20.0**

Table 7: Importance of noise injection timing of **Image-like Retrieval**. **Pre-ϵ** refers to noise injection before retrieval, and **Post-ϵ** refers to noise injection to retrieved features.

k retrieved sentences	COCO B@4	 M	 C	 S
3	28.1	25.7	100.0	19.5
5	**28.5**	**26.0**	**102.0**	**20.0**
7	28.2	**26.0**	101.7	19.8

Table 8: Ablation studies of the number of retrieved captions *k* for **Fusion Module**.

l retrieved sentences	COCO B@4	 M	 C	 S	Flickr B@4	 M	 C	 S
5	29.9	26.4	106.1	20.2	**23.5**	22.2	61.9	16.0
7	30.3	26.5	107.2	**20.3**	**23.5**	**23.0**	**64.4**	**17.0**
9	**30.8**	**26.7**	**108.0**	**20.3**	23.4	22.6	62.9	16.6

Table 9: Ablation studies of the number of retrieved sentences *l* for **Entity Filtering**.

4.4 Cross-domain Captioning

We validate IFCap's transfer ability through diverse domains, including the NoCaps validation set and cross-domain from COCO → Flickr30k and vice versa. In NoCaps, we use the same model trained in the COCO domain to test how the model recognizes unseen objects during training. In the NoCaps validation split, Our IFCap performs the best in every metric and every domain compared to previous state-of-the-art text-only image captioning models [7, 12, 18]. Also, in cross-domain setting between COCO and Flickr, IFCap wins state-of-the-art in most metrics and 2nd best in some metrics.

4.5 Video Captioning

In video captioning, we train our model in the same manner as previous experiments. First, we perform Image-like retrieval on the corpus from each video captioning dataset MSVD [31] and MSR-VTT [32]. For inference time, we sample 5 images from input video bitemporal and calculate the average of their clip image features. We also retrieved 5 sentences from each sampled image, 25 in total, and also calculated the average of clip text features per image. Most of the metrics in both datasets, IFCap, fulfills state-of-the-art performance, except METEOR.

4.6 Ablation Study

We conduct extensive experiments to identify the impact of each key component in IFCap, Image-like Retrieval (**ILR**), Fusion Module(**FM**), and Frequency-based Entity Filtering(**EF**). Also, for each component, we searched the best hyper-parameter in the COCO test split with an in-domain setting.

Key Components: We check the strength of each component via detaching from our best model, which consists of all 3 components Table 5. First, removing **FM**, we simply concatenate the input text feature and retrieved features after applying dimension mapping layer f_{l_1} and f_{l_2} and pass it to the caption decoder. Removing **EF** is simply applying entity extraction via CLIP classifier like [7] does. Demounting **ILR** makes inaccessible to retrieval features solely using input feature; hence **FM** can't exist without **ILR**. Adding more components into the baseline, we can explicitly notice performance improvement. So, using all three key components constitutes a state-of-the-art model, which is IFCap.

Image-like Retrieval: It is crucial to identify adequate timing for injecting noise into text features for successful text-to-text retrieval that imitates image-like retrieval. We can split injecting timing into

Pre-ϵ and Post-ϵ. We find our setting only injects noise before performing retrieval is the best among all possible combinations. We can verify this in Table 7. The first column of the table indicates the way how model performs retrieval, just for easy understanding of noise injection in retrieval.

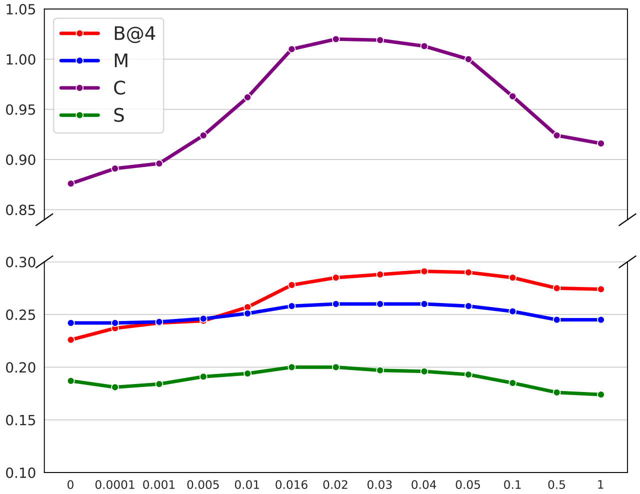


Figure 4: Hyper-parameter search for finding best σ_r used in Image-like Retrieval. All experiments are conducted with a COCO test set. The X-axis denotes σ_r^2, and the Y-axis denotes scores of commonly used captioning metrics B@4, METEOR (M), CIDEr (C), SPICE (S).

Fusion Module: We utilize a cross-attention layer and transformer layer for mapping the network. In Table 6, we try multiple combinations of each layer. The more layers we use, the more performance gain we can get until the layer of transformer is 4. The performance gain is also observed when we use 8 layers of transformer but it is so slight. Increasing the number of cross-attention layers is effective when the transformer layer is small, but the tendency does not last while the transformer layer grows. We conclude using 8 transformer layers and a single cross-attention layer shows the best. For a fair comparison, we detach the EF module. Also, the number of retrieved captions is crucial. We conduct ablation studies to find optimal k, and the result can be found in Table 8.

Frequency-based Entity Filtering: We need to choose 1) how many retrieved sentences l, to use and 2) the threshold τ, for filtering nouns for **EF** to extract accurate and diverse entities. Former can be found in Table 9, note that in different domains, optimal l may vary. For the COCO domain, using l as 9 shows the best performance, while 7 is the best in Flickr30k. More details about choosing threshold including adaptive approach can be found in the supplementary (Section B).

5 Conclusion

In this paper, we propose zero-shot captioning, IFCap, through text-only training. IFCap performs *Image-like Retrieval* to address the gap between image-to-text retrieval and text-to-text retrieval and *Frequency-based Entity Filtering* during inference time to extract frequently occurring entities from the retrieved sentences. Our method can be easily applied to various tasks and provides valuable guidance for retrieval-based methods in a text-only setting. It offers clear and precise information to LLMs without relying on a limited vocabulary. The simplicity and robustness of IFCap are demonstrated through state-of-the-art performance across various datasets in image captioning and video captioning.

6 Limitations

We demonstrate that IFCap exhibits superior performance across various image captioning and video captioning datasets compared to other zero-shot image captioning models with text-only training. However, the optimal value of ϵ_r for image-like retrieval currently requires a heuristic approach to determine. We leave the task of finding a more convenient method for determining the optimal ϵ_r as future work to further improve image captioning models with text-only training.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 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*, pages 382–398. Springer, 2016.

[3] Satanjeev Banerjee and Alon Lavie. 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, 2005.

[4] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/P04-3031`.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

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

[7] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3146, 2023.

[8] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! learning visual tasks using only language supervision. *arXiv preprint arXiv:2211.09778*, 2022.

[9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

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

[13] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[14] Zhiyue Liu, Jinyuan Liu, and Fanrong Ma. Improving cross-modal alignment with synthetic pairs for text-only image captioning, 2023.

[15] Ziyang Luo, Zhipeng Hu, Yadong Xi, Rongsheng Zhang, and Jing Ma. 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)*, pages 1–5. IEEE, 2023.

[16] Feipeng Ma, Yizhou Zhou, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. Image captioning with multi-context synthetic data, 2023.

[17] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

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

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

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

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

[22] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[24] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022.

[25] Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. Zero-shot video captioning with evolving pseudo-tokens. *arXiv preprint arXiv:2207.11100*, 2022.

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

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[29] Junyang Wang, Yi Zhang, Ming Yan, Ji Zhang, and Jitao Sang. Zero-shot image captioning by anchor-augmented vision-language space alignment. *arXiv preprint arXiv:2211.07275*, 2022.

[30] Junyang Wang, Ming Yan, Yi Zhang, and Jitao Sang. From association to generation: Text-only captioning by unsupervised cross-modal mapping. *arXiv preprint arXiv:2304.13273*, 2023.

[31] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. *Deep learning for video classification and captioning*, page 3–29. Association for Computing Machinery and Morgan & Claypool, December 2017. ISBN 9781970001075. doi: 10.1145/3122865.3122867. URL `http://dx.doi.org/10.1145/3122865.3122867`.

[32] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[34] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.

[35] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Zhengjue Wang, and Bo Chen. Meacap: Memory-augmented zero-shot image captioning. *arXiv preprint arXiv:2403.03715*, 2024.

[36] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models, 2021.