MeaCap: Memory-Augmented Zero-shot Image Captioning

Zequn Zeng[*], Yan Xie[*], Hao Zhang[†], Chiyu Chen, Bo Chen
National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, China
{zzequn99, yanxie0904, zhanghao_xidian}@163.com, {chenchiyu, bchen}@mail.xidian.edu.cn

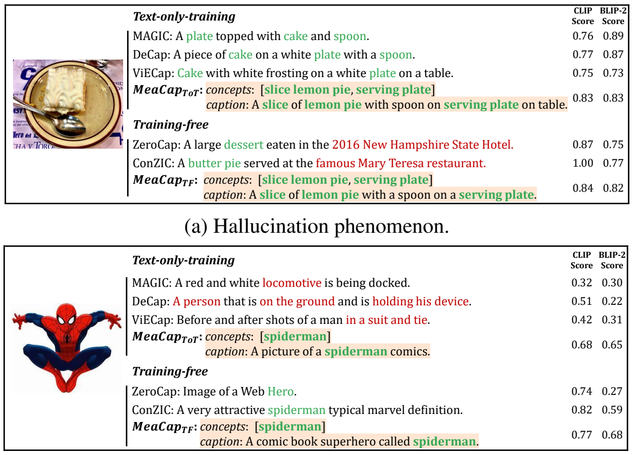
Zhengjue Wang
State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, 710071, China
wangzhengjue@xidian.edu.cn

Abstract

Zero-shot image captioning (IC) without well-paired image-text data can be divided into two categories, training-free and text-only-training. Generally, these two types of methods realize zero-shot IC by integrating pre-trained vision-language models like CLIP for image-text similarity evaluation and a pre-trained language model (LM) for caption generation. The main difference between them is whether using a textual corpus to train the LM. Though achieving attractive performance w.r.t. some metrics, existing methods often exhibit some common drawbacks. Training-free methods tend to produce hallucinations, while text-only-training often lose generalization capability. To move forward, in this paper, we propose a novel ***Me****mory-****A****ugmented zero-shot image* ***Cap****tioning framework (****MeaCap****). Specifically, equipped with a textual memory, we introduce a retrieve-then-filter module to get key concepts that are highly related to the image. By deploying our proposed memory-augmented visual-related fusion score in a keywords-to-sentence LM, MeaCap can generate concept-centered captions that keep high consistency with the image with fewer hallucinations and more world-knowledge. The framework of MeaCap achieves the state-of-the-art performance on a series of zero-shot IC settings. Our code is available at https://github.com/joeyz0z/MeaCap.*

1. Introduction

Image captioning (IC) aims to understand visual content and generate text descriptions. Using well-annotated image-text pairs, supervised models [7, 17, 23, 34, 39, 41, 48, 49, 56] have achieved promising results on typical IC benchmarks [1, 25, 32, 63]. Due to the high costs of annotation, the training sets of these benchmarks often involve limited image styles/contents, which is a hard obstacle for those supervised models to be generalized to images in the wild. To realize IC without human-annotated image-text pairs, recently, zero-shot IC has drawn increasing attention. Existing works can be mainly divided into two groups, training-free methods and text-only-training methods. Training-free approaches [53, 54, 64] realize zero-shot image-to-text generation using pre-trained models without fine-tuning. Specifically, they employ a pre-trained vision-language model like CLIP to guide a pre-trained language model (LM), such as BERT [8] or GPT-2 [44], to generate sentences that match the given image. With iterative inferences, this line of work does not require any training. Though having achieved superior generalization ability and higher CLIPscore [16], these methods show extrinsic hal-



Text-only-training	CLIP Score	BLIP-2 Score
MAGIC: A plate topped with cake and spoon.	0.76	0.89
DeCap: A piece of cake on a white plate with a spoon.	0.77	0.87
ViECap: Cake with white frosting on a white plate on a table.	0.75	0.73
MeaCap$_{ToT}$: *concepts*: [**slice lemon pie**, **serving plate**] *caption*: A **slice** of **lemon pie** with spoon on **serving plate** on table.	0.83	0.83
Training-free		
ZeroCap: A large dessert eaten in the 2016 New Hampshire State Hotel.	0.87	0.75
ConZIC: A butter pie served at the famous Mary Teresa restaurant.	1.00	0.77
MeaCap$_{TF}$: *concepts*: [**slice lemon pie**, **serving plate**] *caption*: A **slice** of **lemon pie** with a spoon on a **serving plate**.	0.84	0.82

(a) Hallucination phenomenon.

Text-only-training	CLIP Score	BLIP-2 Score
MAGIC: A red and white locomotive is being docked.	0.32	0.30
DeCap: A person that is on the ground and is holding his device.	0.51	0.22
ViECap: Before and after shots of a man in a suit and tie.	0.42	0.31
MeaCap$_{ToT}$: *concepts*: [**spiderman**] *caption*: A picture of a **spiderman** comics.	0.68	0.65
Training-free		
ZeroCap: Image of a Web Hero.	0.74	0.27
ConZIC: A very attractive spiderman typical marvel definition.	0.82	0.59
MeaCap$_{TF}$: *concepts*: [**spiderman**] *caption*: A comic book superhero called **spiderman**.	0.77	0.68

(b) Image contains world knowledge.

Figure 1. The motivation of our proposed MeaCap where the red is incorrect and green is correct. (a) Training-free methods associate the *pie* with incorrect location information, which actually get high marks in CLIPscore. This might be due to the fact that CLIP is trained on web-scale noisy image-text data. (b) Existing text-only-training (ToT) methods fail to generate *spiderman* as some training-free methods do, but the ToT version of our method (MeaCap$_{\mathrm{ToT}}$) can also do that.

[*] Equal contribution. [†] Corresponding author

lucination phenomenon, *i.e*., they tend to generate a story containing imaginary information that may not exist in the given image, as shown in Fig. 1a.

To alleviate this issue, another line of works trains or fine-tunes the text decoder based on high-quality text data without corresponding images, termed as text-only-training methods [12, 29, 40, 51, 55]. For testing images containing objects described in the training corpus, text-only-training methods generate captions objectively, achieving significant improvements w.r.t. reference-based scores such as BLEU [42], METEOR [4], and CIDEr [57]. However, due to the limited training corpus, the knowledge contained in the pre-trained LM is gradually forgotten during training, resulting in severe performance degradation on out-of-domain data, as shown in Fig. 1b. Although training on web-scale high-quality corpus is a potential solution, which hence produces extremely high computational costs.

To maintain good generalization ability to images in the wild and to get rid of unreasonable imagination, this paper proposes a novel **Me**mory-**A**ugmented zero-shot image **Cap**tioning framework, namely **MeaCap**, based on the memory-guided mechanism, which provides an alternative scheme to use captioning corpus rather than using it to train the LM. Specifically, from an external textual memory, we develop a retrieve-then-filter module to find key concepts that are highly related to the given image. Introducing our proposed memory-augmented visual-related fusion score to a keywords-to-sentence LM, CBART [15], MeaCap can generate concept-centered captions that keep high consistency with the images. This new visual-related score not only considers image-text cross-modal similarity as most zero-shot IC methods [51, 53–55, 64] do by CLIP but also considers text-text in-modal similarity by evaluating the similarity between captions and retrieved image-related memory. Our proposed MeaCap can be either training-free named $\textbf{MeaCap}_{\mathrm{TF}}$ or text-only-training named $\textbf{MeaCap}_{\mathrm{ToT}}$ by fine-tuning CBART.

Our contributions are summarized as follows:

- We employ the text-only captioning corpus as the external memory to enhance training-free zero-shot IC. To this end, We introduce a retrieve-then-filter module to extract key concepts from the memory and perform concept-centered generations by CBART to alleviate the hallucination issue of previous training-free methods.
- Based on the retrieved textual memory, we develop a memory-augmented visual-related fusion score into CBART, improving the correlation between image and generated captions while reserving the world-knowledge.
- Extensive experiments under zero-shot, in-domain, and cross-domain scenarios demonstrate our proposed memory-augmented design can significantly improve the consistency with image content in both the training-free and text-only-training settings.

2. Related work

2.1. Supervised image captioning

Supervised IC typically uses well-aligned image-text pairs and trains an encoder-decoder model. For example, some early attempts [9, 13, 58, 60] construct CNN-based encoder to extract visual features and RNN/LSTM-based decoder to generate output sentence. For better visual understanding, some methods [3, 7, 17, 18, 26, 43, 59] employ an object detector to extract attentive image regions. To encourage more interactions between two modalities, attention mechanism [7, 17, 39, 41, 48, 49] and graph neural network [61, 62] have been widely adopted.

2.2. Zero-shot image captioning

Recently, zero-shot IC has gained more and more attention, which targets at generating image captions under two cases: *i)* without any data for training named *training-free* zero-shot IC; *ii)* just using text from the captioning dataset to train the LM named *text-only-training* zero-shot IC.

Training-free methods realizes the zero-shot IC via the pre-trained vision-language model [45], to guide the generation of a pre-trained LM. Specifically, ZeroCap [54] and its extension [53] for video captioning are proposed based on the gradient-search iteration. To make the zero-shot IC controllable, ConZIC [64] is proposed by combining Gibbs-sampling with a non-autoregressive LM, improving the diversity and inference speed of IC. Although they achieve superior generalization ability with higher CLIPscore [16], they may generate some descriptions that do not appear in the image, called hallucination as shown in Fig. 1a.

Text-only-training methods train or fine-tune the text decoder just using the corpus from the captioning dataset. Concretely, after fine-tuning the off-the-shelf simCTG [52] directly on the specific corpus, MAGIC [51] and ZEROGEN [55] are proposed by introducing a CLIP-induced score to regularizes the generated process of simCTG, making the caption semantically related to a given image. By regarding the original sentence or sentence embedding as the prompt to train a LM, DeCap [29], CapDec [40] and ViECap [12] are developed by mapping the visual feature to the text feature, which is then fed into the this LM for caption generation.

Our proposed MeaCap can perform both training-free and text-only-training zero-shot IC. For the training-free setting, because we introduce the memory mechanism for key concept identification (Sec. 3.1) and guidance for LM generation (Sec. 3.3), our method can generate more accurate captions with less hallucination. For text-only-training setting, the use of the corpus as the external memory in our method can alleviate the problem of existing methods that forget the world-knowledge learned by pre-trained LM due to the corpus-specific fine-tuning.

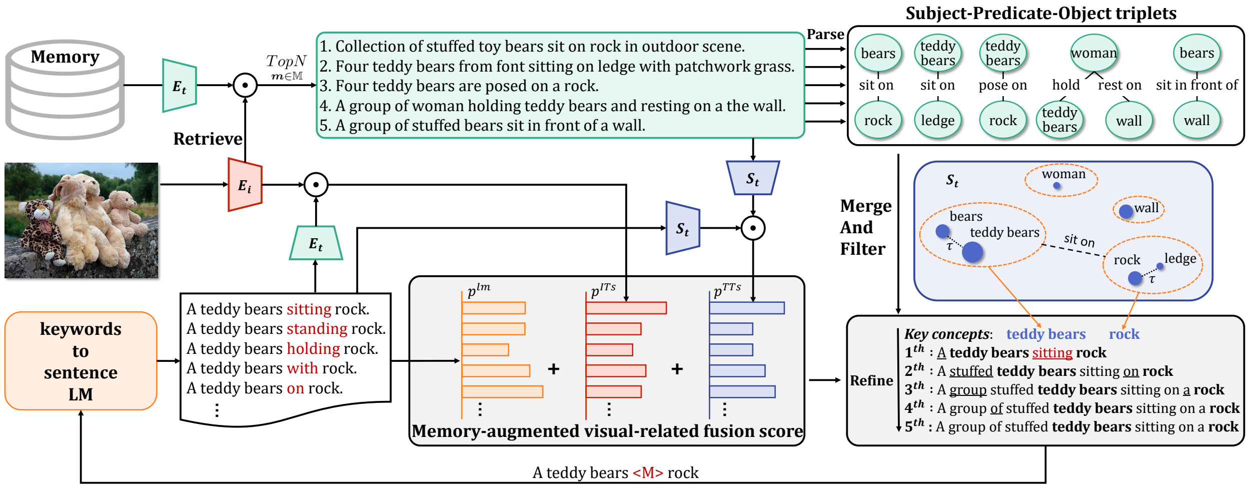

Memory
E_t
$TopN_{m \in \mathbb{M}}$
Retrieve
1. Collection of stuffed toy bears sit on rock in outdoor scene.
2. Four teddy bears from font sitting on ledge with patchwork grass.
3. Four teddy bears are posed on a rock.
4. A group of woman holding teddy bears and resting on a the wall.
5. A group of stuffed bears sit in front of a wall.
Parse
Subject-Predicate-Object triplets
bears
sit on
rock
teddy bears
sit on
ledge
teddy bears
pose on
rock
woman
hold
teddy bears
rest on
wall
bears
sit in front of
wall
E_i
E_t
S_t
S_t
Merge And Filter
S_t
woman
wall
bears
teddy bears
τ
sit on
rock
ledge
τ
keywords to sentence LM
A teddy bears sitting rock.
A teddy bears standing rock.
A teddy bears holding rock.
A teddy bears with rock.
A teddy bears on rock.
p^{lm}
p^{ITs}
p^{TTs}
+
+
Memory-augmented visual-related fusion score
Refine
Key concepts: teddy bears rock
1th : A teddy bears sitting rock
2th : A stuffed teddy bears sitting on rock
3th : A group stuffed teddy bears sitting on a rock
4th : A group of stuffed teddy bears sitting on a rock
5th : A group of stuffed teddy bears sitting on a rock
A teddy bears <M> rock

Figure 2. Overview of our proposed **MeaCap**. The overall data flow is clockwise. *i)* Given an image, we first *retrieve* Top-N relevant descriptions from the memory, which is transformed to the subject-predicate-object triplets; we merge and filter nodes to get the key concepts Sec. 3.1. *ii)* With the memory-augmented visual-related fusion score (Sec. 3.3), starting from key concepts, the keywords-to-sentence LM can complete the image description by iterative refining (Sec. 3.2). E_i, E_t, S_t are CLIP visual encoder, CLIP text encoder, and Sentence-BERT text encoder, respectively. $\odot$ denotes the cosine similarity. The p^{lm}, p^{ITs}, p^{TTs} are fluent score in Eq. (7), image-caption cross-modal similarity Eq. (8), and memory-caption in-modal similarity Eq. (9), respectively.

2.3. External memory in image captioning

It has been proven that introducing external memory is useful for various visual and language tasks, like natural language process [6, 14, 21, 22, 37], visual recognition [33, 35], image synthesis [5, 10], open-domain question-answering [20, 27], and IC included [29, 46]. For instance, SmallCap [46] is a supervised IC method that utilizes CLIP to retrieve a few relevant captions and then takes these captions as the prompt for the LM, demonstrating the memory can help LM to generate accurate captions with fewer training parameters. In zero-shot captioning, DeCap [29] trains a LM to invert the CLIP text embeddings to the corresponding sentence. It projects the CLIP visual embeddings into a weighted sum of the textual memory embeddings and takes the final textual embedding as a soft prompt to the LM to guide caption generation.

Compared with DeCap and SmallCap, which leverage the whole memory sentence as a prompt to guide generation, we propose a training-free filter that removes the noisy information to get the key concepts from retrieved textual memory. Unlike the memory in DeCap is only designed for text-only-training methods, our explicit memory design can be applied to both training-free and text-only-training scenarios and shows superior capability to generate more accurate captions.

3. MeaCap

For better zero-shot IC with less hallucination and reserving more world-knowledge, as shown in Fig. 2, we propose a novel framework called MeaCap. *i)* To solve the problem of existing training-free methods [53, 54, 64] that may bring hallucination in the captions, MeaCap identifies some key concepts from the retrieved textual memory which is highly related to the image, and performs concept-centered captioning (Sec. 3.1). *ii)* We develop a memory-augmented visual-related fusion score (Sec. 3.3), considering both image-text cross-modal similarity and text-text in-modal similarity (between textual memory and captions), which is introduced to the keywords-to-sentence LM, CBART [15] (Sec. 3.2), improving the image-caption correlations.

3.1. Retrieve-then-filter to get key concepts

The existing text-only-training zero-shot IC methods [12, 29, 40, 51, 55] usually train or fine-tune a LM on the texts from the captioning dataset, which brings more suitable descriptions with less hallucination. However, such methods make the generated captions overfit to a specific corpus, lacking the out-of-distribution generalization. Motivated by this phenomenon, instead of training or fine-tuning a LM on the texts, we just build an augmented textual memory to get the key concepts, which can then guide the zero-shot IC.

Build augmented memory. For this end, we firstly construct a large textual memory $\mathbb{M}$ which contains various visual-related sentences with abundant visual concepts. This memory is significant for removing the hallucination for the training-free case and can alleviate the knowledge-forgotten for the text-only-training case.

Retrieve image-related descriptions. Having obtained the memory, given an image **I**, we use CLIP for the evaluation of image-text similarity to retrieve Top-N image-related descriptions from the memory as $\{m_n\}_{n=1}^{N_d}$:

$$\{m_n\}_{n=1}^{N_d} = \underset{\boldsymbol{m}\in\mathbb{M}}{TopN}[\cos(E_i(\mathbf{I}), E_t(\boldsymbol{m}))], \tag{1}$$

where $E_i(\cdot)$ and $E_t(\cdot)$ denote the image and text encoder in the CLIP, respectively; $\cos(\cdot,\cdot)$ is the cosine similarity.

Subject-Predicate-Object triplets. To further reduce the impact of some less-information words in the image-related descriptions $\{m_n\}_{n=1}^{N_d}$, such as article and preposition, we use an off-the-shelf textual parser, TextGraph-Parser [31], to transform each description m_n to a text-graph g_n including multiple subject-predicate-object triplets, where subjects and objects are nodes while predicates are the relation. These nodes are regarded as candidate concepts which will be filtered and merged to form a set of the key concepts. The relations will decide the order between two concepts. We define $\{v_n\}_{n=1}^{N_c}$ as the set of all nodes from all N_d text graphs $\{g_n\}_{n=1}^{N_d}$.

Merge and filter to obtain the key concepts. As shown in Fig. 2, some nodes denoting concepts may represent the same object in the image (*e.g.*, ”bear” and ”teddy bear”), while some ones may be irrelevant to the image (*e.g.*, ”woman”), which should be merged and filtered before getting the key concepts.

i) Merge. With the help of text encoder from Sentence-BERT [47], $S_t(\cdot)$, we can obtain the concept embedding set $\{\boldsymbol{f}_n^c\}_{n=1}^{N_c}$ as $\boldsymbol{f}_n^c = S_t(v_n)$. Then we evaluate the similarity between any two concept embeddings as

$$d_{ij} = \cos(\boldsymbol{f}_i^c, \boldsymbol{f}_j^c); i,j = 1,\cdots,N_c. \tag{2}$$

Then, we set a hyper-parameter τ as the threshold, where $d_{ij} > \tau$ denotes that i-th concept and j-th concept belong to the same cluster. After this, totally we have N_v concept clusters as $\left\{c_n = \{v_i\}_{i=1}^{N_{c_n}}\right\}_{n=1}^{N_v}$, where N_{c_n} denotes the number of nodes in n-th concept cluster c_n.

2) Filter. In this step, we need to decide whether the n-th concept cluster c_n is removed or reserved. For this end, a reasonable assumption is that the word irrelevant to the image has a lower appearance in the retrieved descriptions $\{m_n\}_{n=1}^{N_d}$ in Eq. (1). Therefore, we calculate the concept-cluster frequency $CF(c_n)$ by gradually seeing whether v_i from c_n appearing in m_k as

$$CF(c_n) = \frac{\sum_{i=1}^{N_{c_n}}\sum_{k=1}^{N_d}\delta(v_i \in m_k)}{N_d} \tag{3}$$
$$\delta(v_i \in m_k) = \begin{cases} 1 & v_i \in m_k \\ 0 & v_i \notin m_k \end{cases}.$$

where $CF(c_n)$ indicates the frequency of the n-th cluster appearing in the retrieved descriptions $\{m_n\}_{n=1}^{N_d}$. Empirically, if $CF(c_n) > 0.5$, we reserve this cluster c_n and otherwise delete it. Finally, we filter out n_v key concept clusters from original N_v ones, which are highly related to images.

3) Find key concepts. Having obtained n_v key concept clusters $\{c_n\}_{n=1}^{n_v}$ where each cluster may contain multiple similar concepts, we need to identify one concept to represent this cluster. For this target, we use CLIP to select one concept from one cluster by finding the maximum image-concept similarity as

$$c_n^{key} = \max_{v_j \in c_n}\left[\cos(E_i(\mathbf{I}), E_t(v_j))\right]; n = 1,\cdots,n_v, \tag{4}$$

where c_n^{key} is the selected concept for the cluster c_n.

After these three steps, we have the set of key concepts as $\{c_n^{key}\}_{n=1}^{n_v}$ that is highly visual-related. Before using these concepts to generate captions by the following keywords-to-sentence LM, we need to decide their orders, which is realized by the relations in subject-predicate-object triplets.

3.2. Keywords-to-sentence LM

To generate a fluent visual-related caption starting from key concepts $\{c_n^{key}\}_{n=1}^{n_v}$, we employ a pre-trained lexically constrained language model, CBART [15]. Specifically, CBART is developed to generate a sentence $S = (x_1,...,x_n)$ given the ordered K keywords $\{c_i\}_{i=1}^{K}$ by maximizing the conditional probability.

$$S = \arg\max_S P(x_1,...,x_n|\{c_i\}_{i=1}^{T}), \tag{5}$$

where $x_1,...,x_n$ are words. To this end, CBART has an action encoder and a language decoder for iteratively refining the sentence starting from keywords. At t-th iteration, the encoder is responsible for predicting which word-level action (*copy*, *replacement*, and *insertion*) should be taken. In other words, the encoder takes an incomplete sentence S_t having n' words as input and outputs the corresponding action sequence $L_t = \{l_{t,1},\cdots,l_{t,n'}\}$, where $l_{t,i}$ denotes the action of i-th word at t-th iteration.

i) Copy. Copy means current word remains unchanged.

ii) Replacement. Replacement suggests the current word should be replaced. Specifically, CBART uses a mask token $<\mathrm{M}>$ to replace current word and sample a new word based on the conditional probability $p^{lm}(x_{<\mathrm{M}>}|x_{-<\mathrm{M}>})$, where $x_{-<\mathrm{M}>}$ denotes unmasked tokens.

iii) Insertion. Insertion indicates the decoder should insert a word before the current word. Similar to the replacement action, CBART inserts a $<\mathrm{M}>$ token before the current word and then samples a word from $p^{lm}(x_{<\mathrm{M}>}|x_{-<\mathrm{M}>})$.

Accordingly, the decoder can refine the sentence from S_t to S_{t+1}. Therefore, the complete encoder-decoder sentence refinement by CBART at t-th iteration can be formulated as

$$\begin{aligned} L_t =& \mathrm{LM}_{\mathrm{Encoder}}(S_t) \\ S_{t+1} =& \mathrm{LM}_{\mathrm{Decoder}}(S_t, L_t). \end{aligned} \tag{6}$$

After a few iterations, CBART will terminate the refinement when the encoder outputs a full-copy action sequence.

According to the above introduction, existing CBART does not meet our needs because for replacement and insertion actions, the word only drawn from probability by pretrained LM $p^{lm}(x_{<\mathrm{M}>}|x_{-<\mathrm{M}>})$, which just ensures the fluency but does not consider the visual-text relations.

3.3. Memory-augmented visual-related fusion score

To make the captions highly-related to the given image $\mathbf{I}$, we need a visual guidance for the generation of words in the action of insertion and replacement. Motivated by the widely-used CLIP contrastive score for evaluating the visual-text similarity, we develop a memory-augmented visual-related fusion score to adapt the original word prediction distribution of CBART to tie with the given image, considering both *i) image-text cross-modal* similarity and *ii) text-text in-modal* similarity.

Specifically, when sampling[1] a word x_i at position i, CBART first predicts a conditional probability p^{lm} and select top-K_w candidate words $\{x_{ik}\}_{k=1}^{K_w}$ with the corresponding fluent score, as:

$$p^{lm}(x_{ik}) = p^{LM}(x_{ik}|x_{-i}), k = 1, \cdots, K_w \tag{7}$$

Then K_w candidate sentences $\{s_k = (x_1, ..., x_{ik}, ..., x_n)\}_{k=1}^{K_w}$ are formed by combining candidate word x_{ik} with the context x_{-i}.

i) image-text cross-modal similarity. This similarity is denoted as p^{ITs}, which can be computed by taking candidate sentences $\{s_k\}_{k=1}^{K_w}$ and the image $\mathbf{I}$ as input to calculate the CLIP cross-modality similarity as

$$p^{ITs}(x_{ik}) = \cos(E_i(\mathbf{I}), E_t(s_k)). \tag{8}$$

i) text-text in-modal similarity. Notice that the retrieved memory $\{m_n\}_{n=1}^{N_d}$ in Eq. (1) is also image related. Therefore, we introduce a memory-augmented visual-related similarity as p^{TTs} to further improve the image-caption correlation by using Sentence-BERT text encoder S_t to evaluate the similarity between $\{s_k\}_{k=1}^{K_w}$ and $\{m_n\}_{n=1}^{N_d}$ as

$$p^{TTs}(x_{ik}) = \frac{1}{N_d}\sum_{n=1}^{N_d} \cos(S_t(m_n), S_t(s_k)). \tag{9}$$

Finally, after a weighted sum of Eq. (7), Eq. (8) and Eq. (9), we have the memory-augmented visual-related fusion score as

$$p^{fusion} = \boldsymbol{\alpha} p^{lm} + \boldsymbol{\beta} p^{ITs} + \boldsymbol{\gamma} p^{TTs} \tag{10}$$

As a result, when sampling i-th word for replacement or insertion in CBART for our model, we select the candidate word with the highest fusion score as

$$x_i = \arg\max_{x_{ik}} p^{fusion}(x_{ik}), k = 1, \cdots, K_w \tag{11}$$

Up to now, our proposed MeaCap can achieve training-free zero-shot IC with less hallucination, which is named as **MeaCap**$_{\mathrm{TF}}$ in the experiments.

Moreover, Like most of text-only-training zero-shot IC models [51, 55] that just use text to fine-tune the language model, we can also fine-tune the CBART firstly and then perform text-only zero-shot IC, which is named as **MeaCap**$_{\mathrm{ToT}}$ in the experiments.

4. Experiments

To demonstrate that **MeaCap** can efficiently achieve impressive performance in different zero-shot settings, we follow the previous works [12, 29] to conduct comprehensive experiments on *Task One*: zero-shot IC in Sec. 4.1, and *Task Two*: unpaired IC in Sec. 4.2. For each setting, we report both results of training-free version **MeaCap**$_{\mathrm{TF}}$ and the text-only-training version **MeaCap**$_{\mathrm{ToT}}$. In Sec. 4.3, we further evaluate the validity of our proposed memory-based zero-shot IC framework with other LM. In Sec. 4.4, we conduct detailed ablation studies for MeaCap.

Dataset. We conduct experiments on three widely used image captioning benchmarks, *i.e*. MSCOCO [32], Flickr30K [63], and NoCaps [1]. For MSCOCO and Flickr30K dataset, we follow previous works [7, 11, 12, 29] and use Karpathy split [19]. We use the validation set of NoCaps to evaluate the transferability of IC models trained on other datasets. Besides, for Task One, we follow previous works [29] that transfer the model from a web-scale corpus CC3M [50] to MSCOCO and NoCaps. CC3M contains three million image-description pairs collected from the web and we only use the text for building the memory or finetuning the LM.

Implementation Details. There are various pre-trained modules used in MeaCap. *i) CLIP*: we use the pre-trained VIT-B/32 CLIP. *ii) Sentence-BERT*: we use the pre-trained model from HuggingFace[2]. *iii) CBART*: we use the pre-trained model on One-Billion-Word corpus[3]. *iv) TextGraphParser*: we use the off-the-shelf textual scene graph extractor [31]. For the training-free version MeaCap$_{\mathrm{TF}}$, we concat a prefix "The image above depicts that" at the start position of the sentence. For the text-only-training version MeaCap$_{\mathrm{ToT}}$, we further fine-tune the CBART on the corresponding training corpus with AdamW [24] optimizer. For Task One, we use CC3M to serve as the memory, while for Task Two, we use the training corpus of the source dataset as the memory. More experiments with other textual memory are presented in Appendix C. We set the concept similarity threshold $\tau = 0.55$

[1] No matter for replacement or insertion, the essence is the same, *i.e*., sampling a word to replace the mask.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://www.statmt.org/lm-benchmark/

Methods	Text Corpus		MSCOCO						NoCap val (CIDEr)			
	Training	Memory	B@4	M	C	S	CLIP-S	BLIP2-S	In	Near	Out	Overall
ZeroCap [54]	✗	✗	2.6	11.5	14.6	5.5	0.87	0.70	13.3	14.9	19.7	16.6
Tewel *et al.* [53]	✗	✗	2.2	12.7	17.2	7.3	0.74	0.68	13.7	15.8	18.3	16.9
ConZIC [64]	✗	✗	1.3	11.2	13.3	5.0	**1.00**	0.76	15.4	16.0	20.3	17.5
CLIPRe [29]	✗	CC3M	4.6	13.3	25.6	9.2	0.84	0.70	23.3	26.8	36.5	28.2
DeCap [29]	CC3M	CC3M	8.8	16.0	42.1	10.9	0.76	-	34.8	37.7	49.9	39.7
MeaCap$_{\text{TF}}$	✗	CC3M	7.1	16.6	42.5	11.8	0.84	**0.81**	35.3	39.0	45.1	40.2
MeaCap$_{\text{ToT}}$	CC3M	CC3M	**9.0**	**17.8**	**48.3**	**12.7**	0.79	0.75	**38.5**	**43.6**	**50.0**	**45.1**

Table 1. Zero-shot captioning results on MSCOCO Karpathy-test split and NoCaps validations set. In, Near, and Out denote in-domain, near domain, and out-of-domain. MeaCap$_{\text{TF}}$ is the training-free version and MeaCap$_{\text{ToT}}$ is text-only training version.

Methods	MSCOCO				Flickr30K			
	B@4	M	C	S	B@4	M	C	S
	Training on image-text pairs							
Bottom-Up [3]	36.2	27.0	113.5	20.3	27.3	21.7	56.6	16.0
OSCAR [30]	36.5	30.3	123.7	23.1	-	-	-	-
VinVL [65]	40.9	30.9	140.6	25.1	-	-	-	-
ClipCap [38]	33.5	27.5	113.1	21.1	-	-	-	-
SmallCap [46]	37.0	27.9	119.7	21.3	-	-	-	-
I-Tuning [36]	34.8	28.3	116.7	21.8	25.2	22.8	61.5	16.9
	Text-only-training, zero-shot inference							
ZeroCap† [54]	7.0	15.4	49.3	9.2	5.4	11.8	16.8	6.2
MAGIC [51]	12.9	17.4	49.3	11.3	6.4	13.1	20.4	7.1
ZEROGEN [55]	15.5	18.7	55.4	12.1	13.1	15.2	26.4	8.3
CLIPRe [29]	12.4	20.4	53.4	14.8	9.8	18.2	31.7	12.0
MeaCap$_{\text{TF}}$	9.1	20.6	56.9	15.5	7.2	17.8	36.5	13.1
MeaCap$_{\text{ToT}}$	**17.7**	**24.3**	**84.8**	**18.7**	**15.3**	**20.6**	**50.2**	**14.5**

Table 2. In-domain captioning results on MSCOCO Karpathy-test split and Flickr30K Karpathy-test split. † means text-only re-implemented version from [51].

Methods	MSCOCO → Flickr30k				Flickr30k→MSCOCO			
	B@4	M	C	S	B@4	M	C	S
MAGIC [51]	6.2	12.2	17.5	5.9	5.2	12.5	18.3	5.7
CLIPRe [29]	9.8	16.7	30.1	10.3	6.0	16.0	26.5	10.2
MeaCap$_{\text{TF}}$	7.1	16.6	34.4	11.4	7.4	16.2	46.4	11.2
MeaCap$_{\text{ToT}}$	**13.4**	**18.5**	**40.3**	**12.1**	**9.8**	**17.4**	**51.7**	**12.0**

Table 3. Cross domain captioning results on MSCOCO and Flickr30K Karpathy-test split.

for CC3M memory and $\tau = 0.6$ for other memories. $N_d, K_w, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are set as $5,200, 0.1, 0.4, 0.2$ among all experiments. All experiments are conducted on a single RTX3090 GPU. We preprocess the textual corpus into text embeddings by CLIP text encoder and store text embeddings as our memory for fast retrieval. For example, retrieval on CC3M costs an average of 0.05s on RTX3090 GPU or an average of 1s on CPU. More analysis of computation costs is shown in Appendix E.

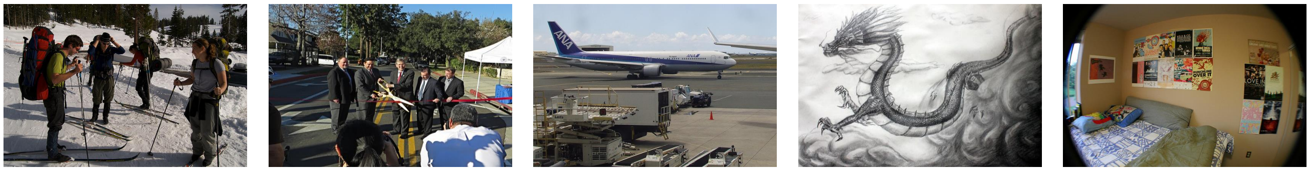
Metrics. To evaluate the accuracy of the generated caption, we use the traditional supervised metrics BLEU (B@n) [42], METEOR (M) [4], CIDEr (C) [57], and SPICE (S) [2] which compute the similarity between candidate sentences and human references. As for training-free methods, we use the CLIPScore (CLIP-S) [16] to measure the image-text similarity. Additionally, considering that CLIP-S is insensitive to the hallucination of those CLIP-based methods as shown in Fig. 1b, we employ another pre-trained large model BLIP-2 [28] to evaluate image-text similarity, *i.e.* BLIP2Score (BLIP2-S). More details are in Appendix D.

4.1. Zero-shot image captioning

In this section, we conduct zero-shot IC experiments to evaluate the ability of models to transfer from a general web-collected corpus to different downstream IC datasets.

Baselines. In this study, we compare two types of baselines. *i)* Training-free methods: ZeroCap [54], Tewel *et al.* [53] and ConZIC [64]. Those methods leverage pre-trained CLIP and freezed LM (BERT or GPT-2) to achieve zero-shot IC. *ii)* Text-only-training methods[4]: DeCap [29], which is also a memory-based method discussed in Sec. 2.3. Instead of using pre-trained LM, DeCap trains a language decoder from scratch. Besides, authors of Decap set up a baseline called CLIPRe, which generate image descriptions by retrieving the most relevant texts from memory directly. Following Decap, for MeaCap$_{\text{TF}}$, we just use CC3M as the memory, and for MeaCap$_{\text{ToT}}$, we use CC3M as the memory and also tuning the CBART. Tab. 1 shows the results on MSCOCO and NoCaps, and MeaCap achieves new state-of-the-art results. **Training-free Results.** Concretely, our training-free version MeaCap$_{\text{TF}}$ has shown superior performance on reference-based metric (B@4, M, C, S) than all previous training-free baselines, ZeroCap, Tewel *et al.* and ConZIC on both MSCOCO and NoCaps datasets by a large margin, demonstrating the effectiveness of our memory-augmented design. For reference-free metrics (CLIP-S and BLIP2-S), MeaCap$_{\text{TF}}$ achieves better results on BLIP2-S and is inferior on CLIP-S. As discussed in the introduction, previous training-free methods are favored by CLIP-S because of the hallucination phenomenon. Besides, Our MeaCap$_{\text{TF}}$ also surpasses the retrieval-based

[4] Other text-only-training methods except DeCap do not have experimented in this setting, we compared them in Task Two (Sec. 4.2)



GT:A group of skiers are gathered together as they get ready to ski.
ConZIC: A California commercial filming undergraduate college students in google photo.
ZeroCap: A video crew showing the scene of a recent study.
MAGIC: Group of people skiing down a snowy hill.
DeCap: A group of people on skis are standing in the snow

[**people**, **ski poles**]
MeaCap$_{\mathrm{TF}}$: Group of **people** with **ski poles** and snow boards outdoors.
MeaCap$_{\mathrm{ToT}}$: A group of **people** standing around with **ski poles** on.

GT:There are many men preparing to cut a red ribbon.
ConZIC: A urban roadway opening with a bicycle wheel beside the marin county park attorney office.
ZeroCap: A recent opening in San Gabriel bikeway.
MAGIC:A picture of a stop sign with a man standing behind it.
DeCap:A man in a suit and tie is giving a bike line to a business sign .

[**ribbon**]
MeaCap$_{\mathrm{TF}}$: Someone cutting the **ribbon**.
MeaCap$_{\mathrm{ToT}}$: A **ribbon** cutting ceremony on a street.

GT: A giant airplane sitting on the tarmac of an airport.
ConZIC: A japanese jet airliner and truck running brown and blue at kyoto midway terminal.
ZeroCap: A Boeing 707 in Japan.
MAGIC:A blue airplane sitting on top of a runway.
DeCap: A passenger jet parked on the tarmac at an airport .

[**airport**, **tarmac**]
MeaCap$_{\mathrm{TF}}$: An airliner parked behind a jet at **airport tarmac**.
MeaCap$_{\mathrm{ToT}}$: A large jet airliner sitting on an **airport tarmac**.

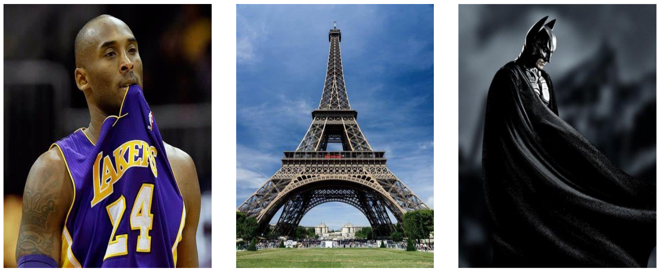
GT: A dragon flying through cloud on a sketch artwork.
ConZIC: A flying dragon genre design completed with charcoal on paper.
ZeroCap: A dragon in the form of the dragon.
MAGIC: A bird riding on a large body of water.
DeCap: A person on a big kite flying through the air.

[**dragon**]
MeaCap$_{\mathrm{TF}}$: A **dragon** sketch of the artwork.
MeaCap$_{\mathrm{ToT}}$: An illustration of a mythical **dragon** in flight.

GT: A bedroom has many posters on the wall.
ConZIC: A bedroom page surrounded by nine pink music student posters.
ZeroCap: A bedroom with posters on the walls.
MAGIC: A picture of a bedroom with a lot of pictures on the wall.
DeCap: A bedroom with a bed and many pictures on the wall .

[**bedroom**, **posters**, **wall**]
MeaCap$_{\mathrm{TF}}$: A **bedroom** with various **posters** and paintings on the **wall**.
MeaCap$_{\mathrm{ToT}}$: A **bedroom** with many **posters** on the **wall** and a bed.

Figure 3. Examples of zero-shot IC compared with other zero-shot baselines. GT denotes the Ground Truth. ConZIC and ZeroCap are training-free, while MAGIC and DeCap are text-only-training. MeaCap displays the extracted concepts in green and generated caption.



ConZIC: A lakers player peeking through his sleeves prior to retiring.
ZeroCap: A great NBA star dead.
MAGIC:A man in a wetsuit with a wetsuit with a big.
DeCap: A man that is in the middle of a game with a tennis racket .

[**basketball, shooting guard**]
MeaCap$_{\mathrm{TF}}$: The **basketball** star **shooting guard** Kobe.
MeaCap$_{\mathrm{ToT}}$: The **basketball** star of **shooting guard**.

ConZIC: Triangular tower picture at virtual domain about france website.
ZeroCap: A French landmark is the name of the song.
MAGIC: A view of a big tower with a clock on it.
DeCap: A tower that is in the center of a tall tower .

[**tower**]
MeaCap$_{\mathrm{TF}}$: The famous Eiffel **tower** in Paris.
MeaCap$_{\mathrm{ToT}}$: The famous **tower** of tourist attraction.

ConZIC: A dark knight representing a gray landscape background shaded.
ZeroCap: A Dark Knight in the film.
MAGIC: A black and white photo of a black and white zebra.
DeCap: A man that is standing in the dark.

[**character**, **batman**]
MeaCap$_{\mathrm{TF}}$: A fictional **character** known as the **batman**.
MeaCap$_{\mathrm{ToT}}$: A character of **batman** in the picture.

Figure 4. Examples of real-world knowledge. MeaCap$_{\mathrm{ToT}}$ can alleviate the world-knowledge-forgotten problem of existing text-only-training methods, such as ”batman” in the third image.

baseline CLIPRe by a large margin, indicating that only retrieving the most relevant caption is deficient in accuracy. Moreover, even compared with the text-only-training method Decap, MeaCap$_{\mathrm{TF}}$ shows superior or comparable performance on both MSCOCO and NoCap.

Text-only-training Results. To explore the potential of our MeaCap with further text-only-training on the web-scale corpus following DeCap, we also fine-tune CBART on CC3M corpus, *i.e.* MeaCap$_{\mathrm{ToT}}$. It can be observed that MeaCap$_{\mathrm{ToT}}$ significantly improves the performance, especially on NoCap. Specifically, under the same training and memory condition, MeaCap$_{\mathrm{ToT}}$ surpasses DeCap in both the MSCOCO dataset and the NoCaps dataset, showing the superiority of our method to use the external memory.

Qualitative results. Besides quantitative compare, we visualize the generated captions in Fig. 1, 3, and 4. Clearly, MeaCap can achieve better captions with more knowledge and less hallucination. More results are in Appendix F.

4.2. Task Two: Unpaired image captioning

4.2.1 In-domain captioning

To explore more potential of MeaCap for in-domain setting, where the training data, the memory, and the test set are from the same dataset, but do not use image-text pairs to build the model and memory.

Baselines. In this study, we compare with other text-only-training methods ZeroCap† [54], MAGIC [51], and ZEROGEN [55] and a retrieval-based approach CLIPRe. ZeroCap is a training-free method which is extended to text-only-training version ZeroCap† [54]. Those methods freeze the CLIP and fine-tune the LM on corresponding training texts. Under the in-domain setting, we also report both the training-free version MeaCap$_{\mathrm{TF}}$, which only employs the training text as memory, and the text-only-training version MeaCap$_{\mathrm{ToT}}$ which utilizes the training text to fine-tune CBART and serve as memory as well.

Results. As shown in Tab. 2, MeaCap$_{\mathrm{TF}}$ outperforms CLIPRe and other text-only-training baselines on C and S scores. Compared with B@4 and M scores, The C and S scores pay more attention to the accuracy of entities and relationships. The superior performance on these two scores demonstrates the high quality of our proposed memory-based retrieval-then-filter method to get the key concepts. Moreover, MeaCap$_{\mathrm{ToT}}$ outperforms all baselines by a large margin, indicating that our proposed method has greater potential with further in-domain training.

4.2.2 Cross-domain captioning

We evaluate the MeaCap for cross-domain IC with training and testing data from different datasets. We use the text from the training set as the memory for MeaCap$_{\mathrm{TF}}$ and MeaCap$_{\mathrm{ToT}}$, and fine-tune the CBART for MeaCap$_{\mathrm{ToT}}$.

Methods	MSCOCO				Flickr30K			
	B@4	M	C	S	B@4	M	C	S
DeCap [29]	24.7	25.0	91.2	18.7	21.2	21.8	56.7	15.2
CapDec [40]	26.4	25.1	91.8	-	17.7	20.0	39.1	-
ViECap [12]	**27.2**	24.8	92.9	18.2	21.4	20.1	47.9	13.6
MeaCap$_{\text{InvLM}}$	**27.2**	**25.3**	**95.4**	**19.0**	**22.3**	**22.3**	**59.4**	**15.6**
	MSCOCO → Flickr30K				Flickr30K→MSCOCO			
DeCap [29]	16.3	17.9	35.7	11.1	12.1	18.0	44.4	10.9
CapDec [40]	17.3	18.6	35.7	-	9.2	16.3	27.3	-
ViECap [12]	17.4	18.0	38.4	11.2	12.6	19.3	54.2	12.5
MeaCap$_{\text{InvLM}}$	**18.5**	**19.5**	**43.9**	**12.8**	**13.1**	**19.7**	**56.4**	**13.2**

Table 4. In-domain and cross-domain captioning results with CLIP-invert language decoder.

Results. We compare MeaCap with the text-only-training baseline MAGIC (fine-tunes GPT-2), and a retrieval-based baseline CLIPRe. Results in Tab. 3 show MAGIC suffers a performance degradation on target data, even worse than the retrieval-based method CLIPRe. Equipped with proposed memory-augmented design, MeaCap$_{\text{TF}}$ surpasses the CLIPRe on most metrics and MeaCap$_{\text{ToT}}$ outperforms all baselines, demonstrating the effectiveness of the proposed memory-augmented design.

4.3. Flexibility of MeaCap with other LM

Our proposed memory mechanism for finding key concepts in Sec. 3.1 is a plug-and-play module to further improve most of the existing text-only-training SOTA methods [12, 29, 40]. For this end, we just replace the CBART (Sec. 3.2) in MeaCap with another LM used in these methods (do not need fusion score in Sec. 3.3) described as follows.

Baselines. DeCap [29], CapDec [40] and ViECap [12] train a LM from scratch to invert the CLIP text encoder, denoted as InvLM in the following. They project the visual embeddings extracted by the CLIP visual encoder to the text embedding space of the CLIP text encoder. Then, they use InvLM to reconstruct the text from text embeddings. To generate descriptions based on our extracted key concepts, we first use a prompt template as "There are $[c_1, c_2, ..., c_n]$ in the image" to inject the concepts into a concept-aware sentence following ViECap, where c_n are the n-th concepts. After encoding the concept-aware sentence to text embeddings by CLIP text encoder, we get a concept-aware prompt. We concat the concept-aware prompt with textual embeddings as the input of InvLM, named as **MeaCap**$_{\text{InvLM}}$.

Results. Tab. 4 shows that MeaCap$_{\text{InvLM}}$ outperforms all baselines on all metrics under in-domain and cross-domain scenarios, demonstrating the effectiveness of our proposed memory-based key concepts, and also indicating its flexibility for various LM and different zero-shot settings, with detailed analysis in the Appendix A.

4.4. Ablation studies

To explore the impact of each key module in MeaCap, *i.e.* the retrieve-then-filter module (**ReF**), the image-text similarity score (**ITs**), and the text-text similarity score (**TTs**), we conduct comprehensive ablation studies on the MSCOCO dataset based on the Task One of zero-shot setting. We evaluate both the training-free version MeaCap$_{\text{TF}}$ and the text-only training version MeaCap$_{\text{ToT}}$ whose results are provided in Tab. 5. As we can see, only combined with the ReF and original LM (the first row) can surpass the only ITs results in the second row (ITs is the only visual guidance of previous training-free methods by CLIP), indicating the key concepts extracted by the ReF module are critical for zero-shot IC. The third row shows that combining ReF with ITs yields more improvements than individual modules alone. Finally, by incorporating the TTs, the performance is further improved, highlighting the efficacy of the memory-augmented visual-related fusion score. We conduct analysis of the effect of memory in Appendix B.

Methods	**ReF**	**ITs**	**TTs**	MSCOCO			
				B@4	M	C	S
MeaCap$_{\text{TF}}$	✔	✗	✗	5.0	13.3	31.1	5.6
	✗	✔	✗	1.8	9.7	12.7	4.8
	✔	✔	✗	5.7	13.6	38.6	8.5
	✔	✔	✔	7.1	16.6	42.5	11.8
MeaCap$_{\text{ToT}}$	✔	✗	✗	7.9	14.9	37.1	10.4
	✗	✔	✗	3.2	9.9	17.3	5.2
	✔	✔	✗	8.1	15.6	44.7	11.1
	✔	✔	✔	9.0	17.8	48.3	12.7

Table 5. Ablation studies on zero-shot IC. ReF, ITs, TTs denote the retrieve-and-filter module, ITs (8) and TTs (1) are image-text and text-text similarity from memory-augmented visual-related score.

5. Conclusion

In this paper, we propose a novel memory-augmented zero-shot IC framework, MeaCap. We introduce a retrieve-then-filter module to extract key concepts from external textual memory. Based on the retrieved textual memory, we further develop a memory-augmented visual-related fusion score to guide the generation of captions. Combined with CBART, we can generate concept-centered descriptions to alleviate the hallucination of previous training-free methods and enhance the accuracy of text-only-training methods. Extensive experiments on various zero-shot captioning settings show that MeaCap outperforms previous methods.

6. Acknowledgements

H. Zhang acknowledges the support of NSFC (62301384), and Excellent Young Scientists Fund (Overseas). Z. Wang acknowledges the support of NSFC (62301407). B. Chen acknowledges the support of the NSFC (U21B2006); Shaanxi Youth Innovation Team Project; 111 Project Grant (B18039); Fundamental Research Funds for the Central Universities QTZX22160.

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. 1, 5

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 6

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2, 6

[4] 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. 2, 6

[5] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. 3

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022. 3

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 1, 2, 5

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[10] Zhibin Duan, Lv Zhiyi, Chaojie Wang, Bo Chen, Bo An, and Mingyuan Zhou. Few-shot generation via recalling brain-inspired episodic-semantic memory. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[11] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18009–18019, 2022. 5

[12] 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. 2, 3, 5, 8

[13] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language cnn for image captioning. In *Proceedings of the IEEE international conference on computer vision*, pages 1222–1231, 2017. 2

[14] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. 3

[15] Xingwei He. Parallel refinements for lexically constrained text generation with bart. *arXiv preprint arXiv:2109.12487*, 2021. 2, 3, 4

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1, 2, 6

[17] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019. 1, 2

[18] Lun Huang, Wenmin Wang, Yaxian Xia, and Jie Chen. Adaptively aligned image captioning via adaptive attention time. *Advances in neural information processing systems*, 32, 2019. 2

[19] 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. 5

[20] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020. 3

[21] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019. 3

[22] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. *arXiv preprint arXiv:2010.00710*, 2020. 3

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

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

[26] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17979, 2022. 2

[27] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, 2019. 3

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[29] 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. 2, 3, 5, 6, 8

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 6

[31] Zhuang Li, Yuyang Chai, Terry Zhuo Yue, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*, 2023. 4, 5

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 5

[33] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 3

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[35] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6959–6969, 2022. 3

[36] 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. 6

[37] Yuxian Meng, Shi Zong, Xiaoya Li, Xiaofei Sun, Tianwei Zhang, Fei Wu, and Jiwei Li. Gnn-lm: Language modeling based on global contexts via gnn. *arXiv preprint arXiv:2110.08743*, 2021. 3

[38] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 6

[39] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022. 1, 2

[40] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 2, 3, 8

[41] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020. 1, 2

[42] 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. 2, 6

[43] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. Look back and predict forward in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8367–8375, 2019. 2

[44] 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. 1

[45] 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. 2

[46] 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. 3, 6

[47] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 4

[48] Idan Schwartz, Alexander Schwing, and Tamir Hazan. High-order attention models for visual question answering. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2

[49] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019. 1, 2

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, 2018. 5

[51] 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. 2, 3, 5, 6, 7

[52] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561, 2022. 2

[53] 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. 1, 2, 3, 6

[54] 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. 1, 2, 3, 6, 7

[55] Haoqin Tu, Bowen Yang, and Xianfeng Zhao. Zerogen: Zero-shot multimodal controllable text generation with multiple oracles. *arXiv preprint arXiv:2306.16649*, 2023. 2, 3, 5, 6, 7

[56] 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. 1

[57] 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. 2, 6

[58] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[59] Li Wang, Zechen Bai, Yonghua Zhang, and Hongtao Lu. Show, recall, and tell: Image captioning with recall mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12176–12183, 2020. 2

[60] 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. 2

[61] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 2

[62] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 2

[63] 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. 1, 5

[64] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23465–23476, 2023. 1, 2, 3, 6

[65] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 6