Retrieval-Augmented Text-Only Training for Image Captioning

Anonymous ACL submission

Abstract

Image captioning has drawn remarkable attention from the natural language processing and computer vision fields. Aiming to reduce the reliance on curated data, several studies have explored image captioning without relying on any humanly-annotated image-text pairs, although existing methods are still outperformed by fully supervised approaches. This paper proposes TTLLCap, i.e. a text-only training method for image captioning, based on prompting a pre-trained language model decoder with information obtained from CLIP representations of the inputs. Specifically, we experimented with the combined use of (a) retrieved examples of captions, (b) relevant concepts for the input, and (c) latent vector representations. Through extensive experiments, we show that TTLLCap outperforms previous training-free methods, and is also competitive with other text-only training methods. We also analyze the impact of different choices regarding the configuration of the retrieval-augmentation component. The source code supporting our experiments is available from a public GitHub repository[1].

1 Introduction

Image captioning concerns generating descriptions for input images. The task has drawn remarkable attention from the natural language processing and computer vision fields, due to its wide applications. Current captioning models based on the encoder-decoder framework have achieved tremendous progress in advancing the state-of-the-art (Alayrac et al., 2022; Li et al., 2023b). These models are usually trained with full supervision, relying on large-scale humanly-annotated training data (i.e., curated image-caption pairs) whose availability depends on expensive labeling work.

Aiming to improve data- and parameter-efficiency, several studies have proposed the use of encoder-decoder models that re-use off-the-shelf pre-trained vision encoders such as CLIP (Radford et al., 2021) and language decoder models such as GPT2 (Radford et al., 2019), keeping the parameters of these components frozen and only training a mapping between the two. In some cases, authors have proposed to leverage retrieval-augmented generation to further improve efficiency, while allowing for training-free domain transfer and the exploration of captioning data in a training-free fashion (Ramos et al., 2023c). To further mitigate the data needs and improve the generalization in real-world scenarios, several studies have explored captioning without relying on any humanly-annotated image-text pairs. These methods can be divided into two groups, namely training-free and text-only training methods. Training-free approaches realize zero-shot captioning using pre-trained models without fine-tuning (e.g., a pre-trained vision-language model like CLIP is used to guide a pre-trained language model such as GPT2, to generate sentences that match the given image). In turn, text-only training methods fine-tune the decoder based on high-quality text data, without relying on corresponding images during training. Despite significant advances, existing training-free methods are given to hallucination problems, and while text-only training can achieve strong results, it is still outperformed by fully supervised approaches.

This paper proposes Text-only Training with Latents plus Language prompt-based Captioning (TTLLCap), an improved text-only training method that combines ideas from several previous studies, namely retrieval-augmented generation as in SmallCap and LMCap (Ramos et al., 2023c,b), a prompting strategy similar to that of the Socratic models framework (Zeng et al., 2022), and also the idea of decoding latent representations from the CLIP model (Gu et al., 2022; Nukrai et al., 2022; Qiu et al., 2024; Wang et al., 2024). Instead of simply using a corpus of textual captions to train a decoder

[1] https://github.com/to-be-made-available

model, we also rely on the text-only training corpus as a retrieval datastore, and use efficient Low-Rank Adaptation (LoRA) to fine-tune the decoder towards generating captions conditioned on a combination of (a) similar captions obtained through retrieval, (b) relevant concepts for the input, and (c) a latent representation of the input. The three aforementioned elements are all obtained from CLIP representations (either from textual captions, during training, or from the input image, during inference), using a simple strategy to circumvent the known CLIP modality gap (Gu et al., 2023). Combining retrieval-augmentation with LoRA training also mitigates the forgetting of the knowledge contained in the pre-trained language model decoder.

Experiments with MSCOCO (Lin et al., 2014) and NoCaps (Agrawal et al., 2019) show that TTLLCap outperforms several previous training-free and text-only training methods, particularly in terms of the generalization capabilities in NoCaps. The retrieval of similar captions is indeed the most impactfull aspect in the proposed combination, and we also analyzed the impact of different choices regarding the configuration of the retrieval-augmentation component, including the number of retrieval examples and trade-offs between retrieved caption similarity and diversity.

2 Related Work

This section reviews relevant previous work related to the proposed approach.

2.1 Zero-Shot Image Captioning

Several previous studies have proposed to re-purpose text-image matching models, such as CLIP (Radford et al., 2021), to generate image captions without any task-specific training (Tewel et al., 2022; Zeng et al., 2023; Ramos et al., 2023b). For instance, ZeroCap (Tewel et al., 2022) uses a pre-trained CLIP model together with a GPT2 language model, being truly zero-shot in the sense that the only optimization that is considered is performed *ex post facto* in the activation space during decoding, without re-training or fine-tuning the model parameters. In particular, ZeroCap uses a customized decoding algorithm, in which the context cache (i.e., all the key and value vectors in the self-attention modules) is updated with the guidance of CLIP and GPT2, for every prediction step.

Other authors have instead proposed zero-shot strategies that do not directly involve the use of image data, instead relying on textual information alone. For instance Zeng et al. (2022) proposed the Socratic models framework, where different pre-trained models communicate via zero-shot or few-shot prompting, without any multimodal training. For the task of image captioning, the GPT-3 language model can be prompted with information about the input image (e.g., information about the number of people present in the image, places, objects, and general classes associated to the image), as obtained with a pre-trained CLIP model. In turn, Ramos et al. (2023b) proposed LMCap, building on similar ideas to those of Socratic models (i.e., LMCap is an image-blind method that generates captions only with basis on textual information provided as input). In this case, CLIP is first used to retrieve captions from similar images, and these captions are then combined into a prompt for a GPT2 language model decoder.

2.2 Text-Only Training for Image Captioning

Instead of zero-shot approaches, several authors have instead explored text-only training for image captioning, e.g. learning a decoder that generates captions from a frozen CLIP text encoder, and using only textual information, unpaired to any images, for model training (Su et al., 2022; Nukrai et al., 2022; Gu et al., 2022; Tam et al., 2023; Li et al., 2023c; Wang et al., 2023; Qiu et al., 2024; Liu et al., 2024a; Wang et al., 2024). The training objective is thus, for a given textual corpus, the reconstruction of each input text from a textual embedding vector produced with CLIP, whereas at inference time we provide the model with a CLIP embedding for an image and the decoder generates the corresponding caption. To rectify the gap between the textual and visual CLIP embedding spaces (i.e., previous studies have shown that the text and image vectors from CLIP can be far apart and for instance, on MSCOCO captions, the average cosine similarity between an image and paired caption is only 0.26, while the average similarity between two unrelated captions is 0.35 (Liang et al., 2022)), most of these studies have explored some form of noise injection during training (e.g., Gaussian noise can be added to the text embeddings produced by CLIP (Gu et al., 2022; Nukrai et al., 2022; Qiu et al., 2024; Wang et al., 2024), in order to define a ball, in the embedding space, that should map to the same image).

Noting that the weak visual guidance in the paradigm described in the previous paragraph can often lead to a modality bias (i.e., the language prior in the language model dominates the decoding process, often leading to descriptions that are unrelated to the corresponding images), some previous studies (Wang et al., 2022; Fei et al., 2023; Zeng et al., 2024; Wang et al., 2024) have explored methods that rely on initializing the decoder with hard textual prompts that encode visual concepts (e.g., nouns extracted from texts during training, or entities retrieved from the input image during inference). The idea is to better guide the language model towards the visual entities, this way enabling more coherent caption generation.

More recently, some authors have proposed approaches that address the modality gap through synthetic image-text pairs (Ma et al., 2024; Liu et al., 2024b). For instance Liu et al. (2024b) used a pre-trained text-to-image model to generate images for a training set of textual captions. Each generated image, corresponding to a training text, is encoded in the CLIP embedding space, and a lightweight decoder model is trained to generate captions from the CLIP representations. Additionally, salient objects in images are recognized with a pre-trained detection model, and the corresponding textual labels are used as hard prompts that enhance the learning of the modality alignment.

2.3 Retrieval-Augmented Image Captioning

The idea of extending the information encoded within language model parameters, through non-parametric knowledge retrieved from datastores (i.e., external memories), has been used extensively in different NLP tasks. The success of retrieval-augmented generation has also inspired some recent studies in image captioning (Sarto et al., 2022; Ramos et al., 2023a,c,b; Yang et al., 2023; Ramos et al., 2024; Li et al., 2023a; Sarto et al., 2024). For instance SmallCap (Ramos et al., 2023c) corresponds to an encoder-decoder model in which a language decoder is prompted with captions retrieved from a datastore through the use of CLIP. The model itself uses a frozen CLIP encoder, a frozen GPT2 decoder, and a cross-attention layer that is trained to map across modalities. The authors of SmallCap also developed the aforementioned LMCap model (Ramos et al., 2023b), which uses a similar strategy for prompting GPT2, but in which only the textual information is used to generate the output captions (i.e., this approach does not involve cross-attention towards the representations produced with a vision encoder, corresponding to an image-blind approach).

Despite its potential, the power of retrieval-augmented generation is highly dependent on configuration options. Considering general NLP applications, some previous studies have analyzed the sensitivity of retrieval-augmented generation to aspects such as the number of retrieved instances, the ordering of the retrieved instances, or the quality and diversity of the retrieval results (Hsia et al., 2024; Cuconasu et al., 2024). Still, within the specific context of multimodal tasks like image captioning, significantly less work has looked into these issues (Peng et al., 2023; Yang et al., 2024). Some authors have noted that retrieved examples adequately describing the salient image objects, with simpler language patterns, seem to improve results. Retrieving examples with high similarity towards the input seems to be more important than the retrieval of diverse examples, while at the same time excessive similarity can lead models to create short-cut inferences from the retrieved instances, potentially misleading generation with low-quality captions. However, we believe that additional studies are needed in order to confirm these statements.

3 Proposed Method

The proposed approach, which we named Text-only Training with Latents plus Language prompt-based Captioning (TTLLCap), combines retrieval-augmented generation as in SmallCap and LMCap (Ramos et al., 2023c,b), together with a prompting strategy similar to that of the Socratic models framework (Zeng et al., 2022), and also the idea of decoding latent representations from CLIP, after noise injection to avoid the modality gap (Gu et al., 2022; Nukrai et al., 2022; Qiu et al., 2024; Wang et al., 2024). The method is based on the use of textual information during training, whereas during inference it uses a textual prompt together with visual information encoded with CLIP.

Specifically, during training, examples of image captions are first encoded with the CLIP text encoder into a vector representation, to which we add Gaussian noise in order to close the CLIP modality gap, spreading out the text vectors so that they better overlap with what would be the corresponding image vectors. The resulting vector representations are used to search for other similar captions in a datastore of examples, and also to build a Socratic prompt with information about the corresponding image (e.g., information about the number of people present in the image, places, objects, and general classes associated to the image). The retrieved

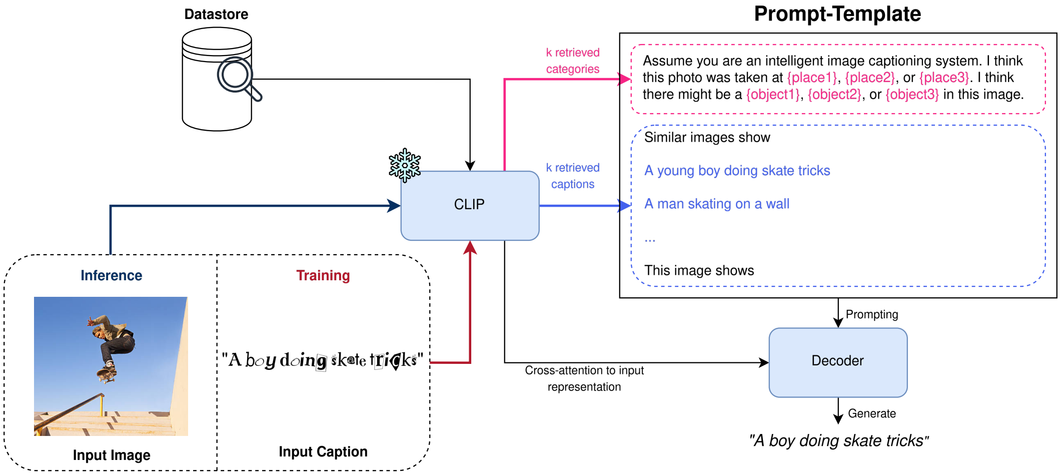

Datastore
Prompt-Template
k retrieved categories
Assume you are an intelligent image captioning system. I think this photo was taken at {place1}, {place2}, or {place3}. I think there might be a {object1}, {object2}, or {object3} in this image.
Similar images show
A young boy doing skate tricks
A man skating on a wall
...
This image shows
k retrieved captions
CLIP
Inference
Training
"A boy doing skate tricks"
Input Image
Input Caption
Cross-attention to input representation
Prompting
Decoder
Generate
"A boy doing skate tricks"

Figure 1: Overview on Text-only Training with Latents plus Language prompt-based Captioning (TTLLCap).

captions and the Socratic prompt are combined together, and used as input to a language model decoder. This decoder is extended with cross-attention layers, which consider single/individual key and value vectors derived from the input vector representation that was also used to perform retrieval. The parameters of CLIP and of the language model decoder are kept frozen during training, but we train the cross-attention layers and also Low-Rank Adaptation (LoRA) layers (Hu et al., 2021) added to the decoder. By using LoRA instead of full-parameter fine-tuning, besides reducing computational costs, we hope to better retain the general knowledge captured in the pre-trained language model, mitigating catastrophic forgetting.

At inference time, given an input image, CLIP is used to encode the visual contents into a vector representation, which is used (a) as input to the cross-attention operations, (b) to build the Socratic prompt, and (c) to find relevant captions in the datastore. The retrieved captions are used to complement the Socratic prompt, and this textual information is provided to the language model, in order to condition the generation of the caption. The main aspects of our approach are shown in Figure 1 and further detailed next.

3.1 Building CLIP Representations

We use a CLIP ViT-L/14 model[2], whose parameters are kept frozen during training, to encode either textual captions (i.e., during training) or the input images (i.e., during inference) into vectors with 768 dimensions. The vectors support the retrieval of relevant information for the input, and also the conditional generation of a caption.

Following previous studies (Gu et al., 2022; Nukrai et al., 2022; Qiu et al., 2024; Wang et al., 2024), we assume that there is a modality gap associated to the CLIP representations, and that the visual embedding corresponding to a text embedding lies somewhere within a ball of small radius ϵ around the text embedding. We would like all text embeddings in this ball to match the same caption, which should also correspond to the visual content mapped to this ball. To implement this intuition, at the same time also improving generalization through additional diversity in the training data, we add zero-mean Gaussian noise multiplied by a random scalar value (i.e., $n \sim \text{Unif}(0, 1) \times \mathcal{N}(0, 1)$) to the text embeddings produced by CLIP, which is equivalent to a random Gaussian distribution with random variances. This approach was introduced by Gu et al. (2023) and, while being relatively simple, was shown to be more effective in closing the modality gap than a simple Gaussian distribution. The exploration of more advanced strategies (e.g, methods like those from Liu et al. (2024a) or Wang et al. (2024), e.g. involving tuning with a small set of image-caption pairs), is left for future work.

3.2 Retrieval of Caption Exemplars

The CLIP ViT-L/14 model is also used for text-text (i.e., during training) or image-text (i.e., during inference) retrieval. Besides encoding the inputs, as described in the previous section, this model is also used to encode a large set of diverse caption exemplars from an external datastore, which is indexed offline with the FAISS nearest-neighbor search library (Johnson et al., 2019), using the index named `IndexIVFFlat` with parameters set to 32 probes at query time, and 256 inverted lists.

Given the encoded data, for each input instance, CLIP is used to retrieve the K most simi-

[2] https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K

lar captions from the datastore. Following SmallCap (Ramos et al., 2023c) and LMCap (Ramos et al., 2023b), most of our experiments used $K = 4$ retrieved captions. The retrieved captions will serve to guide a language model decoder as an example of what the predicted caption should resemble, through the use of a prompt and as described next.

3.3 Building the Prompt

The core aspect in our approach concerns prompting the decoder with retrieved information. Specifically, the top-K retrieved captions are used to fill slots in a prompt template with the following form: `Similar images show {caption₁} ... {captionₖ}. This image shows ___.`

Extending the retrieval prompt, and taking inspiration on the Socratic image captioning approach (Zeng et al., 2022), we also experimented with a template taking the following form: `Assume you are an intelligent image captioning system designed to produce accurate and concise image descriptions. The input image is perhaps showing a {place₁}, {place₂}, or {place₃}. The image likely features an {object₁}, {object₂}, or {object₃}. Similar images show {caption₁} ... {captionₖ}. This image shows ___.`

In the Socratic case, CLIP representations (i.e., obtained from the textual caption, during model training, or the input image, at the inference stage) are used to perform zero-shot detection from large pre-existing libraries of class names, and the template slots are filled with the top-K detected categories (i.e., the top-3 objects and the top-3 places, following the original procedure). The different place categories are collected from Places356 (Zhou et al., 2016) and the object categories from Tencent ML-Images (Wu et al., 2019).

3.4 Training the Language Model

Most of our experiments use GPT2-base as the decoder, connecting it to the CLIP encoder with multi-head cross-attention, through which each layer of the decoder attends to a single vector corresponding to the CLIP representation of the input. Following SmallCap (Ramos et al., 2023c), we control the number of trainable parameters through the dimensionality of the projection matrices in the cross-attention layers, which we denote as d. For GPT2-base, whose hidden representations have a dimensionality of 768 and which involves $h = 12$ cross-attention heads, d defaults to 64 (i.e., the dimensionality of the hidden representations divided by the number of heads). In our experiments, we scale this value down by a factor of four.

The decoder receives the textual prompt, described in the previous section, as input tokens, and it then generates a caption conditioned on the CLIP representation and the prompt. During inference, decoding uses the beam search algorithm with a small beam size of 3 (Cohen and Beck, 2019).

During model training, the weights in the cross-attention layers, and also Low-Rank Adaptation (LoRA) layers added to the query, key and value layers of each self-attention block of the decoder (Hu et al., 2021), are optimized by minimizing the cross-entropy loss towards predictions for the correct tokens in the ground-truth caption, considering the teacher-forcing strategy. LoRA considers hyper-parameters equal to 32 for both α and rank. Additionally, we use an updated version of LoRA, named Rank-stabilized LoRA (rsLoRA), where the adapters are scaled by a factor of $\frac{\alpha}{\sqrt{r}}$ instead of $\frac{\alpha}{r}$, which stabilizes the adapters (Kalajdzievski, 2023). We use the AdamW optimizer (Kingma and Ba, 2014) with an initial learning rate of 1e-4 and a batch size of 64. Training runs for 10 epochs and we use the epoch checkpoint with the best CIDEr score on a held-out validation set. Training with GPT2-base takes up to 9 hours on a single NVIDIA RTX A6000 GPU, using 16 GB of the available memory.

4 Experimental Evaluation

We now describe the experimental evaluation of the proposed approach.

4.1 Datasets and Metrics

Our experiments used two English datasets to assess captioning quality, namely MSCOCO (Lin et al., 2014) and NoCaps (Agrawal et al., 2019).

MSCOCO is a commonly used dataset for assessing image captioning, object detection, and segmentation. We used the Karpathy splits, with 113k/5k/5k images for training, validation, and testing, respectively. Each image is annotated with at least 5 human-generated captions, and we used the textual captions from the training split to train our model. In turn, NoCaps contains 15k images with nearly 400 additional novel classes not represented in MSCOCO, which can be used to evaluate novel object captioning performance.

Similarly to SmallCap (Ramos et al., 2023c), the retrieval datastore features captions from MSCOCO, and from sources such as the Conceptual (i.e., CC3M and CC12M) datasets (Sharma et al., 2018; Changpinyo et al., 2021) and SBU captions (Ordonez et al., 2011), totaling approximately 13 million image descriptions.

For evaluation, we compute the standard metrics of BLEU-1 (B@1), BLEU-4 (B@4) (Papineni et al., 2002), METEOR (M) (Denkowski and Lavie, 2014), and CIDEr (C) (Vedantam et al., 2015), using the MSCOCO evaluation package[3].

4.2 Captioning Results

Table 1 presents experimental results on the MSCOCO and NoCaps datasets, comparing TTLLCap against previous approaches corresponding to training-free (i.e., top set of rows) and text-only training methods (i.e., the set of rows in the middle). The table also shows results for ablated versions of the complete TTLLCap method, respectively corresponding to the following configurations:

- A version similar to the original LMCap method (Ramos et al., 2023b), where the decoder is prompted with retrieved captions only, without any training. This experiment used the OPT-IML-1.3B[4] decoder fine-tuned to follow instructions (Iyer et al., 2022), instead of the less capable GPT2-base model;
- Versions that extend the previous LMCap setting, either by considering the combination of retrieved captions with a Socratic prompt, or by training the decoder using LoRA, in this last case using a GPT2-base model;
- A version that involves training the GPT2 decoder (using LoRA) with cross-attention towards the CLIP representations, using a simple prompt (i.e., the phrase *This image shows*) that does not involve retrieved exemplars;
- Versions that extend the previous setting, using retrieved captions only or combining retrieval with the Socratic prompt;
- Versions featuring an optimized retrieval setting, without the Socratic prompt, in which K=6 or K=8 retrieved captions are re-ranked with basis on the similarity between them, promoting cohesiveness – see the additional experiments reported on Section 4.3;
- A version that is similar to the previous optimized setting, but featuring the larger GPT2-medium decoder model.

The obtained results show that TTLLCap outperforms all zero-shot approaches, including LMCap (Ramos et al., 2023b) and the Socratic framework (Zeng et al., 2022). Better results are obtained when considering retrieval re-ranking with basis on cohesiveness (see Section 4.3 for a deeper discussion on this aspect), but using a larger decoder failed to significantly improve results. The exploration of even larger decoders is left for future work, although the overall results suggest that higher gains can perhaps come from improved approaches for handling the CLIP modality gap, in comparison to larger decoders.

Our reproduction of LMCap, which does not involve any training, achieved lower results than those reported by Ramos et al. (2023b), but this is perhaps due to the smaller decoder and to the fact that we do not include a final selection of the candidate caption with highest CLIP similarity towards the input image, after beam search decoding.

When compared to other text-only training approaches, TTLLCap clearly surpasses well-known methods like MAGIC (Su et al., 2022) and DeCap (Li et al., 2023c), performing similarly to other recent approaches on the MSCOCO dataset, but being outperformed by methods that involve the use of synthetic images for training (Liu et al., 2024b), or more advanced strategies for addressing the CLIP modality gap (e.g., using multi-variate Gaussian distributions estimated from small sets of image-caption pairs (Wang et al., 2024), instead of our approach which does not adjust the parameters of a Gaussian distribution). Combining our retrieval-augmented approach with a better method for handling the CLIP modality gap would likely further improve results. Still, in the NoCaps dataset and particularly in the out-of-domain instances, TTLLCap already significantly surpasses all previous approaches, with results showing that the use of retrieved captions contributes significantly to improved generalization capabilities.

The assessment of different configurations confirms that the use of retrieved captions is indeed the component with the highest impact on performance, with the Socratic prompt, or the use of cross-attention towards the CLIP representations, having only a small influence on the results. In the case of the Socratic prompt, it is interesting to note that it lead to improved results when no training is involved, and to slightly worse results otherwise. We noticed that the text-text retrieval strategy, used during TTLLCap training to build

[3] https://github.com/tylin/coco-caption

[4] https://huggingface.co/facebook/opt-iml-1.3b

			MSCOCO				NoCaps (CIDEr)			
Method	Encoder	Decoder	B@1	B@4	M	C	In	Near	Out	Overall
ConZic (Zeng et al., 2023)	ViT-B/32	BERT-base	–	1.3	11.2	13.3	–	–	–	–
ZeroCap (Tewel et al., 2022)	ViT-B/32	GPT2-medium	49.8	7.0	15.4	34.5	–	–	–	–
Socratic (Zeng et al., 2022)	ViT-L/14	GPT-3	–	6.9	15.0	44.5	–	–	–	–
MeaCap$_{\text{TF}}$ (Zeng et al., 2024)	ViT-B/32	CBART	–	9.1	20.6	56.9	35.3	39.0	45.1	40.2
LMCap (Ramos et al., 2023b)	ViT-H-14	XGLM-2.9B	–	19.9	22.0	75.9	–	–	–	–
MAGIC (Su et al., 2022)	ViT-B/32	GPT2-small	56.8	12.9	17.4	49.3	–	–	–	–
DeCap (Li et al., 2023c)	ViT-B/32	Transformer	–	8.9	17.5	50.6	41.9	41.7	46.2	42.7
CLM (Wang et al., 2022)	–	GPT2	59.3	15.0	18.7	55.7	–	–	–	–
MacCap (Qiu et al., 2024)	ViT-B/32	OPT-1.3B	61.4	17.4	22.3	69.7	–	–	–	–
WS-ClipCap (Tam et al., 2023)	–	GPT2	65.5	22.1	22.2	74.6	–	–	–	–
MeaCap$_{\text{ToT}}$ (Zeng et al., 2024)	ViT-B/32	CBART	–	17.7	24.3	84.8	38.5	43.6	50.0	45.1
CapDec (Nukrai et al., 2022)	ResNet50	GPT2-large	69.2	26.4	25.1	91.8	60.1	50.2	28.7	45.9
CapDec+RLCF-S (Zhao et al., 2023)	ViT-B/16	OPT-125M	–	–	–	–	68.3	58.5	35.3	–
ViECap (Fei et al., 2023)	ViT-B/32	GPT2-base	–	27.2	24.8	92.9	61.1	64.3	65.0	66.2
EntroCap (Yan et al., 2024)	ViT-B/32	GPT2-base	–	27.6	25.3	94.3	62.5	64.5	67.5	67.0
ViECap+ToCa (Zhou et al., 2024)	ViT-B/32	GPT2-base	–	27.1	25.4	95.0	64.6	69.1	70.5	70.9
MeaCap$_{\text{InvLM}}$ (Zeng et al., 2024)	ViT-B/32	GPT2-base	–	27.2	25.3	95.4	–	–	–	–
ICSD (Ma et al., 2024)	ViT-B/32	BERT-base	–	29.9	25.4	96.6	42.9	44.3	35.6	42.7
CLOSE (Gu et al., 2022)	ViT-L/14	T5-base	–	29.5	25.7	97.8	–	–	–	–
ArcSin (Liu et al., 2024a)	ViT-L/14	T5-base	–	30.3	–	99.6	–	–	–	–
SynTIC (Liu et al., 2024b)	ViT-B/32	Transformer	–	29.9	25.8	101.1	–	–	–	–
TipCap (Wang et al., 2024)	ViT-L/14	GPT2-large	73.3	31.4	54.2	106.6	80.2	62.3	39.6	60.3
TTLLCap (no training, retrieval)	ViT-L/14	OPT-IML-1.3B	55.6	15.2	20.6	61.8	52.4	51.1	60.1	53.1
TTLLCap (no training, retrieval + Socratic)	ViT-L/14	OPT-IML-1.3B	55.5	17.0	19.5	64.5	53.7	51.6	57.9	53.2
TTLLCap (training with retrieval)	ViT-L/14	GPT2-base	63.7	20.5	21.7	77.3	63.9	63.8	76.4	66.4
TTLLCap (embedding only)	ViT-L/14	GPT2-base	66.2	24.7	21.3	73.4	41.5	25.6	10.3	24.8
TTLLCap (embedding + retrieval)	ViT-L/14	GPT2-base	63.6	20.4	21.7	77.2	65.6	64.6	76.4	67.1
TTLLCap (embedding + retrieval + Socratic)	ViT-L/14	GPT2-base	63.5	20.3	21.7	76.5	65.4	64.2	76.8	67.0
TTLLCap (K=6, re-ranking with λ=-0.5)	ViT-L/14	GPT2-base	65.5	21.9	23.2	81.3	69.8	69.9	84.1	72.8
TTLLCap (K=8, re-ranking with λ=-0.5)	ViT-L/14	GPT2-base	65.0	21.6	23.2	80.3	69.8	70.1	82.9	72.6
TTLLCap (K=6, re-ranking with λ=-0.5)	ViT-L/14	GPT2-medium	66.7	22.6	23.1	83.1	70.6	71.2	83.4	73.6
TTLLCap (K=8, re-ranking with λ=-0.5)	ViT-L/14	GPT2-medium	66.7	23.3	23.8	85.7	73.1	73.1	87.2	76.0

Table 1: Results for different captioning methods on MSCOCO and NoCaps.

the Socratic prompt, often fails to retrieve correct places and/or objects. Hence, the method likely fails to improve performance due to this noise.

Appendix C complements the results in Table 1 with some qualitative examples for the captions generated with TTLLCap.

Retrieved Captions	B@1	B@4	M	C
K=4	63.6	20.4	21.7	77.2
K=6 (training with K=4)	64.1	20.6	22.9	79.1
K=6 (training)	63.6	20.2	21.8	77.0
K=8 (training with K=4)	63.1	20.5	23.3	78.4

Table 2: Results on MSCOCO when varying the number of retrieved captions from 4 to 8, either with or without training the model to specifically use more captions.

4.3 Impact of Retrieval Augmentation

Besides assessing captioning quality, we also looked at the impact of different choices regarding the configuration of the retrieval-augmentation component. This was made with the TTLLCap model variant that only uses retrieved captions (i.e., without the Socratic prompt) plus the CLIP embeddings, leveraging the smaller and more efficient GPT2-base decoder.

A first aspect that we analyzed concerns the optimal number of retrieved captions to consider. Previous work with SmallCap and LMCap has pointed to $K = 4$ as the best configuration, and we attempted to further validate this value through an experiment in which we compared TTLLCap with K=4, K=6, or K=8 retrieved captions. Results over the MSCOCO dataset are shown in Table 2, indicating that $K = 4$ is the optimal training configuration. However, results slightly increase when retrieving more captions during inference, even if the model was trained with $K = 4$. We speculate that access to more captions collected via text-text retrieval fails to better guide the model during training, while more captions retrieved through image-text retrieval is indeed helpful during inference.

We also looked at the relation between properties of the retrieved captions and result quality, namely by analyzing (a) the average similarity of the retrieved captions towards the input image, as measured from the CLIP representations, and (b) the diversity and complementarity of the $K = 4$ retrieved captions, also as measured by the average similarity between their CLIP representations. Figure 2 presents these results for images in the MSCOCO test split, plotting CIDEr values against the similarity or diversity measurements (without any normalization of the scores). The results show that CIDEr generally increases with a higher similarity between the retrieved captions and the input image, while it decreases with more diverse sets of retrieved captions (i.e., results are better when the retrieved captions are more similar to each other).

Keeping $K = 4$ retrieved captions, we experimented with the use of a re-ranking strategy to adapt the retrieval results, assessing alternatives

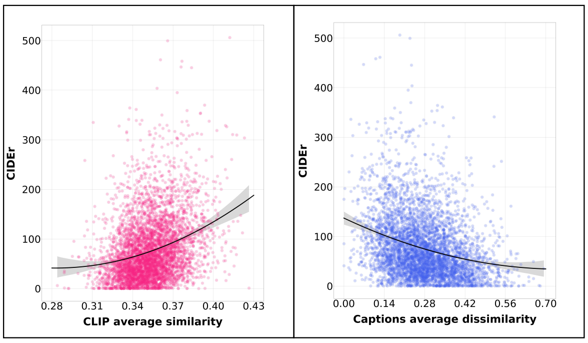


Figure 2: CIDEr values for instances in the MSCOCO test split, versus average similarity between input images and generated captions (left), or diversity in the retrieved captions (right), measured as one minus the average pairwise similarity between the captions.

that either consider a more diverse set of retrieved captions, or instead a more cohesive set. The default retrieval approach only considers similarity towards the input, not attempting to optimize the similarity between the retrieved exemplars themselves. Nothing that an increased diversity on the retrieved captions, or in turn a increased cohesiveness as suggested by Figure 2, can perhaps contribute to improved results, we experimented with a strategy based on the Maximum Marginal Relevance (MMR) approach (Carbonell and Goldstein, 1998), which selects exemplars that are relevant while at the same time controlling for diversity/cohesiveness. If for a given input i we have already selected a set of exemplars $T = \{c_i\}$, following this strategy we will pick up the next exemplar c_j according to:

$$\arg\max_{c_j}(\mathrm{sim}(i, c_j) - \lambda \max_{c_i \in T} \mathrm{sim}(c_j, c_i)), \quad (1)$$

where $\mathrm{sim}()$ denotes CLIP similarity (without any additional normalization), and λ is a parameter that controls the balance between relevance and diversity (which, when negative, promotes cohesiveness). We rely on MMR to iteratively re-rank exemplars from the datastore, scoring the top 50 instances obtained through an initial retrieval.

Table 3 presents the obtained results on the MSCOCO test split, comparing models trained with different λ values in order to promote diversity or cohesiveness. Overall, quality improves when promoting cohesiveness, and the best scores are achieved when $\lambda \approx -0.50$ (i.e., the same value that is used on the main results reported in Table 1). Note that using a negative λ during training and inference, besides promoting cohesiveness, also makes retrieval results depend more on text-text similarity. Making the training and inference stages more similar can contribute to reducing the CLIP modality gap, generally improving performance.

MMR Setting	B@1	B@4	M	C
MMR with $\lambda = 0.15$	63.4	19.8	21.7	75.3
MMR with $\lambda = 0.00$	63.6	20.4	21.7	77.2
MMR with $\lambda = -0.15$	64.8	21.0	22.1	78.4
MMR with $\lambda = -0.30$	65.1	21.5	22.2	79.5
MMR with $\lambda = -0.60$	65.7	21.9	22.4	80.7
MMR with $\lambda = -1.20$	65.0	21.2	22.1	79.0
MMR with $\lambda = -0.50$	65.8	22.2	22.6	81.3

Table 3: Results on MSCOCO when increasing the diversity or the cohesiveness of the retrieved captions.

Appendices A and B further extend the analysis reported in this section, specifically looking at position biases associated to the ordering of the retrieved captions, and looking and how the CLIP modality impacts retrieval quality and, consequently, the captioning results.

5 Conclusions and Future Work

This paper presented TTLLCap, i.e. an improved text-only training method for image captioning based on prompting a pre-trained language model decoder with information derived from CLIP representations of the inputs. Experimental results show that TTLLCap is able to outperform several previous training-free and text-only training methods, especially in terms of out-domain generalization. Out of all the components involved in the proposed approach, the use of retrieved captions is the one that has the highest impact on result quality.

Despite the interesting results, there are also many opportunities for future work. TTLLCap is still outperformed by other similar approaches, e.g. that use synthetic images generated from the textual captions available for training (Ma et al., 2024; Liu et al., 2024b), as well as by fully supervised approaches. We believe that text-only training can be further improved up to the almost same quality as fully supervised techniques, and that the use of other approaches to address the CLIP modality gap will have a fundamental role in this regard (Li et al., 2023c; Wang et al., 2023; Liu et al., 2024a).

In addition, note that our experiments only used English corpora, and it would be important to extend the study to other languages (Ramos et al., 2023b, 2024), in particular considering low-resource languages for which the collection of images paired to textual captions can be harder (i.e., an important motivation for the development of zero-shot or text-only captioning methods is indeed the multilingual captioning scenario).

Limitations and Ethical Considerations

While our work does not raise new ethical issues within the domain of automatic image captioning (e.g., we conducted our experiments on public datasets, carefully designed for academic research and extensively used in previous studies), there are some general important concerns.

For instance image captioning models are notorious for their internal biases, inherited from the training data itself or from the use of pre-trained models such as CLIP. We therefore recommend caution in the use of the approach proposed in this paper, and anticipate further research into model biases, before relying on our work beyond research environments. Still, we observe that balancing a text-only dataset can be easier than collecting balanced text-image pairs, and thus the proposed approach can perhaps offer advantages in terms of mitigating known biases (e.g., if we consider the problem of a dataset containing significantly more images of women in a kitchen than men, collecting more images requires substantial effort, while replacing *woman* with *man*, and their synonyms, in all the training captions is quite simple).

Another important limitation in the work reported on this paper concerns the fact that evaluation is only made on English datasets. Moreover, although the proposed approach can generate image captions without relying on any labeled image-caption training pairs, we still need the independent set of textual captions for model training, which may be difficult to collect in some scenarios (e.g., for some low-resource languages). This might be alleviated in future work, by assessing the use of textual corpora from different sources and/or produced automatically with language models.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. No-Caps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *Proceedings of the International Conference on Machine Learning*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for RAG systems. *arXiv preprint arXiv:2401.14887*.

Michael Denkowski and Alon Lavie. 2014. METEOR Universal: Language specific translation evaluation for any target language. In *Proceedings of the Workshop on Statistical Machine Translation*.

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

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. 2023. Language-only efficient training of zero-shot composed image retrieval. *arXiv preprint arXiv:2312.01998*.

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

Jennifer Hsia, Afreen Shaikh, Zhiruo Wang, and Graham Neubig. 2024. Ragged: Towards informed design of retrieval augmented generation systems. *arXiv preprint arXiv:2403.09040*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3).

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.

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

Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2023a. EVCap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. *arXiv preprint arXiv:2311.15879*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*.

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

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*.

Yang Liu, Xiaomin Yu, Gongyu Zhang, Christos Bergeles, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2024a. ArcSin: Adaptive ranged cosine similarity injected noise for language-driven visual tasks. *arXiv preprint arXiv:2402.17298*.

Zhiyue Liu, Jinyuan Liu, and Fanrong Ma. 2024b. Improving cross-modal alignment with synthetic pairs for text-only image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

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

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Yingzhe Peng, Xu Yang, Haoxuan Ma, Shuo Xu, Chi Zhang, Yucheng Han, and Hanwang Zhang. 2023. Icd-lm: Configuring vision-language in-context demonstrations by language modeling. *arXiv preprint arXiv:2312.10104*.

Longtian Qiu, Shan Ning, and Xuming He. 2024. Mining fine-grained image-text alignment for zero-shot captioning via text-only training. *arXiv preprint arXiv:2401.02347*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

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

Rita Ramos, Emanuele Bugliarello, Bruno Martins, and Desmond Elliott. 2024. PAELLA: Parameter-efficient lightweight language-agnostic captioning model. In *Findings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023a. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023b. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. *arXiv preprint arXiv:2305.19821*.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023c. SmallCap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-augmented transformer for image captioning. In *Proceedings of the International Conference on Content-Based Multimedia Indexing*.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. 2024. Towards retrieval-augmented architectures for image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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

Derek Tam, Colin Raffel, and Mohit Bansal. 2023. Simple weakly-supervised image captioning via CLIP's multimodal embeddings. In *Proceedings of the AAAI Workshop on Creative AI Across Modalities*.

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

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

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

Yiyu Wang, Hao Luo, Jungang Xu, Yingfei Sun, and Fan Wang. 2024. Text data-centric image captioning with interactive prompts. *arXiv preprint arXiv:2403.19193*.

Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. 2019. Tencent ML-images: A large-scale multi-label image database for visual representation learning. *IEEE Access*, 7.

Jie Yan, Yuxiang Xie, Shiwei Zou, Yingmei Wei, and Xidao Luan. 2024. EntroCap: Zero-shot image captioning with entropy-based retrieval. *Social Science Research Network preprint SSRN:4737282*.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2024. Exploring diverse in-context configurations for image captioning. In *Proceedings of the Anual Meeting on Neural Information Processing Systems*.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

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

Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. ConZic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2023. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2305.18010*.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. 2016. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*.

Qing Zhou, Junlin Huang, Qiang Li, Junyu Gao, and Qi Wang. 2024. Text-only synthesis for image captioning. *arXiv preprint arXiv:2405.18258*.

A Analysis of Position Biases

Considering the TTLLCap model variant that only uses retrieved captions (i.e., without the Socratic prompt), and leveraging the smaller and more efficient GPT2-base decoder, we analysed the potential existence of position biases in association to the order of the retrieved captions.

We first note that the GPT2 decoder can have a position bias towards preferring information from retrieved captions at the beginning or the ending of the prompt, independently of the similarity of the captions towards the input. We therefore experimented with changing the ordering of the $K = 4$ captions, either keeping the decoder model trained by default with retrieved captions in decreasing order of similarity, or training the decoder in different settings. Table 4 presents the obtained results on the MSCOCO test split, showing that the order of the captions in the prompt has little effect on the final result. Still, using the retrieved captions in a random order during training achieves slightly better results, i.e. a fact that is perhaps also tied to the CLIP modality gap and to approximating the training and inference stages.

We also looked at how the self-attention layers of the decoder weight information from the

$K = 4$ retrieved captions, considering settings that involve retrieved captions sorted in descending order, or retrieved captions in random order. Figure 3 plots these results, showing the distribution of self-attention weights across all MSCOCO testing instances, averaged (a) separately over the lowest/highest 6 layers of the GPT2-base decoder, and (b) over the tokens that constitute each of the $K = 4$ retrieved captions.

The results show that, when captions are in descending order and the model was trained in this setting (i.e., the top pair of plots in Figure 3), more attention is given to the first captions. The model is giving priority to the captions that have more similarity towards the input, and that are therefore more likely to accurately describe the image. Instead, when captions are ordered randomly, every caption is approximately given an equal amount of attention. This happens independently of whether the model was trained with captions in descending order (i.e., the pair of plots in the middle of Figure 3), or with captions in random order.

B Impact of the CLIP Modality Gap

Our analyses that focused on the impact of different configurations for the retrieval component point to the fact that retrieval quality has a significant role in improving the captioning performance. Moreover, making the training and inference stages more similar, in terms of how they use retrieval results, seems to contribute to improved results, which suggests that the CLIP modality gap can be a limiting factor for the results obtained with our method.

In an attempt to further validate these ideas, we experimented with a setting in which model training is still not directly relying on images, but in which retrieval is always made with basis on image-text similarity (instead of using text-text similar-

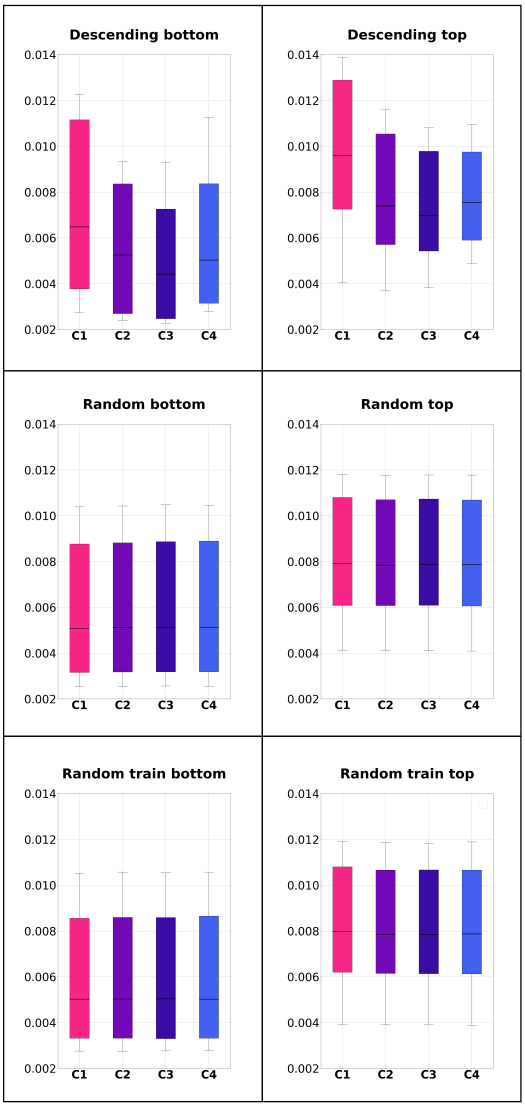


Figure 3: Distribution of average self-attention weights towards tokens associated to the four retrieved captions (i.e., C1 to C4), across all MSCOCO testing instances and separately over the lowest (i.e., the plots on the left) and highest (i.e., the plots on the right) six layers of the GPT2-base decoder. From top to bottom, the plots are derived from models using captions (a) in descending order during training and inference, (b) in descending order during training and random order during inference, and (c) in random order during training and inference.

Caption Ordering.	B@1	B@4	M	C
Descending (default)	63.6	20.4	21.7	77.2
Ascending	63.9	20.8	21.8	78.0
Random	63.8	20.8	21.8	77.6
Descending (default)	63.6	20.4	21.7	77.2
Ascending (training)	63.9	19.9	21.8	75.8
Random (training)	64.4	20.9	22.0	78.3

Table 4: Results when varying the ranking order of the retrieved captions, placing them in the prompt in descending order of similarity (the standard configuration, with the decoder model trained by default in this way), in ascending order of similarity, or in a random order.

Retrieved Captions	B@1	B@4	M	C
T2T	63.6	20.4	21.7	77.2
T2T (with $\lambda = -0.5$)	65.8	22.2	22.6	81.3
I2T	75.1	32.7	26.2	108.5
I2T (with $\lambda = -0.5$)	74.2	31.8	25.8	107.7

Table 5: Results on MSCOCO considering text-text (T2T) or image-text (I2T) similarity, with/without re-ranking the retrieval results to promote cohesiveness.

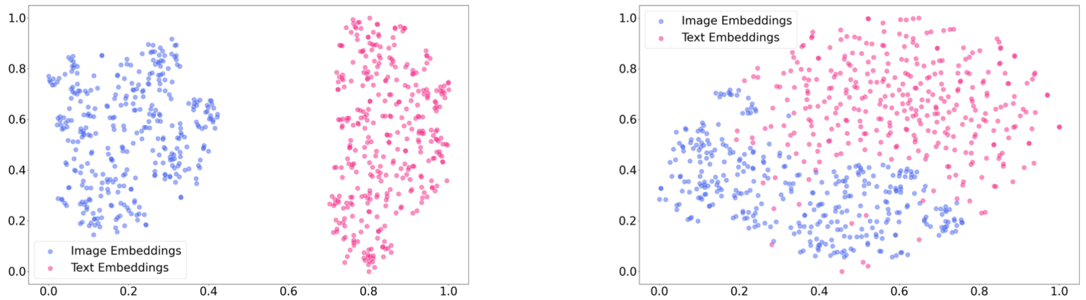


Figure 4: Two-dimensional t-SNE representations for CLIP embeddings of images (i.e., the blue dots) and captions (i.e., the pink dots, corresponding to five captions per image) from the MSCOCO validation split. The plots compare results before (left) and after (right) the addition of Gaussian noise to the caption embeddings.

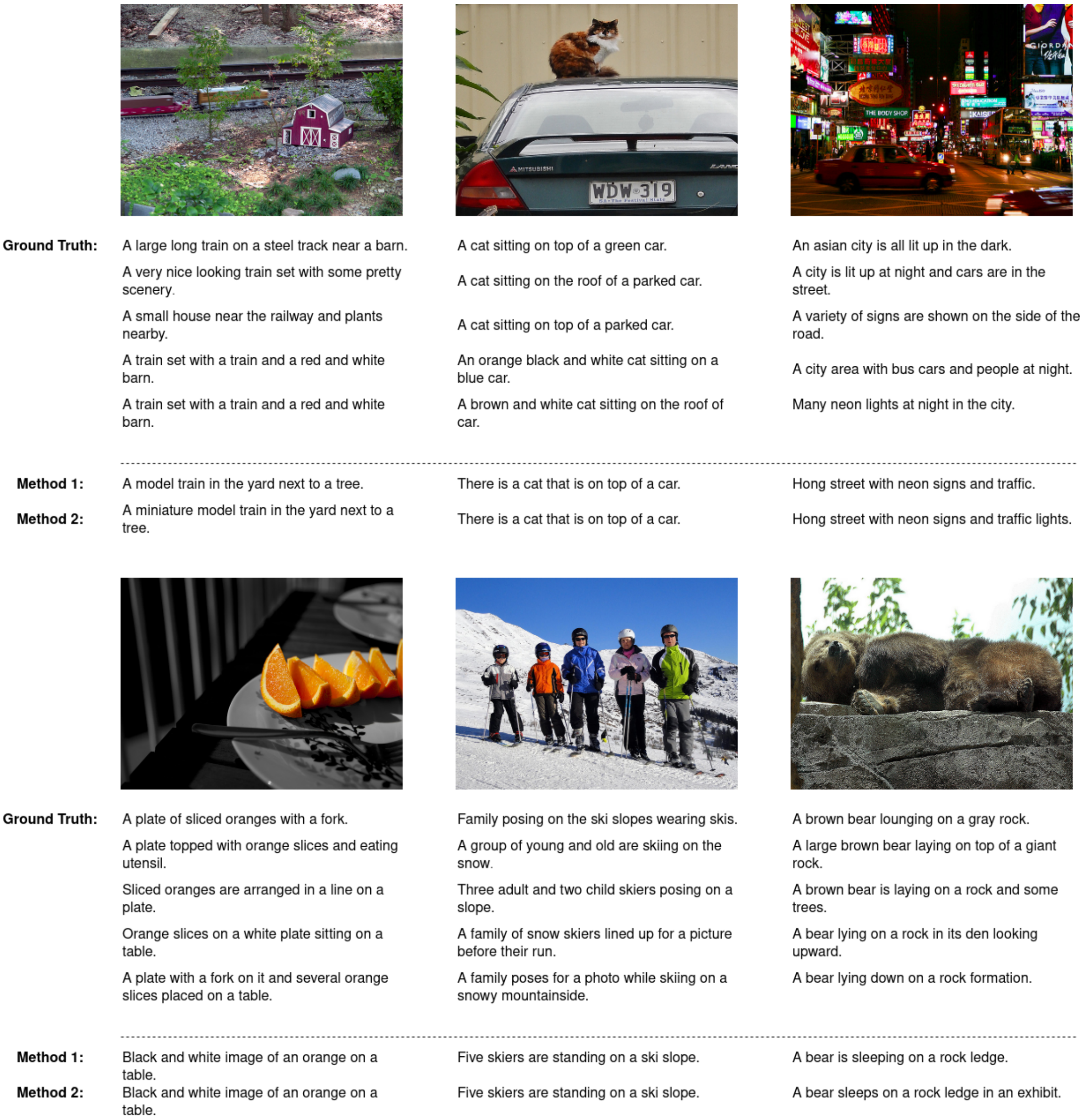
ity during training, and image-text similarity during inference). This setting thus corresponds to a TTLLCap variant that only uses training with the retrieved captions, without cross-attention towards image representations and without the Socratic prompt, relying on the smaller and more efficient GPT2-base decoder. The intuition for exploring this setting relates to the fact that it can perhaps approximate an ideal scenario, in which the CLIP modality gap problem is fully addressed.

Results over the MSCOCO dataset are shown in Table 5, using $K = 4$ retrieved captions and also considering settings that involve re-ranking to promote cohesiveness. The values confirm that image-text similarity indeed leads to a much higher performance, with cohesiveness failing to improve results on this setting. The CLIP modality gap indeed seems like an important limiting factor to our approach, and future work should attempt to further address this aspect, e.g. through the exploration and extension of approaches proposed in previous work (Wang et al., 2024), which nonetheless have the limitation that they require the tuning of parameters with a small set of image-caption pairs.

Figure 4 illustrates the CLIP modality gap, using t-SNE projections (Van der Maaten and Hinton, 2008) to represent in two dimensions the embeddings for the images and captions corresponding to instances in the MSCOCO validation split. The plots show that while the addition of Gaussian noise indeed contributes to reducing the modality gap, there is still significant room for improvement.

C Qualitative Examples

Figure 5 presents several examples of captions generated with the proposed approach, considering two different model configurations. Both of these use cross-attention towards the CLIP embeddings, retrieved captions, and GPT2-base as the decoder, the difference being that Method 2 uses a Socratic prompt, while Method 1 does not. All images were taken from the MSCOCO test split.



Ground Truth:	A large long train on a steel track near a barn.	A cat sitting on top of a green car.	An asian city is all lit up in the dark.
	A very nice looking train set with some pretty scenery.	A cat sitting on the roof of a parked car.	A city is lit up at night and cars are in the street.
	A small house near the railway and plants nearby.	A cat sitting on top of a parked car.	A variety of signs are shown on the side of the road.
	A train set with a train and a red and white barn.	An orange black and white cat sitting on a blue car.	A city area with bus cars and people at night.
	A train set with a train and a red and white barn.	A brown and white cat sitting on the roof of car.	Many neon lights at night in the city.
Method 1:	A model train in the yard next to a tree.	There is a cat that is on top of a car.	Hong street with neon signs and traffic.
Method 2:	A miniature model train in the yard next to a tree.	There is a cat that is on top of a car.	Hong street with neon signs and traffic lights.

Ground Truth:	A plate of sliced oranges with a fork.	Family posing on the ski slopes wearing skis.	A brown bear lounging on a gray rock.
	A plate topped with orange slices and eating utensil.	A group of young and old are skiing on the snow.	A large brown bear laying on top of a giant rock.
	Sliced oranges are arranged in a line on a plate.	Three adult and two child skiers posing on a slope.	A brown bear is laying on a rock and some trees.
	Orange slices on a white plate sitting on a table.	A family of snow skiers lined up for a picture before their run.	A bear lying on a rock in its den looking upward.
	A plate with a fork on it and several orange slices placed on a table.	A family poses for a photo while skiing on a snowy mountainside.	A bear lying down on a rock formation.
Method 1:	Black and white image of an orange on a table.	Five skiers are standing on a ski slope.	A bear is sleeping on a rock ledge.
Method 2:	Black and white image of an orange on a table.	Five skiers are standing on a ski slope.	A bear sleeps on a rock ledge in an exhibit.

Figure 5: Examples of generated captions for images taken from the MSCOCO dataset.