Controllable Image Captioning via Prompting

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, Linlin Li

Huawei Inc.

wn6149@mail.ustc.edu.cn, jh_xie@tongji.edu.cn, {wujihao, jiamingbo, lynn.lilinlin}@huawei.com

Abstract

Despite the remarkable progress of image captioning, existing captioners typically lack the controllable capability to generate desired image captions, e.g., describing the image in a rough or detailed manner, in a factual or emotional view, etc. In this paper, we show that a unified model is qualified to perform well in diverse domains and freely switch among multiple styles. Such a controllable capability is achieved by embedding the prompt learning into the image captioning framework. To be specific, we design a set of prompts to fine-tune the pre-trained image captioner. These prompts allow the model to absorb stylized data from different domains for joint training, without performance degradation in each domain. Furthermore, we optimize the prompts with learnable vectors in the continuous word embedding space, avoiding the heuristic prompt engineering and meanwhile exhibiting superior performance. In the inference stage, our model is able to generate desired stylized captions by choosing the corresponding prompts. Extensive experiments verify the controllable capability of the proposed method. Notably, we achieve outstanding performance on two diverse image captioning benchmarks including COCO Karpathy split and TextCaps using a unified model.

1 Introduction

Image captioning is one of the fundamental tasks in computer vision, which aims to automatically generate natural and readable sentences to describe the image contents. The last decade has witnessed the rapid progress of image captioning, thanks to the development of sophisticated visual representation learning (Zhang et al. 2021; Fang et al. 2021), cross-modal fusion (Pan et al. 2020; Huang et al. 2019; Li et al. 2020), vision-language pre-training (Hu et al. 2021; Li et al. 2022; Wang et al. 2021b), etc. Image captioning is a challenging task that requires the captioners to recognize the objects and attributes, understand their relationships, and properly organize them in the sentence.

Despite the remarkable advances, current image captioning algorithms generally lack the controllable capability to generate desired captions. In other words, once the captioning model is trained, the caption generation process can hardly be influenced. Typical cases include the control of caption length and description style. (1) *Length controllable capability*. Sometimes, a brief description is required to get an overview of the image, while in other circumstances, a detailed caption is preferred to acquire more information. This can be roughly reflected by the controllable capability of the caption length, which is a basic demand in practical applications, but has been largely overlooked in existing methods. (2) *Style controllable capability*. An image can be described in quite different views. For example, given an image with textual contents (e.g., a poster or sign), some people care about the objects, but some may pay more attention to the textual words. Besides, people may generate non-factual captions, e.g., emotional descriptions that contain positive or negative expressions. It is of vital importance to insert different styles in the captioning model to enhance its expressibility. How to simultaneously maintain multiple styles and freely switch among them is an open problem. Existing captioning approaches typically separately handle each scenario, e.g., train a captioner on the COCO dataset (Lin et al. 2014) and train another model on the TextCaps dataset (Sidorov et al. 2020). As a result, these captioners are domain-specific, without style controllability.

In this paper, we show that a unified model is able to generate captions with different lengths and styles. As shown in Figure 1, our approach describes an image semantically

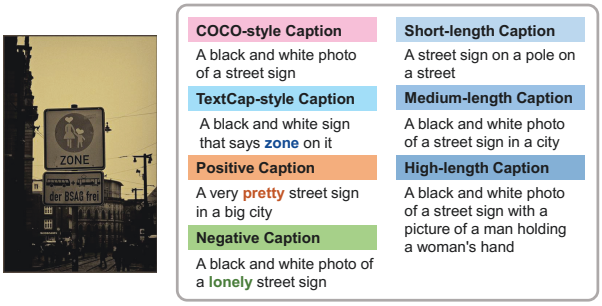

Figure 1: Leveraging a unified model, the proposed method is able to generate diverse captions such as COCO-style [ ], TexCap-style [ ], Positive [ ], Negative [ ], and different caption lengths including Short-length [ ], Medium-length [ ], and High-length [ ]. Best view in color.

accurately in diverse views. This captioning controllable capability is achieved by designing prompts within the cross-modal language model. After large-scale pre-training, the image captioner has already gained the ability to generate diverse captions, but is largely overwhelmed in the downstream fine-tuning, e.g., on a certain stylized dataset such as COCO (Lin et al. 2014). In this work, we aim to unveil the potential hidden in the pre-trained model to flexibly switch captioning styles. Our approach is motivated by the recent advance in prompt learning techniques (Liu et al. 2021) in natural language processing (NLP). In the proposed framework, prompts serve as the anchor points to gather data from different domains, facilitating the multi-domain joint training. By virtue of prompt engineering, captions with different lengths, different styles, and different emotions can be properly separated within a unified model. The prompts, together with the image-text pair, jointly serve as the training corpus to optimize the captioning model. Furthermore, instead of manually designing prompts, we encourage the captioner to automatically learn the prompt embeddings in an end-to-end manner. This continuous auto-prompt learning searches the suitable prompt representations in the entire word embedding space, which not only avoids the heuristic prompt design but also exhibits superior performance.

In the inference stage, different prompts serve as the prediction hints to guide the caption generation. By automatically learning multiple prompt embeddings, the proposed approach has the following merits. Our approach (i) is free of manual prompt engineering, which requires domain expertise and careful word tuning; (ii) is able to generate diverse stylized captions via a single model, which is infeasible for most existing state-of-the-art captioners such as BLIP (Li et al. 2022), LEMON (Hu et al. 2021), and SimVLM (Wang et al. 2021b); (iii) does not degrade the performance on different domains such as COCO (Lin et al. 2014) and TextCaps (Sidorov et al. 2020), and outperforms the traditional training strategy using a prefixed prompt; (iv) is simple and general, which is ready to perform on more domains by incorporating other stylized data.

In summary, the contributions of this work are three-fold:

- To our knowledge, we are the first to propose the prompt-based image captioning framework, which provides a simple yet effective manner to control the caption style.
- We validate the manually designed prompts. We further introduce auto-prompt learning to avoid the heuristic prompt design and achieve superior results.
- Qualitative and quantitative results verify the controllable capability of the proposed framework. Leveraging a unified model, we achieve outstanding performance on several benchmarks including COCO Karpathy set (Lin et al. 2014), NoCaps (Agrawal et al. 2019), and TextCaps (Sidorov et al. 2020).

2 Related Work

General Image Captioning. Image captioning aims to generate a textual description of the image contents (Vinyals et al. 2015), which typically contain a visual encoder to extract the image features and a multi-modal fusion model such as LSTM and Transformer for text generation. To represent the visual contents, previous methods (Huang et al. 2019; Anderson et al. 2018; Deng et al. 2020; Cornia et al. 2020; Fei 2022; Ji et al. 2021) utilize the Region-of-Interest (RoI) features from object detectors (Ren et al. 2016). Recent captioning algorithms (Fang et al. 2021; Xu et al. 2021; Wang et al. 2021b) shed light on the grid features for high efficiency and potentially better performance due to end-to-end training. As for the cross-modal model, classic captioners (Anderson et al. 2018; Huang et al. 2019; Pan et al. 2020; Song et al. 2021) typically utilize the LSTM, while the recent approaches (Li et al. 2020; Zhang et al. 2021; Li et al. 2022; Wang et al. 2021b; Wang, Xu, and Sun 2022; Luo et al. 2021) leverage the attention-based models to fuse vision-language representations and predict the captions.

Controllable Image Captioning. Despite the impressive progress, fewer efforts have been made to control the caption generation. Cornia *et al.* (Cornia, Baraldi, and Cucchiara 2019) utilize image regions to generate region-specific captions. Chen *et al.* (Chen et al. 2020a) propose the abstract scene graph to represent user intention and control the generated image captions. Length-controllable captioning approach is proposed in (Deng et al. 2020), which learns length level embeddings to control the caption length. Shuster *et al.* (Shuster et al. 2019) release an image captioning dataset with personality traits as well as a baseline approach. Zhang *et al.* (Zhang et al. 2022) propose a multi-modal relational graph adversarial inference (MAGIC) framework for diverse text caption. SentiCap (Mathews, Xie, and He 2016) utilizes a switching recurrent neural network with word-level regularization to generate emotional captions. Chen *et al.* (Chen et al. 2018) present a style-factual LSTM to generate captions with diverse styles such as humorous and romantic. However, some of the aforementioned methods (Cornia, Baraldi, and Cucchiara 2019; Chen et al. 2020a, 2018) rely on additional tools or expensive annotations for supervision. In (Kobus, Crego, and Senellart 2016), domain/tag embeddings are involved to control the style, and thus the model architecture is tag-related. Some methods (Mathews, Xie, and He 2016; Chen et al. 2018) can be regarded as the ensemble framework, which include two groups of parameters for factual and stylized branches, increasing the model complexity.

In this work, we control the image captioning style from a different view, i.e., prompt learning. The proposed framework merely involves lightweight learnable prompt embeddings while keeping the baseline architecture unchanged, which is conceptually simple and easy to implement.

Vision-language Pre-training. Vision-language (VL) pre-training is a popular manner to bridge vision and language representations (Dou et al. 2021). CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) use the cross-modal contrastive learning to align the VL representations. Recent VL pre-training approaches (Zhou et al. 2020; Chen et al. 2020b; Huang et al. 2021) generally adopt the attention mechanism (Vaswani et al. 2017) to fuse the VL representations. After large-scale pre-training on the image-text corpus, these models are further fine-tuned on the downstream datasets to conduct a variety of VL tasks such as image captioning. SOHO (Huang et al. 2021) extracts compact image features via a

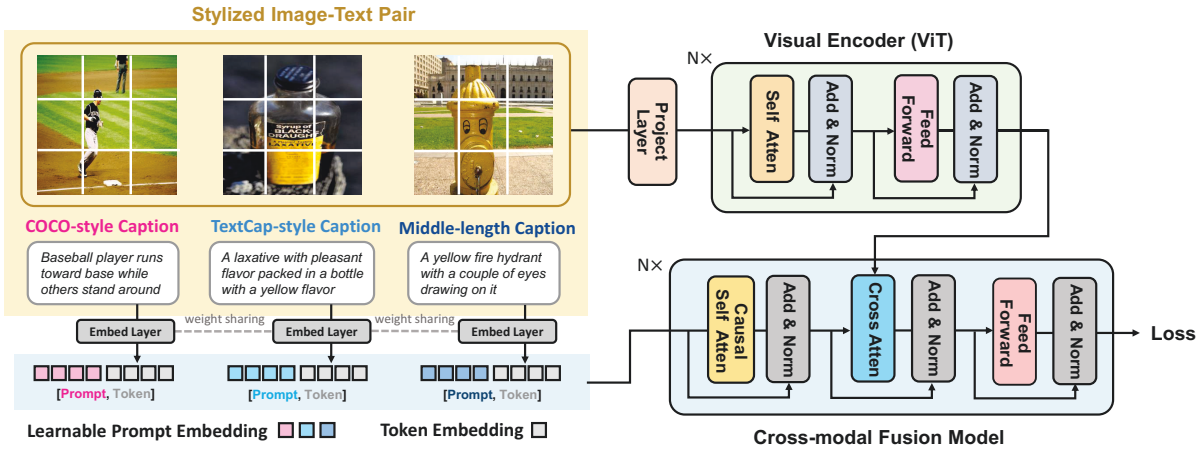

Figure 2: An overview of the proposed prompt-based image captioning framework. Our model optimizes multiple learnable prompt embeddings to absorb stylized data from different domains to jointly train the image captioner. In the inference stage, the model is able to generate diverse captions by feeding different prompts.

learned visual dictionary and trains the whole framework in an end-to-end manner. ALBEF (Li et al. 2021) conducts the cross-modal alignment using contrastive learning technique (Radford et al. 2021) before representation fusion. SimVLM (Wang et al. 2021b) utilizes prefix language modeling for model optimization on the large-scale VL corpus. Inspired by previous arts, we also involve VL pre-training to improve the captioning quality.

Prompt Learning. Prompt learning has gained increasing popularity in natural language processing (NLP) (Liu et al. 2021). Prompt learning allows the language model to be pre-trained on the large-scale corpus, and is able to perform downstream tasks by defining a proper prompting function. Jiang *et al.* (Jiang et al. 2020) propose mining-based and paraphrasing-based approaches to automatically generate high-quality prompts. Shin *et al.* (Shin et al. 2020) search for the proper prompts via a gradient-based approach. Recently, continuous prompt learning has been explored, which directly optimize prompt vectors in the continuous word embedding space (Zhong, Friedman, and Chen 2021; Li and Liang 2021; Lester, Al-Rfou, and Constant 2021; Zhou et al. 2021). It is worth mentioning that prompt learning has been rarely touched in the image captioning. Different from the traditional usage of prompt learning that aims to elicit knowledge for higher performance, we focus on the controllable capability of the captioning algorithm. In the proposed framework, except for the superior performance, the more attractive characteristic is that we can freely switch diverse styles via prompting, which greatly enhances the controllability and expressibility of the image captioner.

3 Approach

In this section, we introduce the method details of the proposed controllable image captioner. First, in Section 3.1, we revisit autoregressive image captioning, which serves as the baseline of our approach. Then, in Section 3.2, we elaborate the manual prompt engineering for image captioning. Finally, we exhibit how to optimize the learnable prompts in Section 3.3 and the inference details in Section 3.4.

3.1 Revisiting Autoregressive Image Captioning

In our method, we adopt the unidirectional language modeling (LM) based image captioning framework as the baseline. Such a framework typically utilizes a transformer block to fuse the image $\boldsymbol{v}$ and text sequence $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$. The token x_t is generated in an autoregressive manner based on the previous tokens $\boldsymbol{x}_{<t}$. The training objective of the cross-modal LM loss is as follows:

$$\mathcal{L}_{\text{LM}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{x})\in\mathcal{D}}\Big[\sum_t \log P\left(x_t|g(\boldsymbol{v}), f(\boldsymbol{x}_{<t})\right)\Big], \quad (1)$$

where $g(\cdot)$ denotes the visual encoder, $f(\cdot)$ represents the word embedding layer, $P(\cdot|\cdot)$ can be regarded as the cross-modal fusion model (e.g., transformer decoder in Figure 2), which receives the visual features $g(\boldsymbol{v})$ and previous token embeddings $f(\boldsymbol{x}_{<t})$ to predict the next word token x_t.

During inference, the autoregressive models take a special token `[BOS]` as input to predict the first token x_1, then x_1 is fed into the model to obtain the next token x_2. This autoregressive prediction process is continued until the special token `[EOS]` is predicted.

3.2 Prompt-based Image Captioning

Model Pre-training. Following previous works (Zhang et al. 2021; Hu et al. 2021; Li et al. 2022; Wang et al. 2021b), we also adopt the large-scale pre-training on the noisy image-text corpus to improve the downstream captioning task. Besides the language modeling (LM) loss, we also adopt the image-text contrastive loss (Radford et al. 2021; Jia et al. 2021) and image-text matching loss (Chen et al. 2020b; Li et al. 2021, 2022) to jointly optimize the visual encoder and cross-modal fusion model, as follows:

$$\mathcal{L}_{\text{Pre-train}} = \mathcal{L}_{\text{Contrast}} + \mathcal{L}_{\text{Match}} + \mathcal{L}_{\text{LM}}. \quad (2)$$

The contrastive loss measures the similarity of the image-text pairs via a light fusion manner such as dot-product, while the matching loss measures the image-text similarity via a heavy fusion manner such as cross-attention. It has been widely recognized that both of them can facilitate cross-modal alignment (Li et al. 2022). Therefore, although we focus on the image captioning, we additionally include the $\mathcal{L}_{\text{Contrast}}$ and $\mathcal{L}_{\text{Match}}$ in the pre-training stage. As for more details, please refer to BLIP (Li et al. 2022).

Prompt Engineering. After pre-training, the model already acquires zero-shot captioning capability thanks to the language modeling loss $\mathcal{L}_{\text{LM}}$. Therefore, previous LM-based image captioners such as SimVLM (Wang et al. 2021b) and BLIP (Li et al. 2022) leverage a pre-defined prompt such as "`a picture of`" or "`a photo of`" to facilitate the image captioning. In this work, we aim to unveil the model potential of generating diverse captions via prompting.

In contrast to single prompt engineering, in the fine-tuning stage, we design multiple prompts as the anchors to distinguish the training data from different domains. In this way, different stylized captions do not disturb their counterparts and together contribute to a stronger model. The manually designed prompts are illustrated in Table 1. (i) For the cross-domain scenario, e.g., evaluating a model on both COCO (Lin et al. 2014) and TextCaps (Sidorov et al. 2020), it is straightforward to assign different prompts for these datasets to learn domain-specific descriptions. (ii) As for the caption length control, we divide the image captions from COCO and TextCaps into three levels depending on the caption length. Captions whose length is in the range $[0, 10)$, $[10, 16)$, and $[16, +\infty)$ are divided. Each of these subsets is assigned with a specific prompt, as shown in Table 1. (iii) Finally, current image captions are typically factual. Nevertheless, each image in the COCO dataset is labeled by five annotators, inevitably containing emotional descriptions. To this end, we collect the positive and negative captions in the COCO dataset to form the non-factual subsets, which contain the pre-defined positive words such as "`great`, `nice`, `cute`" and negative words such as "`ugly`, `terrible`, `disgusting`". Despite these non-factual captions being rare, our method still learns satisfying styles using limited samples, justifying the few-shot learning ability of prompt engineering. The entire positive and negative words, and other potentially effective manual prompts are presented in the *supplementary material*.

Model Fine-tuning. In our framework, multiple training sets are mixed together to train a unified model. Compared to Eq. (1), we predict token x_t based on the visual features $g(\boldsymbol{v})$, prompt token embeddings $f(\boldsymbol{p})$, and previous token embeddings $f(\boldsymbol{x}_{<t})$. Different stylized data is assigned with a specific prompt as illustrated in Table 1. During training, we prepend these hand-crafted prompts to caption tokens as the textual description of the image. We assemble different stylized datasets to jointly train the captioning model using a prompt-based LM loss as follows:

$$-\sum_i \left[\mathbb{E}_{(\boldsymbol{v},\boldsymbol{p}_i,\boldsymbol{x})\in\mathcal{D}_i} \Big[\sum_t \log P\left(x_t | g(\boldsymbol{v}), f(\boldsymbol{p}_i), f(\boldsymbol{x}_{<t})\right) \Big] \right], \tag{3}$$

Caption Style	Manual Prompt $\boldsymbol{p}$
COCO-style	a normal picture that shows
TextCap-style	a textual picture that shows
Short-length	a picture with a short caption that shows
Medium-length	a picture with a medium caption that shows
High-length	a picture with a long caption that shows
Positive	a positive picture that shows
Negative	a negative picture that shows

Table 1: Illustration of the manual prompts.

where $\boldsymbol{p}_i$ denotes the manual prompt for i-th dataset $\mathcal{D}_i$. Note that the prompt tokens $\boldsymbol{p}_i$ and caption tokens $\boldsymbol{x}_{<t}$ share the same embedding mapping layer $f(\cdot)$. In this framework, we keep the baseline model architecture unchanged without additional learnable blocks, which is parameter-efficient.

3.3 Auto-prompt Learning

To avoid the laborious manual prompt engineering in Section 3.2, we further encourage the network to automatically learn the prompts in an end-to-end manner, as shown in Figure 2. Given a sequence of the manual prompt tokens such as "`a textual picture that shows`", the model first maps each token to a unique numeric ID using WordPiece technique. Then, for a BERT-base model, the token IDs are projected to 768-dim word embeddings via the token embedding layer $f(\cdot)$ as the input of the vision-language fusion model, i.e., $f(\boldsymbol{p}) \in \mathbb{R}^{N\times 768}$, where N represents the prompt length. Instead of the manual prompt engineering, we propose to learn the caption prompt embeddings $\mathbf{P}$ as follows:

$$\mathbf{P} = [\mathrm{P}]_1[\mathrm{P}]_2 \cdots [\mathrm{P}]_N, \tag{4}$$

where each embedding vector $[\mathrm{P}]_k$ $(k \in 1, \cdots, N)$ has the same dimension as the word embedding. In other words, $\mathbf{P} \in \mathbb{R}^{N\times 768}$ serves as an alternative of the manual prompt embedding $f(\boldsymbol{p})$. In the training stage, prompt embeddings $\mathbf{P}$ are jointly optimized with the captioning network as follows:

$$-\sum_i \left[\mathbb{E}_{(\boldsymbol{v},\boldsymbol{x})\in\mathcal{D}_i} \Big[\sum_t \log P\left(x_t | g(\boldsymbol{v}), \mathbf{P}_i, f(\boldsymbol{x}_{<t})\right) \Big] \right]. \tag{5}$$

The proposed framework learns specific prompt embeddings $\mathbf{P}_i$ for each domain-specific dataset $\mathcal{D}_i$. During the end-to-end training, the gradients can be effectively back-propagated to optimize the prompt embeddings. To this end, the captioner is able to fully explore the suitable prompt representations in the continuous word embedding space.

3.4 Prompt-based Inference

After prompt learning, our model is able to generate diverse captions using different prompts. In the manual prompt framework, after encoding the special token `[BOS]`, we sequentially embed the prompt tokens via $f(\boldsymbol{p})$ and feed them to the language model to generate the caption in an autoregressive manner. In the auto-prompt framework, we directly concatenate the token embedding of `[BOS]` and learned prompt embeddings $\mathbf{P}$ as the input of the language model.

By switching different prompts, the proposed captioner is able to generate a certain stylized caption.

4 Experiment

4.1 Datasets and Metrics

Pre-training Data. In the experiments, following our baseline approach (Li et al. 2022), we collect the image-text pairs from Visual Genome (Krishna et al. 2017), COCO (Lin et al. 2014), SBU Captions (Ordonez, Kulkarni, and Berg 2011), Conceptual Captions (Sharma et al. 2018), Conceptual 12M (Changpinyo et al. 2021), and a filtered version of LAION (115M images) (Schuhmann et al. 2021) to form the pre-training data. Following BLIP (Li et al. 2022), these data are filtered by a large model to form the high-quality bootstrapped dataset. In total, the pre-training corpus consists of about 129 million images.

Evaluation Datasets and Metrics. We evaluate the proposed method on the COCO caption dataset (Lin et al. 2014) of Karpathy split (Karpathy and Fei-Fei 2015), NoCaps (Agrawal et al. 2019), and TextCaps (Sidorov et al. 2020). To evaluate the quality of the generated captions, we use standard metrics in the image captioning task, including BLEU@4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016). In the inference stage, beam search (beam size = 3) is adopted in all experiments. More details and visualization results can be found in the supplementary material.

4.2 Implementation Details

Our model is implemented in Python with PyTorch. In the pre-training stage, the model is trained on 32 V100 GPUs. The image encoder is initialized from ViT-B/16 pre-trained on the ImageNet (Dosovitskiy et al. 2020), and the text encoder is initialized from BERT-base (Devlin et al. 2018). We pre-train the whole model for 32 epochs using a batch size of 2880. We use AdamW optimizer (Loshchilov and Hutter 2017) with a weight decay of 0.05. The learning rate is warmed-up to 3×10^{-4} and decayed linearly with a rate of 0.85. We take random image crops of resolution 224×224 during pre-training.

In the fine-tuning stage, we train the model using a small learning rate of 1×10^{-5} and linearly decay it. The model is fine-tuned for 5 epochs. Following previous works (Wang et al. 2021b), the image resolution is increased to 384×384 during fine-tuning. As for the prompt embedding $\mathbf{P} \in \mathbb{R}^{N \times 768}$, we randomly initialize it and set $N = 16$. We optimize our algorithm using standard cross-entropy loss *without* reinforcement learning. The proposed Controllable Captioner is denoted as ConCap.

4.3 Ablation Study

Manual Prompt *v.s.* w/o Prompt. Previous works such as BLIP utilize a pre-defined prompt "`a picture of`" to facilitate the caption generation. However, as shown in Table 2, in the zero-shot evaluation without model fine-tuning (① and ②), an empty prompt is even more effective. After downstream model fine-tuning (④ and ⑤), we observe that this hand-crafted prompt is beneficial to COCO dataset but harmful to TextCaps. These results show that the heuristic prompt is not always a good choice, which potentially requires the laborious manual design for different datasets.

Configuration	COCO Test B@4	COCO Test C	TextCaps Val B@4	TextCaps Val C
w/o Fine-tuning (Frozen Model)				
① w/o Prompt	33.9	106.6	18.6	48.6
② Manual Prompt	23.7	83.8	14.5	38.4
③ Learned Prompt	**38.3**	**125.1**	**20.7**	**56.7**
Multi-dataset Individual Training				
④ w/o Prompt	39.1	132.4	30.4	113.4
⑤ Manual Prompt	39.4	132.6	30.1	111.2
⑥ Learned Prompt	**40.5**	**133.5**	**31.2**	**115.9**
Multi-dataset Joint Training				
⑦ w/o Prompt	39.2	131.9	30.1	111.6
⑧ Shared Manual Prompt	39.3	132.2	30.0	110.4
⑨ Multi-prompt	39.6	132.8	30.7	113.5
⑩ Multi-prompt	**40.5**	**133.7**	**31.3**	**116.7**

Table 2: Ablation comparisons on the COCO Karpathy test split (Lin et al. 2014) and TextCaps validation set (Sidorov et al. 2020), where B@4 and C denote BLEU@4 and CIDEr scores, respectively.

Effectiveness of Learned Prompt. For a *frozen* image captioner, we only optimize the prompt embeddings in setting ③ in Table 2. The results show that learned prompt embeddings greatly unveil the potential of a pre-trained model with a good zero-shot performance of 125.1 CIDEr on COCO. After joint training of prompt embeddings and captioning model, the performance of the learned prompt is still superior to the manual prompt (⑥ *v.s.* ⑤).

Multi-dataset Individual Training *v.s.* Joint Training. Previous works typically train the model individually on different datasets. In setting ⑤ in Table 2, we separately fine-tune the image captioner on COCO and TextCaps. In setting ⑧ in Table 2, we merge the datasets of COCO and TextCaps, and use a *shared* prompt "`a picture of`" for both datasets. By analyzing the results of ⑤ and ⑧, we can observe that simply combining two diverse datasets with different styles will degrade the performance. This is consistent with common sense that data from diverse domains will challenge the model training.

Single Prompt *v.s.* Multi-prompt. In setting ⑨ of Table 2, we still combine the COCO and TextCaps to jointly train a unified captioner, but separate multi-domain data using different prompts. It is interesting that "multi-prompt for joint training" (⑨) not only outperforms the "single prompt for joint training" (⑧), but also surpasses "single prompt for individual training" (⑤), indicating that multiple (even manually designed) prompts can effectively separate the data from different domains. Furthermore, the most promising characteristic of "multi-prompt" is that we can control the caption style by feeding different prompts, which is infeasible for the "single prompt" setting. Finally, we encourage the model to jointly optimize multiple learnable prompt embeddings in an end-to-end manner (⑩), which achieves the best results.

Auto-prompt Length. We further validate the influence of

	$N=4$		$N=8$		$N=16$		$N=24$	
	B@4	C	B@4	C	B@4	C	B@4	C
TextCaps Val	30.8	115.4	31.1	115.4	31.3	**116.7**	**31.5**	115.9

Table 3: Ablation of the prompt embedding length N on the TextCaps validation set (Sidorov et al. 2020).

Prompt Style	B@4	M	C	S
Short-length Prompt	39.9	30.2	132.3	23.0
Medium-length Prompt	35.1	30.9	122.9	23.9
High-length Prompt	26.9	30.7	71.6	**25.0**
Positive Prompt	27.0	25.8	97.6	20.7
Negative Prompt	37.0	29.3	121.5	22.9
TextCap-style Prompt	22.1	25.9	66.0	20.5
COCO-style Prompt	**40.5**	**30.9**	**133.7**	23.8

Table 4: Performance comparisons of different prompts on the COCO Karpathy test split (Lin et al. 2014), where B@4, M, C, S denote BLEU@4, METEOR, CIDEr, and SPICE.

prompt embedding length N. In Table 3, the prompt embeddings of different lengths are randomly initialized. We test different lengths of $N = 4, 8, 16, 24$ and observe that increasing the prompt embedding length N can consistently improve the performance. In our experiments, We choose $N = 16$ as it already yields saturated results.

Evaluation of Different Prompts on COCO. Finally, we evaluate the performance of different automatically learned prompts on the COCO Karpathy test split. The results are shown in Table 4. There is no doubt that "COCO-style Prompt" overall performs best, which leverages the entire training set of COCO for model fine-tuning. "TextCap-style Prompt" exhibits poor results on the COCO dataset, which justifies the domain gap between COCO and TextCaps datasets. Finally, it is interesting that CIDEr metric (Vedantam, Lawrence Zitnick, and Parikh 2015) prefers a short caption and "Short-length Prompt" is even comparable to the best "COCO-style Prompt" in CIDEr metric, while SPICE metric (Anderson et al. 2016) prefers a longer caption and "High-length Prompt" clearly outperforms the strong "COCO-style Prompt" in SPICE. Visualization results of different prompts are shown in the next Section 4.4.

4.4 Qualitative Evaluation

Results on COCO (Lin et al. 2014). In Figure 3, we exhibit the captioning results on the COCO dataset. By feeding different prompts, our ConCap method is able to generate diverse captions including COCO-style [■], Positive [■], Negative [■], Short-length [■], Medium-length [■], and High-length [■]. Besides, we observe that the percentage of emotional captions is only about 2% of the entire COCO dataset. The proposed ConCap merely utilizes limited positive or negative captions in COCO to learn such styles. This is consistent with the observation that prompt learning is suitable for few-shot domain transfer (Liu et al. 2021). As shown in Figure 3, our ConCap is able to briefly describe an image or in a more detailed manner. The high-length captions [■] produced by our ConCap are much longer than the ground-truth captions and yield additional

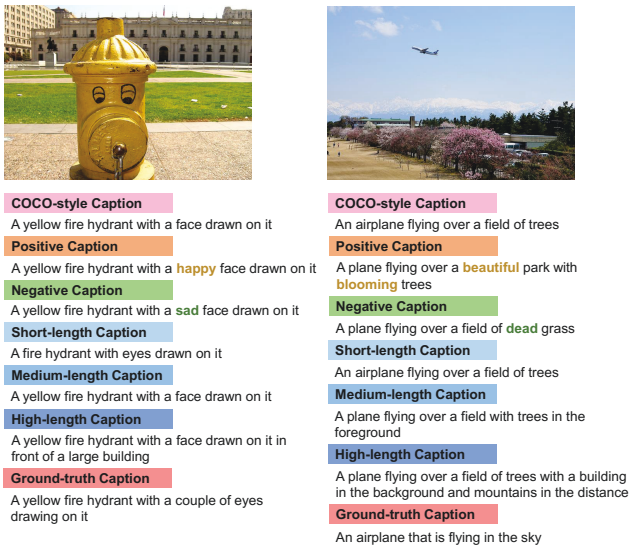

Figure 3: Image captioning examples from COCO (Karpathy and Fei-Fei 2015) with different styles including COCO-style [■], Positive [■], Negative [■], Short-length [■], Medium-length [■] and High-length [■].

meaningful semantics, e.g., "`a large building`" in the first image and "`mountains in the distance`" in the second image. Furthermore, our approach generates the positive words such as "`happy`, `beautiful`" [■] or the negative words such as "`sad`, `dead`" [■] to describe the same image in opposite personality traits. Since the COCO-caption dataset rarely contains the image with OCR contents, we showcase the results of "TextCap-style Prompt" on the TextCaps dataset in Figure 4.

Figure 4: Image captioning examples from TextCaps dataset (Sidorov et al. 2020) with different styles including COCO-style [■] and TexCap-style [■]. Best view in zoom in.

Results on TextCaps (Sidorov et al. 2020). Figure 4 exhibits the results on the TextCaps dataset, where we show the COCO-style [■] and TextCap-style [■] captions for style comparison. Different styles focus on different aspects of the image. For example, in the first image, the TextCap-style caption as well as the ground-truth annotation aim to de-

Method	Pre-training Data	COCO Caption Karpathy Test				NoCaps Validation							
						In-domain		Near-domain		Out-domain		Overall	
		B@4	M	C	S	C	S	C	S	C	S	C	S
BUTD (Anderson et al. 2018)	N/A	36.2	27.0	113.5	20.3	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
AoANet (Huang et al. 2019)	N/A	37.2	28.4	119.8	21.3	-	-	-	-	-	-	-	-
X-LAN (Pan et al. 2020)	N/A	38.2	28.8	122.0	21.9	-	-	-	-	-	-	-	-
$\text{Oscar}_{\text{base}}$ (Li et al. 2020)	7M	36.5	30.3	123.7	23.1	83.4	12.0	81.6	12.0	77.6	10.6	81.1	11.7
ViTCAP (Fang et al. 2021)	10M	36.3	29.3	125.2	22.6	98.7	13.3	92.3	13.3	95.4	12.7	93.8	13.0
$\text{VinVL}_{\text{base}}$ (Zhang et al. 2021)	9M	38.2	30.3	129.3	23.6	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5
$\text{LEMON}_{\text{base}}$ (Hu et al. 2021)	200M	40.3	30.2	133.3	23.3	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1
$\text{BLIP}_{\text{base}}$ (Li et al. 2022)	129M	39.7	-	133.3	23.3	111.8	14.9	**108.6**	**14.8**	111.5	14.2	109.6	14.7
$\text{SimVLM}_{\text{base}}$ (Wang et al. 2021b)	1.8B	39.0	32.9	134.8	24.0	-	-	-	-	-	-	94.8	13.1
ConCap (Ours)	129M	**40.5**	**30.9**	**133.7**	**23.8**	**113.4**	**14.9**	108.4	14.6	**113.2**	**14.4**	**110.2**	**14.8**

Table 5: Performance comparisons on the COCO Karpathy test split (Lin et al. 2014) and NoCaps validation split (Agrawal et al. 2019), where B@4, M, C, S denote BLEU@4, METEOR, CIDEr, and SPICE scores. For a fair comparison, all the methods only adopt the standard cross-entropy without CIDEr optimization.

scribe the words in the sign (e.g., "`heart break`") while ignoring the objects such as "`trash cans`". In contrast, the COCO-style pays more attention to the objects and environment, e.g., "`tall building`" in the second image.

4.5 Quantitative Evaluation

COCO (Lin et al. 2014). In Table 5, we present the performance of state-of-the-art captioning methods on the COCO-caption Karpathy test split (Karpathy and Fei-Fei 2015). Compared with the recent LEMON (Hu et al. 2021) that leverages more pre-training data, our method achieves superior performance. BLIP (Li et al. 2022) can be regarded as the baseline of our approach. Compared with BLIP, our ConCap outperforms it on all metrics. More importantly, the proposed ConCap is able to simultaneously handle other domains and generate captions with different lengths and styles for each image, which is infeasible for BLIP. The recent SimVLM approach (Wang et al. 2021b) leverages a large-scale pre-training corpus including 1.8 billion image-text pairs, which is $10\times$ larger than ours. Besides, SimVLM combines the ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2020) models as the visual extractor, which is stronger than our pure ViT structure.

NoCaps (Agrawal et al. 2019). NoCaps dataset covers more than 600 object categories and nearly 2/3 of them are unseen from the training set in COCO. The images in NoCaps are categorized into in-domain, near-domain, and out-of-domain based on whether these images are seen in the COCO training set. On this benchmark, we evaluate our ConCap using the "COCO-style" prompt. As shown in Table 5, the proposed ConCap outperforms all existing methods in terms of the overall performance, which verifies the generalizability of our method.

TextCaps (Sidorov et al. 2020). TextCaps is a recently proposed dataset containing 28K images and 145K captions, which is more challenging than COCO due to the existence of complex textual words. We compare the proposed method with the classic captioner such as AoANet (Huang et al. 2019) and the recent state-of-the-art methods including MMA-SR (Wang, Tang, and Luo 2020), CNMT (Wang et al. 2021a), and TAP (Yang et al. 2021).

Method	OCR Input	Validation		Test	
		B@4	C	B@4	C
BUTD (Anderson et al. 2018)	✗	20.1	41.9	14.9	33.8
AoANet (Huang et al. 2019)	✗	20.4	42.7	15.9	34.6
M4C (Hu et al. 2020)	✓	23.3	89.6	18.9	81.0
CNMT (Wang et al. 2021a)	✓	24.8	101.7	20.0	93.0
TAP (Yang et al. 2021)	✓	25.8	109.2	21.9	103.2
ConCap (Ours)	✗	**31.3**	**116.7**	**27.4**	**105.6**

Table 6: Comparsion results on the TextCaps validation set and test set (Sidorov et al. 2020), where B@4 and C denote BLEU@4 and CIDEr scores, respectively.

The comparison results are shown in Table 6. Our approach significantly outperforms the classic methods without pre-training such as AoANet (Huang et al. 2019). To the best of our knowledge, TAP (Yang et al. 2021) represents the current performance leader on the TextCaps dataset. TAP approach collects high-quality OCR-based image-text pre-training data, and performs the text-aware pre-training. Besides, TAP feeds the OCR detection results to the model, while our approach is free of such necessity. Without knowing the OCR results, our approach still surpasses the current state-of-the-art TAP method by a large margin of 7.5 CIDEr on the validation set. It is worth noting that our ConCap is not specially designed for TextCaps and is able to perform well on multiple domains including COCO, NoCaps, and TextCaps using a single model.

5 Conclusion

In this paper, we propose a conceptually simple yet effective prompt-based image captioning framework, which has been rarely investigated in the captioning community. By prompt engineering, the proposed approach is able to generate captions with diverse styles. To further explore the potential of prompt learning, we encourage the network to automatically learn the suitable prompt vectors in the continuous word embedding space. Extensive qualitative and quantitative experiments verify the effectiveness of the proposed framework.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *ICCV*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020a. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*.

Chen, T.; Zhang, Z.; You, Q.; Fang, C.; Wang, Z.; Jin, H.; and Luo, J. 2018. "Factual"or"Emotional": Stylized Image Captioning with Adaptive Learning and Attention. In *ECCV*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Universal image-text representation learning. In *ECCV*.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-memory transformer for image captioning. In *CVPR*.

Deng, C.; Ding, N.; Tan, M.; and Wu, Q. 2020. Length-controllable image captioning. In *ECCV*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Liu, Z.; Zeng, M.; et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387*.

Fang, Z.; Wang, J.; Hu, X.; Liang, L.; Gan, Z.; Wang, L.; Yang, Y.; and Liu, Z. 2021. Injecting semantic concepts into end-to-end image captioning. *arXiv preprint arXiv:2112.05230*.

Fei, Z. 2022. Attention-Aligned Transformer for Image Captioning. In *AAAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2021. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *ICCV*.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *AAAI*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *TACL*, 8: 423–438.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kobus, C.; Crego, J.; and Senellart, J. 2016. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1): 32–73.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; and Ji, R. 2021. Dual-level collaborative transformer for image captioning. In *AAAI*.

Mathews, A.; Xie, L.; and He, X. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-linear attention networks for image captioning. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6): 1137–1149.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; and Weston, J. 2019. Engaging image captioning via personality. In *CVPR*.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*.

Song, Z.; Zhou, X.; Mao, Z.; and Tan, J. 2021. Image captioning with context-aware auxiliary guidance. In *AAAI*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, J.; Tang, J.; and Luo, J. 2020. Multimodal attention with image text spatial relationship for ocr-based image captioning. In *ACM MM*.

Wang, Y.; Xu, J.; and Sun, Y. 2022. End-to-End Transformer Based Model for Image Captioning. In *AAAI*.

Wang, Z.; Bao, R.; Wu, Q.; and Liu, S. 2021a. Confidence-aware non-repetitive multimodal transformers for textcaps. In *AAAI*.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; and Huang, F. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804*.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florencio, D.; Wang, L.; Zhang, C.; Zhang, L.; and Luo, J. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *CVPR*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Zhang, W.; Shi, H.; Guo, J.; Zhang, S.; Cai, Q.; Li, J.; Luo, S.; and Zhuang, Y. 2022. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *AAAI*.

Zhong, Z.; Friedman, D.; and Chen, D. 2021. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*.