Transferable Decoding with Visual Entities for Zero-Shot Image Captioning

Junjie Fei[1*], Teng Wang[1,2*], Jinrui Zhang[1], Zhenyu He[3], Chengjie Wang[4,5], Feng Zheng[1†]
[1]Southern University of Science and Technology [2]The University of Hong Kong
[3]Harbin Institute of Technology (Shenzhen) [4]Tencent [5]Shanghai Jiao Tong University
junjiefei@outlook.com tengwang@connect.hku.hk zhangjr2018@mail.sustech.edu.cn
zhenyuhe@hit.edu.cn jasoncjwang@tencent.com f.zheng@ieee.org

Abstract

*Image-to-text generation aims to describe images using natural language. Recently, zero-shot image captioning based on pre-trained vision-language models (VLMs) and large language models (LLMs) has made significant progress. However, we have observed and empirically demonstrated that these methods are susceptible to modality bias induced by LLMs and tend to generate descriptions containing objects (entities) that do not actually exist in the image but frequently appear during training (*i.e.*, object hallucination). In this paper, we propose ViECap, a transferable decoding model that leverages entity-aware decoding to generate descriptions in both seen and unseen scenarios. ViECap incorporates entity-aware hard prompts to guide LLMs' attention toward the visual entities present in the image, enabling coherent caption generation across diverse scenes. With entity-aware hard prompts, ViECap is capable of maintaining performance when transferring from in-domain to out-of-domain scenarios. Extensive experiments demonstrate that ViECap sets a new state-of-the-art cross-domain (transferable) captioning and performs competitively in-domain captioning compared to previous VLMs-based zero-shot methods. Our code is available at:* https://github.com/FeiElysia/ViECap

1. Introduction

Large-scale pre-trained vision-language models (VLMs) like CLIP [37] and ALIGN [22] showcase impressive zero-shot transferability in various discriminative downstream tasks (*e.g.*, classification [37], segmentation [26, 49], and detection [16, 27]). However, effectively adapting these pre-trained VLMs into zero-shot generative tasks (*e.g.*, text and image generation) remains an open question that requires further exploration. Recently, some works [43, 42] have leveraged large language models (LLMs), *e.g.*, GPT [38, 5], to achieve CLIP-based zero-shot image-to-text generation. They follow a late-guidance paradigm where visual information is injected after completing word prediction. However, the weak visual guidance in this paradigm often results in modality bias, *i.e.*, the language prior in LLMs dominates the decoding process and therefore generates descriptions that are unrelated to the corresponding images. Fig. 1(a) shows incorrect predictions made by late-guidance decoding, *e.g.*, even if *"jump"* and *"cobblestone"* are unrelated to the image, they finally appear in the predictions due to their close association with predicted words *"skateboarder"* and *"snowy"*. Similarly, another example in Fig. 1(b) shows that *"donut"* and *"rocks"* are primarily generated by language prior instead of visual guidance.

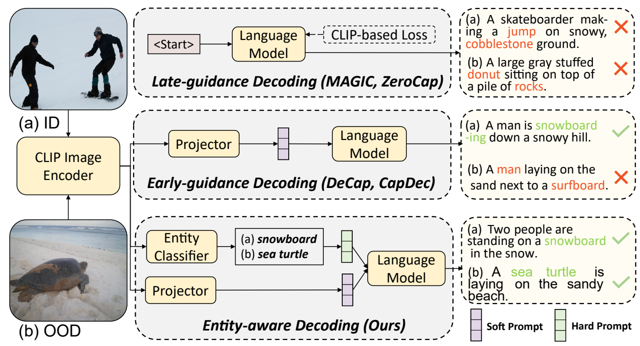

Figure 1. Decoding paradigm for zero-shot image-to-text generation. (a) in-domain (ID) image, (b) out-of-domain (OOD) image. ID refers to objects appearing in the image included in the training corpus, while OOD indicates that they are not included. ✗ and ✓ refers to the incorrect and correct predictions, respectively. Late-guidance methods generate descriptions irrelevant to the image, *e.g.*, *"jump"* and *"donut"*, while early-guidance models often tend to hallucinate objects that are not actually present in the OOD image, *e.g.*, *"surfboard"*. In contrast, our model utilizes entities as additional prompts to describe novel objects in the image, leading to superior transferability in OOD settings, *e.g.*, *"sea turtle"*.

Early-guidance methods [28, 35, 53] provide explicit

* Equal contribution † Corresponding author

guidance for word generation in LLMs by prefixing visual prompts to the text tokens. Typically, visual prompts are projected from the CLIP image embedding using a learnable projector. This early-guidance paradigm significantly alleviates the modality bias and boosts the alignments between the image and the generated captions. However, the learnable (soft) visual prompts are prone to overfitting when trained on a limited corpus, leading to poor performance in describing a diverse range of objects (visual entities). This, in turn, may cause object hallucination in the generated captions. Specifically, when transferring these models to unseen scenarios beyond the training corpus, novel entities are often misrecognized as similar entities frequently appearing in the training corpus. As Fig. 1 shows, early-guidance decoding is capable of understanding in-domain (ID) images but tends to hallucinate entities that do not actually exist in out-of-domain (OOD) images (*i.e.*, hallucinating *"sea turtle"* with *"surfboard"*, where *"surfboard"* frequently appears in the training corpus). Consequently, the transferability of the well-learned CLIP latent space is degraded into current decoding strategies, significantly limiting their applicability in real-world scenarios. We further validate the observed modality bias and object hallucination issues when adapting pre-trained VLMs and LLMs for image-to-text generation through experiments in Sec. 3.

To address the observed issues, we propose ViECap, which incorporates entity-aware hard prompts to compensate for the degradation of the CLIP latent space caused by learning soft prompts on a specific training corpus. This method is motivated by our observation that the CLIP-based entity classifier can accurately classify both ID and OOD images (*e.g.*, *"snowboard"* and *"sea turtle"* in Fig. 1). The entity-aware hard prompts enable transferable language decoding from the CLIP latent space. Fig. 1 shows that the proposed entity-aware decoding approach is capable of describing both seen and unseen entities in diverse images. Specifically, ViECap builds on early-guidance decoding methods, *e.g.*, CapDec [35]. Unlike these models, which can only describe entities present in the training corpus, our model can generate captions in diverse scenarios. Following CapDec, we train ViECap only using text data. The entity-aware hard prompt is the critical design enabling the transferability of our model to diverse captioning scenarios. The hard prompts, constructed by nouns extracted from texts during training or entities retrieved from images during inference, can prompt the LLMs to attend training-agnostic entities based on open vocabulary retrieval through CLIP. As we find that a naive integration of entities pushes ViECap to learn a copy-then-paste shortcut (*i.e.*, directly copying the entities to captions), we introduce a simple yet efficient entity masking strategy when incorporating the entity-aware hard prompts into early-guidance decoding.

We extensively evaluate ViECap on four benchmarks, NoCaps [1], COCO [30, 8], Flickr30k [52], and FlickrStyle10K [15]. The experimental results demonstrate that ViECap outperforms all other text-only methods and sets a new state-of-the-art in the cross-domain (transferable) setting while remaining competitive with them in the in-domain setting. In out-of-domain scenarios (NoCaps), we achieve a margin of 39.2 and 36.3 improvements, respectively, compared to DeCap and CapDec. We even surpass some supervised methods, indicating our model generalizes well to novel entities. In the experiment on FlickrStyle10K, ViECap effectively generates captions in different styles corresponding to the styles of the training set. Additionally, the data-efficient experiment shows ViECap's applicability in low-data settings, further highlighting its versatility and effectiveness across various scenarios.

To summarize, our contributions are as follows: 1) We shed light on the observations and underlying reasons behind the degraded generalizability when adapting pre-trained VLMs and LLMs into image-to-text generation, *i.e.*, modality bias and object hallucination, providing timely and valuable insights for pre-trained large-scale model adaptation. 2) We introduce entity-aware decoding to improve the transferability of zero-shot captioning. Specifically, aided by VLMs, we integrate entity-aware hard prompts with entity masking strategy into the decoding process, guiding LLMs to attend both seen and unseen entities. 3) Extensive experiments show the remarkable zero-shot transferability of ViECap, even in low-data scenarios.

2. Related Work

Supervision in Image Captioning. We classify image captioning models into supervised and unsupervised methods based on whether the image-text alignment information is provided during training. Supervised image captioning methods [50, 23, 21, 4, 11, 45, 7] are trained with paired (well-aligned) image-text data and typically adopt the encoder-decoder architecture. Initially, diverse vision backbones (*e.g.*, CNN [18], ViT [13]) are utilized to extract visual features, which are then fed into a language decoder (*e.g.*, LSTM [20], Transformer [44]) to generate coherent sentences. Various attention mechanisms [50, 21, 51, 7, 31] are commonly designed to capture vision-language alignment cues. However, the high cost associated with collecting paired image-text data limits the applicability of these models. In contrast, Unsupervised image captioning methods [25, 14] train the model using unpaired image-text data and primarily rely on visual concepts as anchor points to establish pseudo image-text alignment. Our proposed approach, on the other hand, only requires text data for model training. Compared to previous methods, our method further reduces the data collection cost while exhibiting superior efficiency by eliminating the need for image encoding during training.

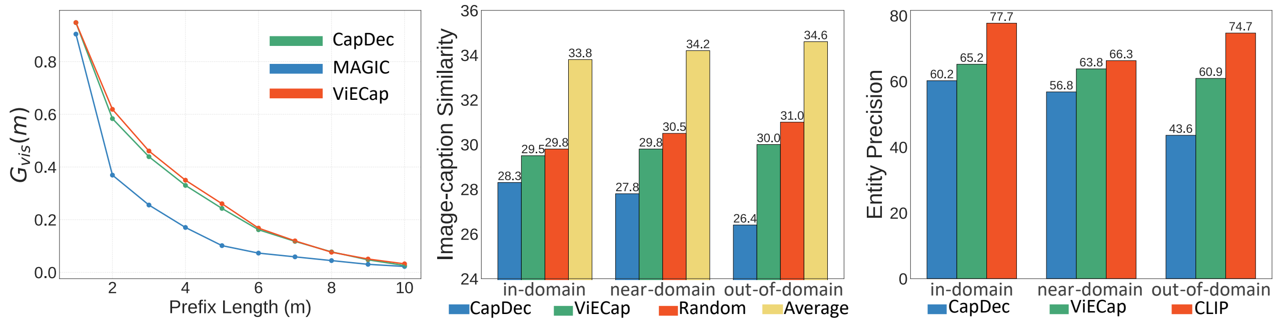

Figure 2. **Left**: Visual guidance with varied m on COCO. **Middle**: CLIP similarity between image and captions on NoCaps. "Random" refers to randomly sampling one ground-truth caption to calculate similarity scores with the paired image, and "Average" involves calculating the similarity scores by averaging the embeddings of all ground-truth captions. **Right**: Precision of detected entities using CLIP and captioning models on NoCaps. For both CapDec and ViECap, we evaluate the precision of entity words in the generated captions.

Zero-shot Image Captioning. Zero-shot image captioning aims to generate image captions without relying on human-annotated data [28]. Some methods [6, 47] in this area pre-train the model on large-scale weak image-text pairs and then evaluate the model on target benchmarks without further fine-tuning. Another set of methods [43, 35, 28, 42, 53] achieves zero-shot image captioning by combining large VLMs and LLMs. Specifically, VLMs provide vision-aware language guidance, which guides LLMs to generate image-related captions. We divide these methods into two paradigms: 1) late-guidance methods (ZeroCap [43] and MAGIC [42]) inject visual guidance after word prediction, and 2) early-guidance methods (SMs [53], CapDec [35], and DeCap [35]) retain visual information in several tokens using VLMs, prompting LLMs to generate image-aware words. Compared with these methods, we follow the early-guidance paradigm but integrate additional entity-aware hard prompts with an entity masking strategy, which significantly reduces the problem of object hallucination when describing images containing novel objects.

Novel Object Captioning This task aims to generate descriptions for images containing unseen objects during training [9, 2, 1, 32, 19, 46]. DCC [19] and NOC [46] leverage object recognition networks to recognize novel concepts. Other methods rely on object detectors (*e.g.*, Faster R-CNN [41], Mask R-CNN [17]) to recognize unseen entities in images [32, 29, 54]. Recently, captioning models leveraging the CLIP-based entity classifier [40, 10] have shown even more promising performance in describing novel concepts in images. Despite their success, these methods are trained on limited image-text pairs, making data collection challenging and rendering them susceptible to overfitting to the text style of the training corpus. Consequently, their capability to generate diverse sentences is restricted. In this study, we extend novel object captioning to a more data-efficient setting. Unlike the aforementioned methods, our model can seamlessly adapt to a new domain by simply fine-tuning with text-only data, thereby improving its transferability and diversity.

3. Empirical Observations

This section demonstrates the existence of modality bias and object hallucination when adapting VLMs and LLMs for image-to-text generation. It serves as a starting point for the proposed ViECap, which can address such limitations.

Modality Bias. A good captioning model should strike a balance between visual guidance and language contexts. To evaluate the influence of visual guidance, we design a two-stage decoding strategy: first, we use the captioning model (*e.g.*, MAGIC [42], CapDec [35]) to generate the first m words of the caption, then we feed these prefix words into a pre-trained language model to obtain the subsequent words based on pure language contexts. The accuracy of generated captions, measured by CIDEr [45], is denoted as CIDEr(m). We define the importance of visual guidance $G_{\text{vis}}(m)$ as:

$$G_{\text{vis}}(m) = 1 - G_{\text{lang}}(m) = 1 - \frac{\text{CIDEr}(m)}{\text{CIDEr}_{\text{model}}} \quad (1)$$

where $\text{CIDEr}_{\text{model}}$ is the accuracy of captions generated without a pure language model (*i.e.*, m equals sentence length). $G_{\text{lang}}(m)$ is the importance of language contexts.

If a captioning model is dominated by language priors, $G_{\text{vis}}(m)$ will be small as a pure language model can accurately predict captions. As Fig. 2 (left) shows, the late-guidance method MAGIC gains much lower $G_{\text{vis}}(m)$ compared to early-guidance methods CapDec and our ViECap, especially when m is small. This observation confirms modality bias towards language in late-guidance methods.

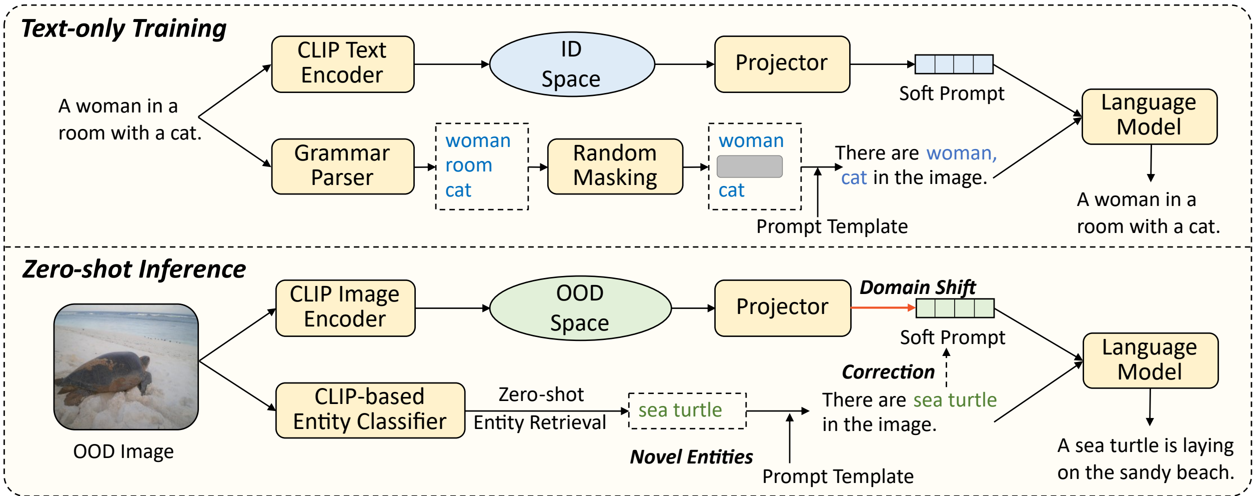

Figure 3. The overview of the proposed ViECap framework. During training, with text-only corpus, nouns are extracted from the sentence by a grammar parser to construct the hard prompt. Then, the soft prompt encodes the overall contexts of the sentence by CLIP text encoder followed by a learnable projector. Two types of prompts are concatenated together as the input for the language model to predict captions. During inference, given a test image, we input the CLIP image embedding into the projector to obtain the soft prompt and introduce a CLIP-based entity classifier to construct the entity-aware hard prompt. With the strong transferability from the training-agnostic hard prompt, our framework is robust to the shift of image domain, achieving excellent captioning performance in both ID and OOD images.

Object Hallucination. While early-guidance decoding alleviates the problem of modality bias effectively, previous models still show limited generalizability towards OOD images containing novel concepts. To illustrate the degradation of transferability in current methods, we calculate the cosine similarity using CLIP between the image and the corresponding generated caption. Fig. 2 (middle) shows that CapDec experiences a gradual performance drop when transferring from ID to OOD settings, while our ViECap exhibits a more robust capability in describing images with different domains.

Object hallucination leads to incorrect entities in the generated caption. We further analyze the precision of entities detected by different methods in Fig. 2 (right). While the CLIP embedding shows remarkable transferability, it is degraded in the caption generation process of CapDec. The accuracy of CapDec drops significantly (60.2 $\rightarrow$ 43.6) when transferring from ID to OOD images. By explicitly introducing visual entities, ViECap demonstrates the capability to describe both seen and unseen entities in images. Specifically, the accuracy of correctly detecting entities decreases slightly by 4.3 compared to the accuracy predicted by CLIP, which is a reduction of 3.

4. ViECap

The proposed ViECap is a transferable captioning framework based on CLIP and trained on a text-only corpus. Specifically, We train a language decoder to decode the CLIP text embedding of sentences and incorporate entity-aware prompts to enable transferable captioning. For zero-shot inference, we directly feed the CLIP image embedding of a given image into the trained decoder to generate captions. Fig. 3 illustrates the overall framework of ViECap.

4.1. Entity-aware Transferable Decoding

Given the text-only data, our goal is to train an entity-aware language decoder with promising transferability. To this end, we extract two types of visual-aware guidance from the ground-truth caption: 1) nouns in the caption, which serve as anchors for grounding entities in the image. These nouns (*i.e.*, discrete category names) are capable of capturing salient and static visual cues, such as humans, animals, and objects. 2) CLIP text embedding of the caption, which is implicitly aligned with the image embedding, provides the overall contexts across all images, such as scenes and interactions between objects. We transform entities and the text embedding into prompt tokens to guide the language model (*i.e.*, GPT-2) in predicting captions. During training, we freeze the parameters of the CLIP text encoder to maximize its transferability. We train the projector from scratch and fine-tune the language model using an auto-regressive loss (details can be found in the Appendix).

Hard Prompt. We first construct a vocabulary of entities, denoted as $\mathcal{V}$. Nouns in the caption, regarded as visual entities, are recognized by the NLTK grammar parser and filtered by this vocabulary. The extracted entities are then inserted into a prompt template "There are e_1, ..., e_N in the image.", where e_n refers to the nth entity. The entity-aware

COCO ⇒ NoCaps val									
Methods	Pre-trained Model	in-domain		near-domain		out-of-domain		Overall	
		CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Paired image-text training, zero-shot inference									
$\text{OSCAR}_{\text{Base}}$ [29] ECCV'20	Faster R-CNN + $\text{BERT}_{\text{Base}}$	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
ClipCap [34] ArXiv'21	ViT-B/32 + $\text{GPT-2}_{\text{Large}}$	84.9	12.1	66.8	10.9	49.1	9.6	65.8	10.9
$\text{I-Tuning}_{\text{Base}}$ [33] ICASSP'23	ViT-B/16 + $\text{GPT-2}_{\text{Base}}$	83.9	12.4	70.3	11.7	48.1	9.5	67.8	11.4
SmallCap* [40] CVPR'23	ViT-B/32 + $\text{GPT-2}_{\text{Base}}$	83.3	-	77.1	-	65.0	-	75.8	-
Text-only training, zero-shot inference									
DeCap* [28] ICLR'22	ViT-B/32 + Transformer	**65.2**	-	47.8	-	25.8	-	45.9	-
CapDec† [35] EMNLP'22	ViT-B/32 + $\text{GPT-2}_{\text{Base}}$	60.1	10.2	50.2	9.3	28.7	6.0	45.9	8.3
ViECap ICCV'23	ViT-B/32 + $\text{GPT-2}_{\text{Base}}$	61.1	**10.4**	**64.3**	**9.9**	**65.0**	**8.6**	**66.2**	**9.5**

Table 1. Cross-domain captioning results on the NoCaps validation set. †represents our re-implemented results. * refers to the use of a **memory bank**. Note that SmallCap reports results on the NoCaps test set, while other methods report results on the NoCaps validation set.

Method	COCO ⇒ Flickr30k				Flickr30k ⇒ COCO			
	B@4	M	C	S	B@4	M	C	S
MAGIC [42]	6.2	12.2	17.5	5.9	5.2	12.5	18.3	5.7
DeCap [28]	16.3	17.9	35.7	11.1	12.1	18.0	44.4	10.9
CapDec [35]	17.3	**18.6**	35.7	-	9.2	16.3	27.3	-
ViECap	**17.4**	18.0	**38.4**	**11.2**	**12.6**	**19.3**	**54.2**	**12.5**

Table 2. Cross-domain captioning results on the Flickr30k test set and COCO test set. All methods in this table use text-only training.

hard prompt is constructed by a training-agnostic module, enabling strong robustness to the domain shift from ID to OOD images.

Soft Prompt. We first inject Gaussian noise into the CLIP text embedding to alleviate the modality gap as indicated in CapDec [35]. A trainable projector then transforms the CLIP text embedding to generate the soft prompt. The projector is implemented as a lightweight transformer with L learnable queries as in ClipCap [34]. The output features of L query tokens are considered as the soft prompt.

Entity masking. We observe that naively integrating nouns to construct hard prompts tends to learn a copy-then-paste shortcut during training, where all nouns are input together to generate a caption, *i.e.*, the model simply pastes the input nouns without making any modification. Consequently, the captioning prediction task becomes trivial, and the model's generalizability is severely impaired, particularly when confronting incorrect entities during inference. To address this issue, we propose a simple yet effective entity masking strategy that randomly drops a certain proportion of nouns with the masking ratio r_{mask} during training. This strategy significantly alleviates the learning collapse and boosts captioning performance in both ID and OOD settings. The effectiveness of the masking strategy is verified in Tab. 6.

4.2. Zero-shot Inference

Once the decoder is trained, we can leverage it for zero-shot captioning inference. Given a test image, we first extract its visual embedding using the CLIP image encoder. We then employ the trained projector to convert the visual embedding into the soft prompt. For the hard prompt, we again use visual embedding for entity classification. Specifically, we use the manual template "A photo of {entity}" as the entity description for each category in $\mathcal{V}$. Then we rank and select the top M entities with the highest similarity scores between different entity descriptions and the visual embedding to construct the entity-aware hard prompt. Finally, the soft prompt and hard prompt are concatenated together in sequential order and input into the language model to predict the caption auto-regressively.

It should be noted that there exists a training-inference gap in the model structure. Two strategies during training are adopted to address this gap and improve the model performance. Firstly, we use noisy text embedding to bridge the gap between visual and text embedding. Secondly, we propose a non-trivial entity masking mechanism to avoid the copy-then-paste shortcut, meanwhile pushing the model to recover the missing entities from soft prompts.

Transferability on OOD images. Trained on limited ID data, the projector may overfit to the ID dataset, leading to a significant performance degradation of the soft prompt for OOD inputs. In contrast, the entity-aware hard prompt, predicted by the frozen CLIP, inherits the powerful transferability from CLIP embeddings. The GPT could flexibly combine the soft and entity-aware hard prompts for a better trade-off between ID and OOD performance.

5. Experiments

We conduct extensive experiments to evaluate the performance of ViECap in diverse zero-shot image captioning settings, including 1) cross-domain captioning, 2) in-domain captioning, and 3) data-efficient captioning. The experiments are organized as follows: In Sec. 5.1, we assess the transferability of ViECap through the cross-domain setting. Here, the model is trained on a corpus from the source domain and evaluated on a target domain. In Sec. 5.2, we focus on the generalizability of ViECap under the in-domain scenario, where the model is trained and evaluated on the same dataset. We conduct data-efficient experiments to assess the applicability of our model in low-data scenarios in Sec. 5.3. In Sec. 5.4, we perform various ablation experiments to assess the effectiveness of entity-aware decoding. Furthermore, we qualitatively evaluate ViECap in Sec. 5.5.

Implementation Details. We use CLIP-ViT-B/32 as our backbone. The language model is GPT-2_{base} implemented by Wolf et al. [48]. The projector comprises an 8-layer transformer with 8 attention heads and a hidden size of 768. The length of learnable soft prompts is set to 10. During training, we freeze the CLIP text encoder and only train GPT-2 and projector using AdamW [24] optimizer for all experiments. For caption generation, we use beam search with a beam size of 5. Details are shown in the Appendix.

Datasets and Metrics. We conduct experiments on four widely used image captioning benchmarks, *i.e.*, NoCaps [1], COCO [30, 8], Flikcr30k [52], and FlickrStyle10K [15]. For COCO and Flickr30k, we follow the commonly used Karpathy split [23]. For NoCaps, we train our model on the COCO training set and report the results on the validation set, as suggested by OSCAR [29]. As for FlickrStyle10K, we follow MemCap [56], randomly sampling 6,000 captions as our training set and using the remaining image-text pairs for testing. We report results with common used captioning metrics BLEU@n (B@n) [36], METEOR (M) [12], CIDEr (C) [45] and SPICE (S) [3]. Refer to the Appendix for details about these datasets.

Methods. We include several captioning methods as follows: 1) BUTD [4] and OSCAR [29] as classic supervised methods, 2) ClipCap [34], I-Tuning [33], and SmallCap [40] as lightweight paired captioning methods that utilize GPT-2 for CLIP-based captioning, 3) ZeroCap [43] as a training-free method, 4) StyleNet [15] and MemCap [56] as classic methods for style captioning, and 5) MAGIC [42], CapDec [35], and DeCap [28] as text-only training methods, which are in line with our work. Specifically, MAGIC employs late-guidance decoding. CapDec and DeCap adopt early-guidance decoding, which learns soft prompts for caption generation. Notably, DeCap leverages an additional memory bank, and CapDec serves as our baseline.

5.1. Cross-Domain Captioning

In this section, we demonstrate the transferability of ViECap in cross-domain captioning. We evaluate ViECap's ability to describe novel entities in images by training it on the COCO training set and testing on the NoCaps validation set without any additional fine-tuning. As Tab. 1 shows, ViECap outperforms all other text-only methods by a large margin and even achieves competitive performance compared to some supervised methods in the *out-of-domain* and *Overall* setting, indicating that incorporating the entities-aware hard prompt is beneficial for the model to describe unseen entities. While other methods experience a notable drop in CIDEr score from the *in-domain* to *out-of-domain* setting in NoCaps, ViECap maintains a minimal fluctuation across different domains, showcasing the remarkable transferability of our model. In real-world scenarios, the target domain of images is typically agnostic, making the evaluation based on the *Overall* results a better reflection of the models' effectiveness in practical applications. Surprisingly, with text-only corpus, ViECap achieves comparable results to supervised methods, obtaining a CIDEr score of 66.2 compared to 63.8 for OSCAR and 65.8 for ClipCap on the *Overall*. Furthermore, ViECap significantly outperforms the unpaired methods DeCap and CapDec by a large margin of 20.3 CIDEr, demonstrating our model can gener-

In-Domain Captioning

Method	COCO B@4	COCO M	COCO C	COCO S	Flickr30k B@4	Flickr30k M	Flickr30k C	Flickr30k S
Paired image-text training								
BUTD [4] CVPR'18	36.2	27.0	113.5	20.3	27.3	21.7	56.6	16.0
OSCAR [29] ECCV'20	36.5	30.3	123.7	23.1	-	-	-	-
ClipCap [34] ArXiv'21	33.5	27.5	113.1	21.1	-	-	-	-
I-Tuning [33] ICASSP'23	34.8	28.3	116.7	21.8	25.2	22.8	61.5	16.9
SmallCap* [40] CVPR'23	37.0	27.9	119.7	21.3	-	-	-	-
Text-only training, zero-shot inference								
ZeroCap [43] CVPR'22	7.0	15.4	34.5	9.2	5.4	11.8	16.8	6.2
MAGIC [42] ArXiv'22	12.9	17.4	49.3	11.3	6.4	13.1	20.4	7.1
DeCap* [28] ICLR'22	24.7	25.0	91.2	**18.7**	21.2	**21.8**	**56.7**	**15.2**
CapDec [35] EMNLP'22	26.4	**25.1**	91.8	-	17.7	20.0	39.1	-
ViECap ICCV'23	**27.2**	24.8	**92.9**	18.2	**21.4**	20.1	47.9	13.6

Table 3. In-domain captioning results on the COCO test set and Flickr30k test set. * denotes using a **memory bank**. It should be noted that the result of ZeroCap is copied from MAGIC, and the results of OSCAR and I-Tuning are from their base backbone.

Method	Romantic B@1	Romantic B@3	Romantic M	Romantic C	Humorous B@1	Humorous B@3	Humorous M	Humorous C
StyleNet [15]	13.1	1.5	4.5	7.2	13.4	0.9	4.3	11.3
MemCap [56]	21.2	4.8	8.4	22.4	19.9	4.3	7.4	19.4
CapDec [35]	21.4	5.0	9.6	26.9	**24.9**	6.0	10.2	34.1
ViECap	**25.7**	**6.5**	**10.4**	**33.6**	24.3	**6.5**	**10.4**	**35.0**

Table 4. In-domain captioning results on the FlickrStyle10K.

Data	Method	COCO Test	NoCaps val In	Near	Out	Overall
0.1%	CapDec	24.0	13.2	11.0	6.2	10.4
	ViECap	**32.3**	**20.9**	**27.6**	**34.9**	**30.2**
1%	CapDec	55.8	29.6	20.5	9.8	18.9
	ViECap	**63.9**	**34.6**	**39.9**	**39.3**	**40.4**
10%	CapDec	**83.6**	47.3	39.8	19.1	35.4
	ViECap	83.4	**45.9**	**51.8**	**48.7**	**53.3**
100%	CapDec	92.7	60.1	50.2	28.7	45.9
	ViECap	**92.9**	**61.1**	**64.3**	**65.0**	**66.2**

Table 5. Data-efficient captioning results.

Method	COCO Test	NoCaps val In	Near	Out	Overall
Baseline	92.7	60.1	50.2	28.7	45.9
+ Entity	53.2	32.4	40.3	53.1	44.8
+ Entity + Masking (20%)	88.1	54.1	60.5	63.5	62.7
+ Entity + Masking (40%)	92.9	**61.1**	**64.3**	**65.0**	**66.2**
+ Entity + Masking (60%)	**94.6**	59.1	64.0	63.9	65.5
+ Entity + Masking (80%)	94.5	57.8	**64.3**	63.5	65.3

Table 6. Ablation studies of entity masking.

Method	COCO Test	NoCaps val In	Near	Out	Overall
Baseline (Soft-only)	92.7	60.1	50.2	28.7	45.9
Entity-only	61.4	29.9	41.0	49.8	44.4
Entity + Soft	92.5	60.8	63.6	**65.1**	65.7
Soft + Entity (w/o ensemble)	92.3	57.3	59.5	59.1	61.0
Soft + Entity	**92.9**	**61.1**	**64.3**	65.0	**66.2**

Table 7. Ablation studies of prompts.

ate captions with stable quality in diverse domains.

Tab. 2 showcases results in more cross-domain settings, where ViECap sets a new state-of-the-art performance on all metrics from Flickr30k to COCO and on most metrics from COCO to Flickr30k.

Both Tab. 1 and Tab. 2 demonstrate the remarkable zero-shot transferability of our model. ViECap is capable of describing images that deviate from the distribution of the training sets, as well as those that do not, making it highly useful when applied in real-world scenarios.

5.2. In-Domain Captioning

To further assess the generalizability of ViECap, we conduct evaluations on COCO, Flickr30k, and FlickrStyle10K in the in-domain setting, where the training and testing data are from the same dataset. As shown in Tab. 3, our proposed model outperforms CapDec, our baseline method, in most metrics. We attribute this improvement to our entity-aware hard prompt, which explicitly emphasizes infrequent object concepts, thereby mitigating the long-tail problem associated with the existing dataset. It is worth noting that DeCap utilizes a large memory bank to bridge the modality gap, which may not be practical in real-world scenarios. In contrast, our approach achieves comparable performance with an acceptable memory complexity, highlighting the effectiveness of the proposed ViECap. Tab. 4 shows that ViECap achieves state-of-the-art performance on Flick-Style10K. These results demonstrate that ViECap can adapt well to diverse style text data, showcasing its versatility and strong generalizability.

5.3. Data-Efficient Captioning

In this section, we explore ViECap's capability to learn from low-data scenarios. Specifically, we randomly sample different scales of data from the COCO training set to train ViECap. For simplicity, we leverage In, Near, and Out to denote *in-domain*, *near-domain*, and *out-of-domain*, respectively. As shown in Tab. 5, ViECap outperforms CapDec across all data scales. Even with as little as 0.1% of the data, ViECap can still generate reasonable captions and maintain transferability (CIDEr: 32.3 on the COCO testing set vs. 30.2 on the *Overall* of NoCaps validation set), suggesting ViECap is data-efficient and applicable in low-data settings.

5.4. Ablation Studies

We conduct comprehensive ablation studies to explore the influence of entities and prompt structures on our model. Additionally, we evaluate the advantages of using a larger-scale language model for ViECap. In all these experiments, we assess the performance of our model on both the COCO testing set and the NoCaps validation set.

Masking Rate. We begin by investigating the impact of entities on ViECap by randomly masking entities at different rates. As shown in Tab. 6, incorporating entities enhances our model's ability to perceive unseen entities (from 28.7 to 53.1). However, as mentioned before, the model may learn a shortcut from detected entities, leading to a rapid decline in ID performance (from 92.7 to 53.2). The CIDEr score on the COCO testing set gradually increases as the masking rate increases, indicating that entity masking can prevent ViECap from relying heavily on entities. For the NoCaps validation set, the performance of ViECap first increases and then decreases, showing that a moderate entity masking rate benefits caption prediction in unseen scenarios. The results of this experiment demonstrate that the proposed entity masking strategy boosts the captioning performance of ViECap across diverse scenes.

Prompts. We then explore the impact of different prompt generation methods on the performance of ViECap. As

Figure 4. Visualization of generated captions of some images from the out-of-domain setting on the NoCaps validation set, which contain unseen entities. GT refers to the Ground Truth, CapDec and Ours denote caption generated by CapDec and ViECap, respectively.

Model	pt.	#para	**COCO** Test	**NoCaps val** In	Near	Out	Overall
Tuned language model							
$\mathrm{GPT}_{\mathrm{Base}}$ (4-layer)	×	67M	89.8	54.5	54.2	50.2	54.6
$\mathrm{GPT}_{\mathrm{Base}}$	√	124M	92.9	61.1	64.3	65.0	66.2
Frozen language model							
$\mathrm{GPT}_{\mathrm{Base}}$	√	124M	88.0	57.7	60.2	61.4	62.0
$\mathrm{GPT}_{\mathrm{Large}}$	√	774M	91.7	59.4	64.4	68.4	66.9
$\mathrm{GPT}_{\mathrm{XL}}$	√	1.5B	94.5	63.4	66.6	68.9	68.9
$\mathrm{OPT}_{1.3\mathrm{B}}$	√	1.3B	95.6	64.9	68.7	69.9	70.7
$\mathrm{OPT}_{2.7\mathrm{B}}$	√	2.7B	96.9	64.7	70.2	71.9	72.1

Table 8. ViECap with different scales of language models. "pt." represents using pre-trained weights.

shown in Tab. 7, learning only soft prompts leads to overfitting to in-domain captioning, resulting in poor performance when describing novel entities. Incorporating hard prompts improves captioning performance for unseen images, but solely modeling entity information leads to reduced performance on in-domain captioning. When combining soft and hard prompts, we achieve comparable ID performance with soft prompts-only methods and superior performance on OOD scenarios. Additionally, we find that the order of soft and hard prompts does not affect ViECap's performance. To construct hard prompts with visual entities, we leverage CLIP-based retrieval, where the accuracy of retrieval benefits from the prompt ensemble [1].

Scaling up ViECap. We assess diverse pre-trained LLMs, ranging from GPT-2 to OPT [55], to investigate the impact of scaling up language models on the capability of ViECap to describe novel entities in images. The results are shown in Tab. 8. As the model parameters increase, the performance of ViECap continues to improve. Notably, we freeze the language model for more effective training. To our surprise, ViECap can effectively leverage the information from the language model without the need for further fine-tuning.

[1]The prompt engineering is released by OpenAI. https://github.com/openai/CLIP/blob/main/notebooks

5.5. Qualitative Evaluation

Fig. 4 displays ground truth captions and captions generated by CapDec and ViECap trained on COCO. Images are from the *out-of-domain* set in the NoCaps validation set, which contains unseen entities. In the first image in Fig. 4, ViECap correctly recognizes the specific entity "jay" while CapDec mistakenly identifies a generic entity "bird". Similar outcomes can be observed in the other instances shown in the figure. CapDec exhibits object hallucination in its textual descriptions, while ViECap demonstrates the ability to generate high-quality descriptions of entities in novel scenarios, illustrating the effectiveness of using hard prompts to guide LLMs in attending to entities in images. More visualization results can be found in the Appendix.

6. Conclusion

In this paper, we have comprehensively investigated the challenges of adapting pre-trained VLMs and LLMs for image-to-text generation. Our empirical findings demonstrate the existence of modality bias and object hallucination, highlighting the limited transferability when adapting pre-trained models to downstream tasks and providing valuable insight for further research in this area. We propose an entity-aware decoding approach to address the observed issues. By leveraging the CLIP latent space to prompt GPT-2 during caption generation, our method, ViECap, demonstrates remarkable performance in both seen and unseen scenarios. Extensive experiments showcase that our ViECap outperforms existing zero-shot methods in transferability and achieves competitive performance on zero-shot in-domain captioning. Moreover, ViECap proves to be data-efficient in low-data settings on COCO. Experiment on FlickrStyle10K shows that our model can also generate captions in different styles based on the training corpus style.

Acknowledgments. This work was supported by the National Key R&D Program of China (Grant NO. 2022YFF1202903) and the National Natural Science Foundation of China (Grant NO. 62122035).

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 2, 3, 6

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016. 3

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398, 2016. 6

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

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 3

[7] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 2

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 6

[9] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015. 3

[10] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. Universal captioner: Inducing content-style separation in vision-and-language model training. *arXiv preprint arXiv:2111.12727*, 2021. 3

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

[12] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 6

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019. 2

[15] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3137–3146, 2017. 2, 6

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[19] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016. 3

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

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

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning*, pages 4904–4916, 2021. 1

[23] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2, 6, 13

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424, 2019. 2

[26] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 1

[27] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[28] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning. *International Conference on Learning Representations*, 2023. 1, 3, 5, 6

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137, 2020. 3, 5, 6, 13

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755, 2014. 2, 6

[31] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 2

[32] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 3

[33] Ziyang Luo, Zhipeng Hu, Yadong Xi, Rongsheng Zhang, and Jing Ma. I-tuning: Tuning frozen language models with image for lightweight image captioning. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 5, 6

[34] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 5, 6

[35] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 1, 2, 3, 5, 6

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International conference on machine learning*, pages 8748–8763, 2021. 1

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

[39] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588, 2019. 14

[40] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023. 3, 5, 6

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[42] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 1, 3, 5, 6

[43] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*, 2021. 1, 3, 6

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[45] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2, 3, 6

[46] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2017. 3

[47] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3

[48] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 6

[49] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 1

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning*, pages 2048–2057, 2015. 2

[51] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[52] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2, 6

[53] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 1, 3

[54] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 3, 13

[55] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 8

[56] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. Memcap: Memorizing style knowledge for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12984–12992, 2020. 6, 13