

Zero-shot Image Captioning by Anchor-augmented Vision-Language Space Alignment

Junyang Wang Yi Zhang Ming Yan Ji Zhang Jitao Sang

November 15, 2022

Abstract

CLIP (Contrastive Language-Image Pre-Training) has shown remarkable zero-shot transfer capabilities in cross-modal correlation tasks such as visual classification and image retrieval. However, its performance in cross-modal generation tasks like zero-shot image captioning remains unsatisfied. In this work, we discuss that directly employing CLIP for zero-shot image captioning relies more on the textual modality in context and largely ignores the visual information, which we call *contextual language prior*. To address this, we propose Cross-modal Language Models (CLMs) to facilitate unsupervised cross-modal learning. We further propose Anchor Augment to guide the generative model’s attention to the fine-grained information in the representation of CLIP. Experiments on MS COCO and Flickr 30K validate the promising performance of proposed approach in both captioning quality and computational efficiency.

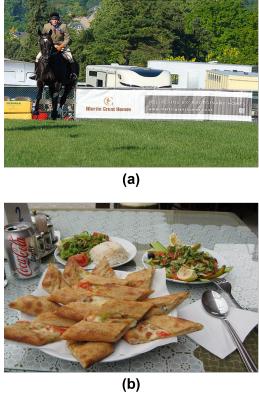
1 Introduction

Vision-Language Pre-training (VLP) has advanced the research of multi-modal modeling in recent years [28, 5, 17, 16], among which CLIP [22] has drawn increasing attention for its transferable visual representation learning. Benefiting from contrastive learning on a large-scale web image-text dataset, CLIP independently encodes images and text and maps them into a vision-language space with common semantics, thus making the zero-shot transfer between the two modalities possible [33, 36, 26, 24, 34, 10]. Impres-

sive zero-shot image classification capability (76.2% accuracy on ImageNet) was demonstrated by CLIP [22].

The zero-shot classification ability of CLIP has encouraged research on zero-shot image captioning. Existing CLIP-based zero-shot image captioning approaches [29, 27] use a language model by the means of next-token prediction method to first suggest candidate words and then calculate the representation similarities of CLIP between each candidate word and image to select the generated word. These approaches use two constraints: one on constraining the generated caption to be as similar as possible to the given image in the CLIP representation space, and the other on constraining the generated caption to have as low a loss as possible on the language model. It is easy to see that the second constraint is only imposed on the textual modality. We discuss that this uni-modal candidate word-centric solution is prone to produce *contextual language prior*: The language model suggests candidate words only based on the context of the generated caption, by exploiting the prior activated in the language model. While image captioning is essentially a cross-modal task, the above contextual language prior issue can make the caption generation ignore the visual information of the given image and thus generates irrelevant content as illustrated in Figure 1.

Cross-modal learning is thus needed in CLIP-based zero-shot image captioning. The problem turns to address the unavailability of supervised cross-modal data under zero-shot settings. We are inspired by the fact that CLIP aligns visual and textual representations in cross-modal embedding space, and propose



- Reference:** A horseback rider is somewhat in the background of the field.
- ZeroCap:** Image of a 2013 derby **riding** on top of Berlin-based tour-style track.
- MAGIC:** A view of a **busy city street** with a **horse riding event**.
- Ours:** A person **riding a horse** in front of a car.
- Reference:** A meal containing soda, salad pizza and rice on a table.
- ZeroCap:** Image of a photourn (meal). Photo taken 2006.
- MAGIC:** A **table with food** and a glass of wine.
- Ours:** A pizza sitting on a table next to a fork.

Figure 1: We show different captions generated by three zero-shot image captioning approaches: ZeroCap [29], MAGIC [27], and ours. The red-colored words refer to relevant information, while the blue-colored words refer to irrelevant information that cannot be inferred from the image. Observations include: (1) ZeroCap fails to capture the relevant information due to the overpowering contextual language prior; (2) MAGIC manages to encode the relevant information, but generates irrelevant information such as the “street” due to “busy” and the “glass of wine” due to “table” and “food” via ungrounded correlation from contextual language.

Cross-modal Language Models (CLMs) to transform uni-modal data in cross-modal representations to facilitate unsupervised cross-modal learning. Specifically, we first use CLIP to obtain the representations of the sentences in the unsupervised corpus and place the representations in the first position of the language model as prefix tokens. And then we use the original sentences as self-supervised labels for auto-regressive language modeling training.

With the cross-modal learning of CLIP representation, CLMs manages to employ both textual and visual modalities for caption generation at a global level. However, the fine-grained information in cross-modal representation of CLIP is still not fully exploited. To improve the attention of the generative model to the fine-grained information, we propose Anchor Augment. In the training phase, we extract nouns from the sentences as anchors and use them for CLMs. We hope that anchors can be used as cues for the fine-grained information in the representation of CLIP. To improve the robustness of anchors, we further introduce anchor random dropout: dropping out all anchors in a training sample with a certain probability. In the generation phase, we use the object detector to extract the labels of the objects in the images as anchors, which improves the quality of caption generation.

We summarize the contributions as follows:

- We position the contextual language prior issue in CLIP-based zero-shot image captioning and propose Cross-modal Language Models (CLMs) to address it from unsupervised data in an auto-regressive fashion. By CLMs, the generative model learns cross-modal knowledge through the representation of CLIP.
- We propose Anchor Augment to further improve the attention of the generative model to the fine-grained information in representation of CLIP. We extract anchors from the original data and use them in CLMs training to guide the generative model.
- Our approach achieves superior results than existing zero-shot image captioning approaches on

both MS COCO and Flickr 30K. In addition, it has a significant advantage in generative speed.

2 Background and Related Work

2.1 CLIP

By contrastive learning on a dataset of 400 million (image, text) pairs, CLIP [22] can encode data from visual and textual modalities independently into a common vision-language semantic space. This means that similar image and text are aligned in this space, thus bridging the gap between visual and textual modalities and enabling the learning of transferable cross-modal representation. With the strong generalization obtained in the pre-training phase, CLIP can be used for zero-shot tasks such as classification, retrieval, etc [22]. The zero-shot performance is claimed to be close to or even better than fine-tuned models [22]. Many works have applied the zero-shot ability of CLIP to specific application scenarios such as image segmentation [33], image generation [24], and object detection [36].

2.2 Zero-Shot Image Captioning

Early implementations of this task relied on visual features extracted by convolutional neural network[31, 32]. [2] pioneered the use of an object detector and a bottom-up approach to extract visual features, which lead to a better attention mechanism for the model. With the rise of vision-language pre-training models, researchers have commonly adopted the approach of learning cross-modal universal knowledge by using large-scale image-text pairs in the pre-training phase and fine-tuning on the image captioning datasets[28, 5, 17, 16, 6, 19].

As the development of supervised image captioning is greatly limited due to the high data collection costs, unsupervised approaches attract a lot of attention. The powerful zero-shot ability of CLIP has encouraged research on the zero-shot image captioning. [29] proposes to employ the native GPT-2 [23] to

provide candidate words for the next position based on the context of generated text. The generated captions and each candidate word are then combined into a new set of candidate captions. Finally, the CLIP similarity between these candidate captions and the given image is calculated separately, and the candidate caption with the highest similarity is selected as the newly generated caption at the current position. [27] improves upon this by introducing target domain contextual language prior. They first fine-tune the GPT-2 on the unsupervised caption corpus of the target domain. Then, they combined fine-tuned GPT-2 and CLIP to compute similarity scores for different word candidates for a given image, and selected the word with the highest similarity score as the content of the caption.

The above works use the GPT-2 to generate candidate words based not on the image but only on the uni-modal context of generated text. This results in heavy influence caused by the *contextual language prior*, which may generate captions that do not correspond to the images. We discuss that, for a cross-modal task such as image captioning, the generative model needs to adequately consider both the image and the context together. This motivates us to achieve a cross-modal understanding of the generative model.

3 Method

In this section, we first describe the zero-shot image captioning and introduce the vision-language pre-training dual-encoder model CLIP and language model GPT-2 employed by our approach, which are used to extract cross-modal representations and textualize cross-modal representations, respectively. We propose a cross-modal language modeling framework, which does not need to use any supervised data in training, and generates captions based on the cross-modal representations extracted by CLIP (Section.3.2). Based on the framework, we propose Anchor Augment (Section.3.3) to improve the attention to the fine-grained information. The overview of our approach is shown in Figure 2.

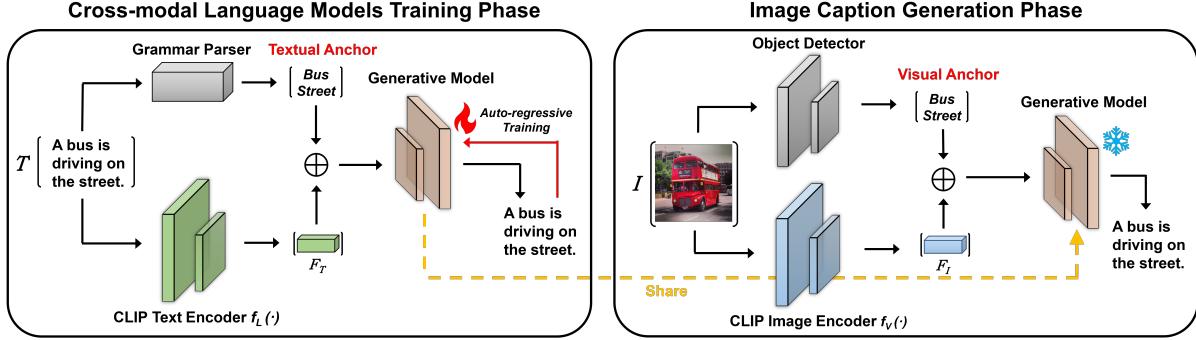


Figure 2: The overview of our approach. In the language model training phase, we use Cross-modal Language Models (CLMs) with anchors to train a generative model that can generate the training text based on textual representation extracted by CLIP text encoder and anchors extracted by a grammar parser. The process is self-supervised and without supervised data. In the caption generation phase, we use the trained generative model to generate the caption depending on the visual representation extracted by CLIP image encoder and anchors extracted by a object detector.

3.1 Preliminaries

Notations. We first describe supervised data and unsupervised data in the image captioning problem. The supervised dataset $\mathcal{D}_s = \{(x_1, y_1), \dots, (x_n, y_n)\}$ consisting of n pairs with images x_i and reference captions $y_i = \{c_i^1, \dots, c_i^{|y_i|}\}$, where y_i is a set of the captions that describe the x_i from different perspectives, and c_i^j denotes the j_{th} caption of y_i . The un-supervised data includes unlabeled image dataset $\mathcal{D}_u^I = \{x_1, \dots, x_i\}$ and text datasets $\mathcal{D}_u^T = \{y_1, \dots, y_j\}$. Traditional image captioning approaches use supervised dataset \mathcal{D}_s for training, while zero-shot image captioning only assumes the availability of unlabeled dataset \mathcal{D}_u^I and \mathcal{D}_u^T .

Base models. CLIP is a VLP model with dual-encoder architecture. It consists of two independent encoders for visual and textual modalities. Similarities between vision and language representations on large-scale image-text pairs are used to pre-train CLIP, bridging the gap between vision-language semantics in the representation space of CLIP. The sim-

ilarity is calculated as

$$\begin{aligned} F_I &= f_V(I) \\ F_T &= f_L(T) \end{aligned} \quad (1)$$

$$\text{Similarity}(I, T) = \cos < F_I, F_T > = \frac{F_I}{|F_I|} \cdot \frac{F_T}{|F_T|}$$

where f_V and f_L is the image encoder and text encoder of CLIP respectively.

GPT-2 is a transformer-based language model trained on a large-scale text corpus by an auto-regressive pre-training task. It learns the relationship of sentence context by the next-token prediction. The pre-trained GPT-2 can be fine-tuned on downstream datasets and thus used for continuous specific text generation, such as literature abstract, dialogue, story, etc. The pre-training loss of GPT-2 is the Maximum Likelihood Estimation (MLE) and calculated as

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|T|} \sum_{i=1}^{|T|} \log D_\theta(T_i | T_1 T_2 \dots T_{i-1}) \quad (2)$$

where θ denotes the parameter that needs to be optimized for model D .

3.2 Cross-modal Language Models by Auto-regressive Training

The visual and textual representations of CLIP are aligned through the contrastive learning on large-scale image-text pairs. That is, the textual representation on CLIP can be seen as the visual representation of the matched image on CLIP, even though the CLIP representations of visual modality and textual modality are separately encoded by two uni-modal encoders. Inspired by this, we propose the Cross-modal Language Models (CLMs) to learn to textualize understanding for visual representations without any image training data. The overview of CLMs training is shown on the left of Figure 2.

GPT-2 architecture is employed as the generative model D_θ to generate the textual description for CLIP representations. Based on text encoder $f_L(\cdot)$ of CLIP, we first obtain the representation F_T of the training text T . Then, we put the representation into the position of the first token of GPT-2 and generate the original text by the constraint of auto-regressive loss. Note that the token dimension of the GPT-2 architecture we are using is the same as the CLIP representation dimension, which allows the representation of CLIP to be fed into GPT-2 like a normal token. For settings with inconsistent dimensions, an additional adapter module is required to unify the dimensions. The training loss is formulated as follows

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|T|} \sum_{i=1}^{|T|} \log D_\theta(T_i | F_T T_1 T_2 \dots T_{i-1}) \quad (3)$$

The input format is formulated as follows

$$[\text{cls}][F_T][\text{sep}][T_1] \dots [T_n][\text{cls}] \quad (4)$$

where $[\text{cls}]$ denotes a special token used at the beginning and the end of the input and $[\text{sep}]$ denotes a special token for splitting different parts of the input. For example, if the training text is “A man with a red helmet is riding a motorbike on a dirt road”, the input will be $[\text{cls}][F_T][\text{sep}][A][\text{man}][\text{with}] \dots [a][\text{dirt}][\text{road}][\text{cls}]$.

3.3 Improving the Attention with Anchor Augment

To improve the attention of the generative model to the key information in the cross-modal representation of CLIP, we propose Anchor Augment. We want to explicitly supply the generative model with the fine-grained information in the input data so that the generative model can focus more attention on the information. Therefore, we extract the key information from the input text or image as anchors and apply them to the CLMs training. To make the anchors accurate, we directly use a grammar parser to obtain nouns in the training text as anchors as shown in the left of Figure 2. After applying the anchors, the training loss is shown by the following equation

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|T|} \sum_{i=1}^{|T|} \log D_\theta(T_i | \mathbf{F}_T \mathbf{A}_1 \dots \mathbf{A}_n T_1 \dots T_{i-1}) \quad (5)$$

where A_n denotes the n_{th} anchor in the sentence.

The input format is formulated as follows

$$[\text{cls}][F_T][\text{sep}][A_1] \dots [A_n][\text{sep}][T_1] \dots [T_n][\text{cls}] \quad (6)$$

For example, if the training text is “A man with a red helmet is riding a motorbike on a dirt road” whose anchor is “man”, “helmat”, “motorbike” and “road”, the input will be $[\text{cls}][F_T][\text{sep}][\text{man}][\text{helmat}][\text{motorbike}][\text{road}][\text{sep}][A][\text{man}][\text{with}] \dots [a][\text{dirt}][\text{road}][\text{cls}]$.

In the generation phase, we use an object detector to extract the labels of the image’s ROI (Region of Interest). Since the anchors extracted by the object detector are noisy, the model needs to be robust to anchors. To achieve this, we use anchor random dropout. We drop out all anchors with q probability randomly for each training text. In the caption generation phase, we do not use the anchor random dropout.

3.4 Zero-shot Image Captioning

The overview of the image caption generation phase is shown on the right of Figure 2. Similar to the

training phase, the input images are fed into two modules: image encoder $f_V(\cdot)$ of CLIP and object detector. The image encoder is used to obtain the CLIP representation of the image and the object detector is used to obtain the anchors from the image. For anchor extraction, we set a confidence threshold p for the object detector. Only objects with a confidence greater than p are used for generation. A larger value of p may lose information in the image, while a smaller value of p may introduce noise. After confidence threshold filtering, we extracted the labels of these ROIs as visual anchors. Since the inputs of the training and generation phases are identical in format, we directly share the generative model from CLMs training without any other training. The input format is formulated as follows

$$[cls][F_I][sep][A_1] \dots [A_n][sep] \quad (7)$$

We leave the position of the self-supervised label $[T_1] \dots [T_n][cls]$ empty. Finally, the model generates the image caption in these positions based on the visual representation of CLIP and anchors.

4 Experiment

In this section, we conduct quantitative and qualitative experiments to evaluate our approach. In the subsection of quantitative experiments, we first compare the performance of our approach with the baselines on various evaluation metrics of image captioning. Then, we conduct ablation experiments to prove the significance of our approach design. Finally, we analyze the effect of the cross-modal transfer by quantitatively observing the performance of the generative model on text generation and image caption generation during the training phase. In the subsection of qualitative experiments, we analyze some example captions generated by the baselines and our approach.

Evaluation Benchmarks. We conduct experiments on two widely used benchmarks: MS-COCO [18] and Flickr30k [21]. For both datasets, we set up the training, validation, and test splits according to the splits of Karpathy et al [13].

Implementation Details. We use the text in the training split as corpus for CLMs training. We optimize the generative model with the Adam optimizer [14] and a learning rate of 5e-7. Notably, this procedure is computationally negligible, i.e., less than 3 hours with 1 NVIDIA 3080 GPU. We decide the epoch number of training based on the performance of the model on the validation set. For object detector, we choose the mainstream model Faster-RCNN [25]. For the confidence threshold p of the object detector, we chose three values of 0.5, 0.7, and 0.9. For the probability q of anchor random dropout, we chose 0, 0.25, 0.5, 0.75, and 1. For the approach of search, we choose beam search, where the branches of the beam are chosen as 5 which is as same as other approaches.

Baselines. First we choose some weakly supervised approaches UIC [11], IC-SME [15], S2S-SS and S2S-GCC [12]. The starting point of these approaches is to address the problem of data limitation, which is similar to the unsupervised approaches. However, the weakly supervised approaches still require an amount of supervised data to generate unsupervised data. Then we compare with a CLIP-based approach, called CLIPRe. Given an image, it retrieves the most related caption from a caption corpus based on the image-text similarity as measured by CLIP. We also compare three unsupervised approaches from the last year, ZeroCap [29], MAGIC [27], and SMs [35].

Evaluation Metrics. Following the common practice in the literature, we perform evaluation using BLEU-1 (B@1), BLEU-4 (B@4) [20], METEOR (M) [9], ROUGE-L (R-L) [18], CIDEr [30].

4.1 Quantitative Experiments

4.1.1 Image Captioning

Table 1 shows the results on zero-shot image captioning. First, we observe that the unsupervised approaches based on CLIP have better performances compared to weak supervision approaches. This is because the weak supervision approaches need to expand the amount of data according to a small amount of supervision data. Therefore, the weak supervision

Approach	MS-COCO						Flickr30k					
	B@1	B@4	M	R-L	CIDEr	SPICE	B@1	B@4	M	R-L	CIDEr	SPICE
<i>Weakly Supervised Approach</i>												
UIC [11]	41.0	5.6	12.4	28.7	28.6	8.1	-	-	-	-	-	-
IC-SME [15]	-	6.5	12.9	35.1	22.7	-	-	7.9	13.0	32.8	9.9	-
S2S-SS [12]	49.5	6.3	14.0	34.5	31.9	8.6	-	-	-	-	-	-
S2S-GCC [12]	50.4	7.6	13.5	37.3	31.8	8.4	-	-	-	-	-	-
<i>Unsupervised Approach</i>												
CLIPRe [27]	39.5	4.9	11.4	29.0	13.6	5.3	38.5	5.2	11.6	27.6	10.0	5.7
ZeroCap [29]	49.8	7.0	15.4	31.8	34.5	9.2	44.7	5.4	11.8	27.3	16.8	6.2
SMs [35]	-	6.9	15.0	34.1	44.5	10.1	-	-	-	-	-	-
MAGIC [27]	56.8	12.9	17.4	39.9	49.3	11.3	44.5	6.4	13.1	31.6	20.4	7.1
Ours	59.3	15.0	18.7	41.8	55.7	10.9	58.3	16.8	16.2	39.6	22.5	9.8

Table 1: Image captioning performances of different approaches on MS-COCO and Flickr30k, where the B@1, B@4, M, and R-L represent BLEU@1, BLEU@4, METEOR, and Rouge-L respectively.

Model	MS-COCO \Rightarrow Flickr30k						Flickr30k \Rightarrow MS-COCO					
	B@1	B@4	M	R-L	CIDEr	SPICE	B@1	B@4	M	R-L	CIDEr	SPICE
CLIPRe	38.7	4.4	9.6	27.2	5.9	4.2	31.1	3.0	9.9	22.8	8.5	3.9
MAGIC	46.4	6.2	12.2	31.3	17.5	5.9	41.4	5.2	12.5	30.7	18.3	5.7
Ours	49.2	10.1	12.5	33.8	12.7	5.7	47.6	7.7	14.9	35.9	38.5	8.2

Table 2: Cross-Domain Evaluation. X \Rightarrow Y means source domain \Rightarrow target domain.

Model	Speed (second)
ZeroCap	76.8
MAGIC	2.89
Ours	1.46 (OD Phase) + 0.27 (ICG Phase)

Table 3: The caption generation speed of different approaches on a sample. We count the speed of our approach for the object detection (OD) phase and image caption generation (ICG) phase separately.

approach does not really get rid of the data limitations. Second, we observe that the performances of generation-based approaches (ZeroCap, MAGIC, and ours) are far better than the retrieval-based approach (CLIPRe). The retrieval-based approach is limited by the capacity of the text corpus, so it can hardly generate the caption for unseen samples. There is also a problem with the supervision approaches, but the supervision approaches can make

the model fit the data so that the model has a certain generalization of the unseen sample. Finally, our approach achieved excellent performance of all approaches, with 11 of its 12 metrics being the best. Not only that, we have a significant improvement on some metrics (such as the B@4 and SPICE on Flickr30K). This is due to the fact that our approach makes the generative model truly attend to the content in the image through cross-modal learning, unlike other approaches that are based on contextual language prior.

Generalization. We also focus on the generalization ability. The result of the cross-domain evaluation is shown in Table 2. We apply the generative model trained on the training text corpus of the source domain to perform inference on the test set of the target domain. The experimental results show that our approach has a strong cross-domain ability.

Efficiency. We also pay attention to generative efficiency. The experimental results of generative efficiency are shown in Table 3. We can observe that

MS-COCO						Flickr30k				
q	0	0.25	0.5	0.75	1	0	0.25	0.5	0.75	1
$p = 0.5$	31.7	32.9	32.5	32.0		19.0	26.5	27.1	27.0	
$p = 0.7$	32.3	33.7	33.6	33.0		18.6	25.2	27.2	26.8	
$p = 0.9$	31.8	33.4	33.5	32.9	24.6	18.7	25.2	27.2	26.8	26.4
$p = 1$	11.6	20.9	22.3	22.0		18.0	25.8	26.7	26.6	

Table 4: We have chosen different values of p and q for the ablation analysis of our approach. The values in the table represent the average of all evaluated metrics. p represents the confidence threshold of the object detector. The objects with confidence greater than this threshold are used as anchors. Where $p = 1$ means that no anchor is provided in the generation phase. q represents the probability of anchor random dropout in the training phase. $q = 1$ means that no anchor is added in the training phase. In order to align the settings of the two phases, the generation phase in this case also does not provide anchors, so the value is independent of p .

our approach is much faster than the ZeroCap. That is because ZeroCap involves computationally inefficient operations like gradient updates [7, 29]. This efficiency seriously limits its actual application of it. We observe that our approach is more efficient than baselines and the time cost in the image caption generation phase is almost negligible. The efficiency of our approach is mainly limited by the object detector. With the development of faster object detectors, our approach holds promise for use in real-time scenarios.

4.1.2 Ablation Experiments

Table 4 shows the results of the ablation experiments. We probe the influence of anchors on the model by setting the confidence threshold p of the object detector and the probability q of anchor random dropout. For the convenience of observation, we calculate the mean values of the 6 image captioning evaluation metrics.

Confidence Threshold of Object Detector.

First, we probe the influence of p by comparing the value of the same column. From the results, we can see that the performance difference is not significant at p equal to 0.5, 0.7, and 0.9. This verifies that the model is robust to the anchors by anchor random dropout. Even if we provide noisy anchors by a low p for the model in the generation phase, the model is still able to select the ones relevant to the CLIP representation from these anchors. This is because

anchor random dropout makes it mandatory for the model to learn the connection between CLIP representation and anchors. However, when we set p to 1, i.e., we do not provide any anchor to the model in the generation phase, the performance of the model will be significantly declined. This shows that using anchors only in the training phase does not effectively improve the performance of the model.

Probability of Anchor Random Dropout.

Second, We probe the influence of q by comparing the value of the same line, where q equals 0 means no anchor dropout is used. It is worth noting that q equals 1 means that there is no anchor to participate in training. In this case, we no longer provide anchors for the model in the generation stage. We use a value to represent the performance because in this case, the performance is independent of p . We observe that both without the use of anchor random dropout and without anchors involved in the training phase, the performance decreased. This illustrates both the need for the existence of anchor augmentation and the need for caution in the use of anchors. When there is no anchor, the model can only extract information from the CLIP representation that has a low dimension compared to a sentence, which increases the difficulty of generation. And when anchors are fully used, the model creates a dependency on the anchors and thus ignores the information in the CLIP representation. Therefore, an appropriate dropout probability is particularly important. From

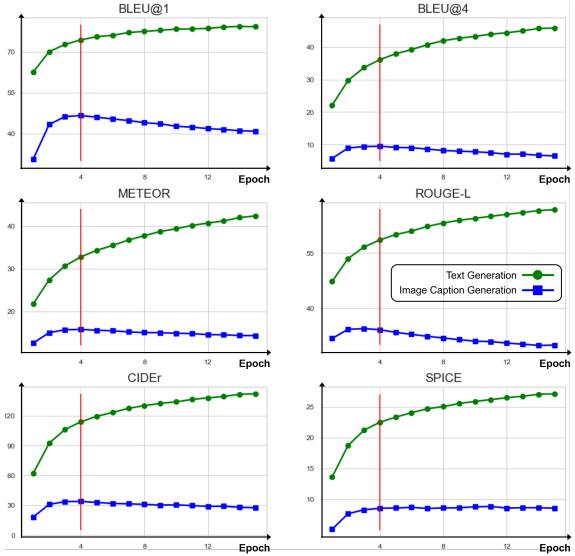


Figure 3: The performance for text generation and image caption generation varies with the number of training epochs.

the results, a dropout probability of 0.5 is a desirable setting. With this setting, the model is able to learn the information in CLIP representation and anchors in a balanced way.

4.1.3 Analysis of Cross-modal Transfer

Without the supervised data, we achieve the cross-modal transfer of knowledge through the vision-language space of CLIP. However, the vision-language space of CLIP has been shown to contain biases [3, 8, 1]. We hope the cross-modal transfer to gravitate towards non-biased knowledge as possible. Therefore, we wanted to understand the influence of CLMs training on the cross-modal transfer of knowledge. We apply the generative model for text generation and image caption generation, respectively, where the input format for image caption generation is consistent with Equation 7 and for text generation is the replacement of F_I with F_T in Equation 7. We calculate the generation quality for both cases separately based on the reference caption and plot line graphs of the different metrics with the epochs

of training. The results are shown in Figure 3. We observe that at the early stage of training (before the red line), CLMs training greatly promotes the cross-modal transfer of knowledge. However, as the training progresses (after the red line), CLMs inhibit cross-modal transfer. This means we need to stop CLMs training at an appropriate point.

4.2 Qualitative Experiments

Figure 4 shows visual comparisons between our approach and the two zero-shot baselines along with the reference caption. The results demonstrate that our approach can generate fluent captions.

We can clearly see that ZeroCap is almost failing to generate quality captions. The generated captions contain a lot of unsupported content. Although the authors argue that ZeroCap’s generation style makes sense for unsupervised image captioning because it can greatly enrich the diversity of captions, we believe that such unsupported content is somewhat risky. For example, by inputting an image, ZeroCap runs the risk of generating some private content, such as the location where it was taken.

The reason why MAGIC and our approach are much better in this respect is that we use clean text for training, which allows effective control over the generation style of the generative model. However, MAGIC’s lack of cross-modal learning makes its language model only a uni-modal model. This means that the caption generated by MAGIC contains severe context language prior. For example in Figure 4 a, there is a context language prior between the elephant and the grass in the language training corpus, which is also reflected in the generated results. When the model generated the “elephant”, the subsequent generation did not consider the fact that the elephant is standing on the ground in the image but generated the “grassy” based on the experience of the language model. In all of the examples of Figure 4, our approach achieves accurate generation. In addition, our approach is more accurate in identifying objects. For example, in Figure 4 b, our approach accurately identifies the “baby” that is not identified by other approaches. This is made possible by the use of anchors.

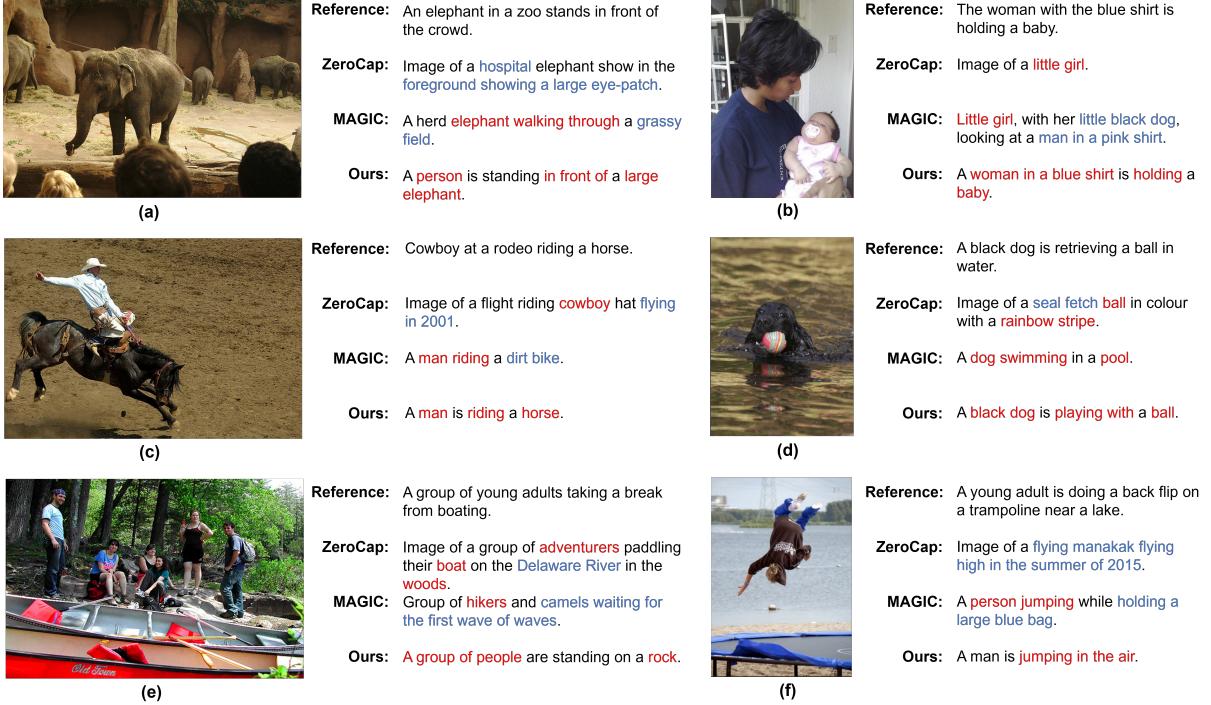


Figure 4: We show more captions generated by three zero-shot image captioning approaches: ZeroCap, MAGIC, and ours. The red-colored words refer to relevant information, while the blue-colored words refer to irrelevant information that cannot be inferred from the image.

$I_{man} = \left[\begin{array}{ c } \hline \text{Image of a man} \\ \hline \end{array} \right]$	$D(I_{man}) :$ An image of a man wearing a checkered patterned tie.
$T_{man} = \left[\begin{array}{ c } \hline \text{A photo of} \\ \hline \text{a man} \\ \hline \end{array} \right]$	$D(I_{man} - T_{man} + T_{woman}) :$ An image of a woman wearing a floral patterned neck tie.
$T_{woman} = \left[\begin{array}{ c } \hline \text{A photo of} \\ \hline \text{a woman} \\ \hline \end{array} \right]$	
$I_{dog} = \left[\begin{array}{ c } \hline \text{Image of a dog} \\ \hline \end{array} \right]$	$D(I_{dog}) :$ This is a close up of a pug dog with a frisbee .
$T_{dog} = \left[\begin{array}{ c } \hline \text{A photo of} \\ \hline \text{a dog} \\ \hline \end{array} \right]$	$D(I_{dog} - T_{dog} + T_{cat}) :$ This is a very cute cat that is getting ready to use a tennis racket .
$T_{cat} = \left[\begin{array}{ c } \hline \text{A photo of} \\ \hline \text{a cat} \\ \hline \end{array} \right]$	

Figure 5: Illustration for explainable bias study by cross-modal counterfactual operation. The blue and red font represent the difference between the captions generated before and after counterfactual operation, respectively.

5 Discussion

Currently, CLIP has been widely used as the base model for various tasks. All along, researchers have tried to figure out what CLIP learned from the web data. [4] observed that a subspace of concepts can be obtained by performing linear operations on the word vectors in the embedding space. We first use this property to perform a linear operation on the textual representation of CLIP to obtain a subspace that represents a concept change to another concept. Then, by applying this subspace to the visual representation of an image, a counterfactual representation is obtained. We use the example of detecting the stereotypes in CLIP as shown in Figure 5. First, we obtain the subspace of *man* to *woman* and *dog* to *cat* by linear operations. Then, we generate captions for the original representations and counterfactual representations, respectively. We observe that there are attributes that should be not related to the concept of subspace change. For example, when we change the gender concept in the image from *man* to *woman*, the *checkered* change to *floral*, even though the pattern of

the tie should be not related to the gender. This suggests spurious stereotypes in CLIP. Accordingly, our approach can be used to assess/interpret dependencies and connections between concepts in CLIP-style models, and can be generalized as a tool to provide new insights to explore the modeling of information in opaque VLP models.

6 Conclusion

In this paper, we propose Anchor-augmented Vision-Language Space Alignment for zero-shot image captioning. To avoid the unintentional introduction of contextual language priors caused by uni-modal language models in previous approaches, we propose Cross-modal Language Models training task. In addition, to improve the attention of the generative model to the fine-grained information in the cross-modal representation of CLIP, we propose Anchor Augment. The training process of our approach is efficient and does not require annotated data. The experiment results demonstrate that our approach can achieve SOTA in generation quality compared with other baselines. Also because of the non-query-based architecture, our approach achieves a high generation efficiency.

References

- [1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 9
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3
- [3] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022. 9
- [4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016. 11
- [5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 1, 3
- [6] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022. 3
- [7] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019. 8
- [8] Nassim Dehouche. Implicit stereotypes in pre-trained classifiers. *IEEE Access*, 9:167936–167947, 2021. 9
- [9] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 6
- [10] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, 2022. 1
- [11] 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. 6, 7
- [12] Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. *arXiv preprint arXiv:2104.13872*, 2021. 6, 7
- [13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 6
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [15] 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. 6, 7
- [16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 3
- [17] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 3
- [18] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004. 6
- [19] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3
- [20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

- [21] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [23] 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. 3
- [24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3
- [25] 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. 6
- [26] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022. 1
- [27] 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, 2, 3, 6, 7
- [28] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 1, 3
- [29] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022. 1, 2, 3, 6, 7, 8
- [30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6
- [31] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 3
- [32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 3
- [33] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 1, 3
- [34] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3637–3645, 2022. 1
- [35] 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. 6, 7
- [36] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1, 3