# ConZIC: Controllable Zero-shot Image Captioning by Sampling-Based Polishing

Zequn Zeng*, Hao Zhang*, Ruiying Lu, Dongsheng Wang, Bo Chen†
National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, China
{zzequn99, zhanghao_xidian}@163.com, bchen@mail.xidian.edu.cn

Zhengjue Wang
State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, 710071, China
zhengjuewang@163.com

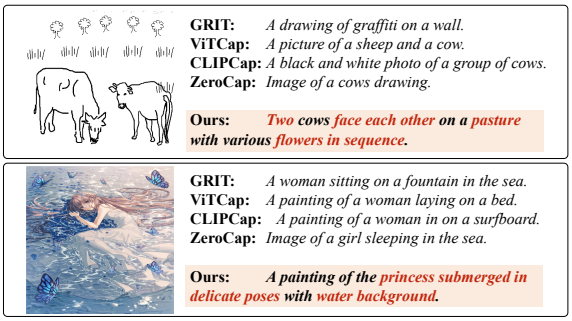
## Abstract

*Zero-shot capability has been considered as a new revolution of deep learning, letting machines work on tasks without curated training data. As a good start and the only existing outcome of zero-shot image captioning (IC), ZeroCap abandons supervised training and sequentially searches every word in the caption using the knowledge of large-scale pre-trained models. Though effective, its autoregressive generation and gradient-directed searching mechanism limit the diversity of captions and inference speed, respectively. Moreover, ZeroCap does not consider the controllability issue of zero-shot IC. To move forward, we propose a framework for* ***Con****trollable* ***Z****ero-shot* ***IC****, named* ***ConZIC****. The core of ConZIC is a novel sampling-based non-autoregressive language model named GibbsBERT, which can generate and continuously polish every word. Extensive quantitative and qualitative results demonstrate the superior performance of our proposed ConZIC for both zero-shot IC and controllable zero-shot IC. Especially, ConZIC achieves about 5× faster generation speed than ZeroCap, and about 1.5× higher diversity scores, with accurate generation given different control signals. Our code is available at https://github.com/joeyz0z/ConZIC.*

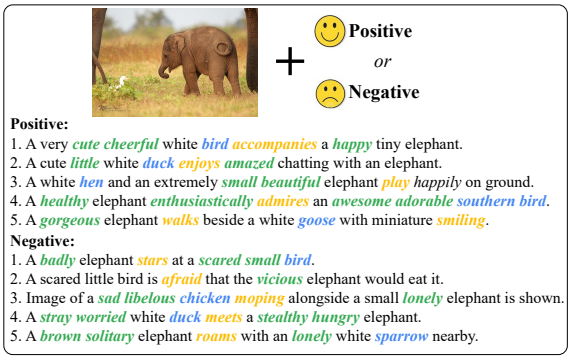
## 1. Introduction

Image captioning (IC) is a visual-language task, which targets at automatically describing an image by generating a coherent sentence. By performing supervised learning on human-annotated datasets, such as MS-COCO [43], many methods [22, 33, 49, 50] have achieved impressive evaluation scores on metrics like BLEU [52], METEOR [7], CIDERr [66], and SPICE [3]. However, these methods still lag behind human capability of zero-shot IC.

Specifically, those supervised methods extremely rely on well-designed image-captions pairs. However, it is likely impossible to construct a large enough dataset, including paired images and high-quality captions covering various styles/contents. As a result, it is challenging for the machine to caption images that are outliers with respect to the training distribution, which is common in real applications (see




(a) Examples of zero-shot image captioning.




(b) Diversity of ConZIC.

Figure 1. The highlights of our proposed method. (a) shows two examples of zero-shot image captioning on several SOTA methods. Specifically, GRIT [50] and ViTCAP [22] are two supervised methods without pre-trained models. ClipCap [49] is a supervised method using pre-trained CLIP. GRIT, ViTCAP, and CLIPCap are firstly trained on MSCOCO and then do testing. ZeroCap [65] is the zero-shot method without any training. (b) shows the diversity of our proposed ConZIC, which manifests two aspects: *semantic* (diverse words: different colors denoting different parts-of-speech) and *syntactic* (diverse sentence patterns).

*Equal contribution. †Corresponding authors

examples in Fig. 1a). On the contrary, humans can perform IC without any specific training, *i.e*., realizing zero-shot IC. Because humans can integrate what they see, *i.e*., the image, and what they know, *i.e*., the knowledge.

Recently, large-scale pretraining models have shown a strong capability of learning knowledge from super-large-scale data, showing great potential in various downstream tasks [10, 27, 54, 57, 63]. Equipped with the visual-language knowledge learned by CLIP [57] and linguistic knowledge from GPT-2 [58], ZeroCap [65] is the first and the only zero-shot IC method, which proposes a searching-based strategy and is free of training on extra supervised data. Specifically, ZeroCap searches the caption words one by one and from left to right, guided by CLIP-induced score for image-text matching and GPT-2 word distribution for caption fluency. ZeroCap is a good start and inspires us to explore how to search for the optimal caption in a better way.

*i) More flexible.* ZeroCap utilizes GPT-2 to perform left-to-right autoregressive generation. Once a word is fixed, there is no chance to modify it when we move to the next position. In other words, such generation order is not flexible enough to consider the full context information.

*ii) More efficient.* The searching at every position is realized by iteratively updating the parameters of GPT-2, which is time-consuming, as shown in Fig. 3c.

*iii) More diverse.* IC is an open problem. Given an image, different persons may have different visual attentions [14] and language describing styles [24, 47, 73], thus resulting in diverse descriptions. ZeroCap employs beam search to generate several candidate sentences, which, however, have similar syntactic patterns (see Appendix D).

*iv) More controllable.* To endow captioning models with human-like controllability, *e.g*., sentiment, personality, a recent surge of efforts [12, 19, 24, 47] resort to introducing extra control signals as constraints of the generated captions, called Controllable IC. However, controllable zero-shot IC has not been explored yet.

Bearing all these four-aspect concerns in mind, we propose a novel framework for controllable zero-shot IC, named ConZIC, as shown in Fig. 2. Specifically, after analyzing the relationship between Gibbs sampling and masked language models (MLMs, currently we use BERT) [11, 20, 70], we firstly develop a new language model (LM) called Gibbs-BERT to realize the zero-shot IC by sampling-based search. Compared with autoregressive models, Gibbs-BERT has more a flexible generation order, bringing the self-correct capability by bidirectional attention with faster and more diverse generations. After integrating Gibbs-BERT with the CLIP that is used to evaluate the similarity between image and text, our proposed framework can perform zero-shot IC. By further introducing a task-specific discriminator for control signal into our framework, our proposed framework can perform controllable zero-shot IC.

The main contributions of this paper are:

- We propose to solve the controllable zero-shot IC task in a polishing way. By combining Gibbs sampling with a MLM, we can randomly initialize the caption and then polish every word based on the full context (bidirectional information) in the caption.
- ConZIC is free of parameter updates, achieving about 5× faster generation speed than the SOTA method, ZeroCap.
- Equipped with Gibbs-BERT, ConZIC can perform flexible searching, thus generating sentences with higher diversity, as shown in Table. 1.
- To the best of our knowledge, ConZIC is the first controllable zero-shot IC method. Four classes of controllable signals, including length, infilling, styles, and parts-of-speech, are evaluated in our experiments.

## 2. Related work

### 2.1. Supervised Image captioning

To generate text description given an input image, traditional image captioning (IC) often relies on curated image-caption pairs to train an encoder-decoder model. For example, some early attempts [21, 28, 69, 75] construct CNN-based encoder to extract visual features and RNN/LSTM-based decoder to generate output sentence. For better visual understanding, some methods [4, 17, 35, 36, 40, 56, 71] employ an object detector to extract attentive image regions. To encourage more interactions between two modalities, attention mechanism [17, 35, 50, 51, 59, 60] and graph neural network [76, 77] have been widely adopted.

Recently, a series of large-scale visual-language pre-training models [33, 37, 41, 57, 79, 82] have been built, showing remarkable performance in various downstream tasks, IC included. Equipped with these pre-trained models, [22, 33, 49, 64, 80] have shown SOTA performances. However, these methods still need supervised fine-tuning on human-annotated datasets, such as MS-COCO.

### 2.2. Zero-shot Image Captioning

Zero-shot capability with large-scale pretrained models has drawn much attention in a wide range of research fields in computer vision and natural language processing [8, 10, 55, 57, 58, 78], showing great potential of transferring knowledge to tasks without supervised training data. However, for IC, the zero-shot ability is under-explored. To the best of our knowledge, the only achievement was made by Tewel *et al*. in [65], where a method called ZeroCap is proposed. Additionally, several approaches address the task of describing novel objects which are not present in paired image-sentence training data by incorporating external unsupervised data [1, 31, 67] or object tagger [2, 23, 34, 42, 46], namely novel object image captioning (NOIC). In some cases, NOIC is also called zero-shot IC, but is very different from what we study in this work. A detailed compari-

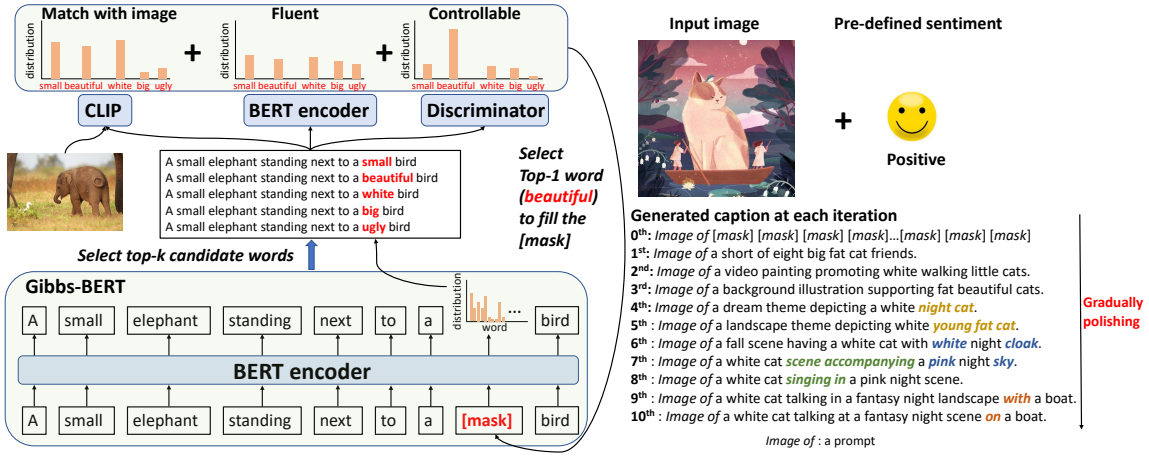



Figure 2. An overview of our approach. ConZIC starts from a prompt "Image of" and a [*mask*] sequence and then iteratively updates the caption by sampling each word (see right). As an example (see left), ConZIC selects the word "beautiful" by considering image-matching score (in Sec. 3.3), fluent score (in Sec. 3.2), and controllable score (in Sec. 3.4) whose specific algorithm is in Algorithm 1. ConZIC can correct itself for zero-shot generation, as illustrated with the same color between two iterations at right.

son is in Appendix A. Specifically, ZeroCap uses a freezed GPT-2 and then for the generation at each position, they update the context cache by minimizing the image-text matching loss measured by CLIP at inference stage. Note that CLIP is a visual-language pretraining model trained on an automatically collected noisy web-scale dataset, rather than human-annotated IC datasets. As a result, ZeroCap realizes zero-shot IC by gradient-directed searching without training. However, due to the autoregressive nature, searching for the word at current position only considers the information from the left side, not the full context. Besides, the autoregressive nature tends to bring mode collapse problem [74], resulting in captions with less diversity. Moreover, the time-cost of iterative gradient-update is high, especially for long captions. Further, ZeroCap has not been considered for the task of controllable zero-shot IC.

### 2.3. Diversity and Controllability

Diversity [68, 72, 73] and controllability [14–16, 18, 19, 24, 25, 30, 39] are two important properties that have drawn much attention in previous IC researches. Recent findings [13] show that the captions generated by supervised methods tend to be biased toward the "average" caption, capturing the most general linguistic patterns and words in the training corpus, *i.e.*, the so-called mode collapse problem. In other words, semantically, "diversity" refers to various words, and syntactically, "diversity" refers to abundant sentence patterns. Without the limitation of supervised data and using the knowledge from CLIP, ZeroCap increases the vocabulary size of generated captions. Nevertheless, its autoregressive nature brings a phenomenon that the candidate sentences for a given image often have similar syntactic patterns, *i.e.*, less syntactical diversity.

To imitate human controllability, lots of works have been made to control the caption generation by introducing control signals into the supervised training process, such as the subjective signals like sentiments [25, 30, 81], emotions [24, 48], personality [15, 62] and the objective signals like length level [18], parts-of-speech [19], object region [16, 44], and visual relation [39]. However, how to introduce control signals without training and realize controllable zero-shot IC has not been explored yet.

## 3. Method

### 3.1. Framework of ConZIC

Given an image $I$, zero-shot image captioning (IC) aims to generate a linguistic description $\boldsymbol{x}_{<1,n>}$ containing $n$ words, without training on the supervised database. This process can be formalized as searching $\boldsymbol{x}_{<1,n>}$ by maximizing the data likelihood $p(\boldsymbol{x}_{<1,n>}|I)$.

To further consider the influence of a control signal $C$, controllable zero-shot IC focuses on searching $\boldsymbol{x}_{<1,n>}$ by maximizing $p(\boldsymbol{x}_{<1,n>}|I, C)$. According to the Bayes rule, the log data likelihood can be derived as:

$$\begin{aligned} & \log p(\boldsymbol{x}_{<1,n>}|I, C) \\ \propto\ & \log p(\boldsymbol{x}_{<1,n>}, I, C) \\ =\ & \log p(I|\boldsymbol{x}_{<1,n>}) + \log p(C|\boldsymbol{x}_{<1,n>}) + \log p(\boldsymbol{x}_{<1,n>}), \end{aligned} \tag{1}$$

which implies three basic rules to guide the searching process, realized by three modules, respectively. Specifically, *i)* a language model (LM), evaluating $p(\boldsymbol{x}_{<1,n>})$, helps with searching for captions with high-level fluency; *ii)* a matching network, measuring the similarity between the input image and the generated caption, *i.e.*, $p(I|\boldsymbol{x}_{<1,n>})$, helps with searching for captions highly related to the input image; and *ii)* a discriminator, measuring $p(C|\boldsymbol{x}_{<1,n>})$, helps with searching for captions that meet the control signal.

These three modules constitute our proposed controllable zero-shot IC framework, ConZIC, which will be further introduced in the following subsections. ConZIC tries to solve the controllable zero-shot IC problem by iteratively polishing the words at every position.

### 3.2. Sampling-based language model for $p(\boldsymbol{x}_{<1,n>})$

To model $p(\boldsymbol{x}_{<1,n>})$, existing IC methods (including zero-shot and supervised ones) often adopt sequential autoregressive generations, as:

$$p(\boldsymbol{x}_{<1,n>}) = p(x_n|\boldsymbol{x}_{<n}) \cdots p(x_2|x_1)p(x_1). \quad (2)$$

However, such autoregressive generation often results in issues such as sequential error accumulation and lack of diversity [13, 74]. Further, for zero-shot IC, the sequential searching-order is lack of flexible. See related work for more detailed discussions. To move beyond, inspired by our analysis of the relation between Gibbs sampling and the design of masked language models (MLMs), we develop a sampling-based LM for $p(\boldsymbol{x}_{<1,n>})$.

Specifically, Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm, which aims to collect samples from the joint data distribution $p(\boldsymbol{x}_{<1,n>})$ by sampling each variable $x_i$ (word in our case) iteratively from its conditional probability $p(x_i|\boldsymbol{x}_{-i})$, where $\boldsymbol{x}_{-i}$ denotes all other random variables in $p(\boldsymbol{x}_{<1,n>})$ except $x_i$. In practice, Gibbs sampling brings flexible sampling orders like

$$\begin{aligned} &p(x_n|\boldsymbol{x}_{-n}) \rightarrow p(x_{n-1}|\boldsymbol{x}_{-(n-1)}) \rightarrow \cdots \rightarrow p(x_1|\boldsymbol{x}_{-1}) \\ &p(x_1|\boldsymbol{x}_{-1}) \rightarrow p(x_2|\boldsymbol{x}_{-2}) \rightarrow \cdots \rightarrow p(x_n|\boldsymbol{x}_{-n}) \\ &p(x_t|\boldsymbol{x}_{-t}) \rightarrow \cdots \rightarrow p(x_j|\boldsymbol{x}_{-j}). \end{aligned} \quad (3)$$

Such flexible order gives Gibbs sampling the ability to walk out the collapsed modes (a key problem of lacking diversity in IC [13]), resulting in more diverse generations.

From another view to analyze Eq. 3, each item is associated with the learning of MLMs. Specifically, given a sentence, MLMs set several words as the [*MASK*] denoted by $\boldsymbol{x}_{\mathbb{M}}$, and then use other words $\boldsymbol{x}_{-\mathbb{M}}$ to predict these masked words. Mathematically, the target of MLMs is to learn the conditional distribution $p(\boldsymbol{x}_{\mathbb{M}}|\boldsymbol{x}_{-\mathbb{M}})$ from the corpus. Therefore, if we just set $i$-th word $x_i$ as the [*MASK*], MLMs and Gibbs sampling are equivalent to predict $p(x_i|\boldsymbol{x}_{-i})$. Currently, we use BERT as MLMs and therefore we call this new LM as Gibbs-BERT to model $p(\boldsymbol{x}_{<1,n>})$.

The specific algorithm of Gibbs-BERT for sampling a sentence $\boldsymbol{x}_{<1,n>}$ from $p(\boldsymbol{x}_{<1,n>})$ is shown in Algorithm 2 in Appendix B. After randomly choosing the generation order, Gibbs-BERT starts from a full noisy sentence (*e.g.*, all [*MASK*] tokens). At each iteration, Gibbs-BERT progressively samples each word by putting [*MASK*] at this position and then selecting the top-1 word from the predicted word distribution over the vocabulary by BERT. The result of $t$-th iteration is the initialization of the $(t+1)$-th iteration.

**Algorithm 1:** Algorithm of our proposed ConZIC.

**Data:** initial caption:$\boldsymbol{x}^0_{<1,n>} = (x^0_1, ..., x^0_n)$;
iterations=$T$, candidates=$K$;
position sequence $P$ = *Shuffle*$([1, ..., n])$;
**Result:** the final caption: $\boldsymbol{x}^T_{<1,n>} = (x^T_1, ..., x^T_n)$;
**for** *iteration* $t \in [1, ..., T]$ **do**
  state: $\boldsymbol{x}^{t-1}_{<1,n>} = (x^{t-1}_1, ..., x^{t-1}_n)$;
  **for** *position* $i \in P$ **do**
    1. Replace $x^{t-1}_i$ with [*MASK*];
    2. Predict the word distribution over vocabulary by Gibbs-BERT: $p(x_i|\boldsymbol{x}^{t-1}_{-i})$;
    3. Select top-$K$ candidate words $\{x^t_{ik}\}_{k=1}$ by $p(x_i|\boldsymbol{x}^{t-1}_{-i})$, whose probability is $p^{Bert}_k$;
    4. Get $K$ candidate sentences $\{s_k\}^K_{k=1}$: $(x^{t-1}_1, ..., , x^{t-1}_{i-1}, x^t_{ik}, x^{t-1}_{i+1}, ..., x^{t-1}_n)^K_{k=1}$;
    5. Compute the CLIP and classifier score for $\{s_k\}^K_{k=1}$ by Eq. 4 and 5: $p^{Clip}_k$ and $p^{Cls}_k$.
    6. Select $x^t_i$ with largest probability by $\alpha p^{Bert}_k + \beta p^{Clip}_k + \gamma p^{Cls}_k$;
    7. Replace $x^{t-1}_i$ with $x^t_i$;
  **end**
  state: $\boldsymbol{x}^t_{<1,n>} = (x^t_1, ..., x^t_n)$;
**end**

### 3.3. Image-text matching network for $p(I|\boldsymbol{x}_{<1,n>})$

To make the generated caption highly related to the image, our framework needs a matching network that can measure the similarity between images and texts. Recently, pre-trained on the sufficiently large-scale image-text pairs, CLIP [57] learns abundant world knowledge for measuring their similarity. Therefore, we introduce the pre-trained CLIP into our framework for modeling $p(I|\boldsymbol{x}_{<1,n>})$.

Specifically, when sampling each word as $p(x_i|\boldsymbol{x}_{-i}; I)$, Gibbs-BERT firstly provides top-$K$ candidate words according to its predicted word distribution over the vocabulary. Then we replace the [*MASK*] token for $i$-th position with these $K$ candidate words, forming $K$ candidate sentences $\{s_k = (x_1, ..., x_{ik}, ..., x_n)\}^K_{k=1}; x_{ik}$ =[*MASK*]. The CLIP matching score $p(I|s_k)$ can be computed as $CLIP(s_k, I)$, where a higher score represents that image and text are better aligned. Using the Softmax, we obtain a predicted distribution over these $K$ candidate words as

$$p(I|\{s_k\}^K_{k=1}) \propto Softmax[CLIP(s_k, I)]. \quad (4)$$

According to Eq. 4, we select the top-1 word (largest probability) as $x_i$, forming the sentence with other words $\boldsymbol{x}_{-i}$.

Up to now, our framework has already realized the zero-shot IC without the control signal. Next, we will introduce how to integrate a discriminator $p(C|\boldsymbol{x}_{<1,n>})$ of the control signal $C$ for controllable zero-shot IC.

### 3.4. Discriminator for control signal $p(C|\boldsymbol{x}_{<1,n>})$

As for controllable IC, we need to generate text related to both image and given control signal $C$. For some types of control signals, like sentiment or parts-of-speech (POS), we need an extra discriminator $p(C|\boldsymbol{x}_{<1,n>})$ to evaluate the correlation between caption and control signals. Specifically, similar to what we do for $p(I|\boldsymbol{x}_{<1,n>})$, after selecting top-$K$ sentences $s_k$ by Gibbs-BERT, we use a pre-trained classifier with Softmax function to model $p(C|\{s_k\}_{k=1}^K)$ as

$$p(C|\{s_k\}_{k=1}^K) \propto Softmax[Classifier(s_k)]. \quad (5)$$

The classifier is different for different tasks, which will be detailed in the experiments.

### 3.5. Overall algorithm

The overall algorithm of the framework for controllable zero-shot IC is shown in Algorithm 1. We firstly need to initial the caption (currently we use all $[MASK]$ tokens) and set several hyper-parameters. Starting from the output of the previous iteration, for each position $i$ of this iteration, Gibbs-BERT firstly provides top-$K$ candidate words, forming $K$ sentences with other words denoted as $\{s_k\}_{k=1}^K$. Then, according to Eq. 4 and Eq. 5, we can obtain the text-image and text-control matching scores $p(I|\boldsymbol{x}_{<1,n>})$ and $p(C|\boldsymbol{x}_{<1,n>})$. After integrating these two distributions with Gibbs-BERT predicted distributions $p(x_i|\boldsymbol{x}_{-i})$ by trade-off parameters $\{\alpha, \beta, \gamma\}$, we can get a final distribution, from which we select the word with the largest probability as $x_i$.

There are three points about our proposed framework that we need to clarify. *i)* deleting the item $p(C|\boldsymbol{x}_{<1,n>})$, our framework can do standard zero-shot IC (without control signal). *ii)* for some tasks of controllable IC, such as length control, there is no need to use $p(C|\boldsymbol{x}_{<1,n>})$, whose details are in experiments. *iii)* our framework is free of the specific modules. As our future exploration, we will study whether better pre-trained models, like using RoBERTa [45] to replace BERT and ALIGN [37] to replace CLIP, can further improve the performance of ConZIC.

## 4. Experiments

### 4.1. Datasets

**MSCOCO caption** [43]: MSCOCO caption is a large IC dataset. we use the 'Karpathy' splits [38] that have been used extensively for reporting results in most of prior works. This split contains 113,287 training and validation images, and 5,000 for testing. Each image contains five captions.

**SentiCap** [47]: SentiCap is a sentiment IC dataset developed from the MSCOCO dataset. Each image is labeled by 3 positive and/or 3 negative sentiment captions. As a result, the positive (negative) set contains 998 (673) and 997 (503) images for training (testing), respectively.

**FlickrStyle10k** [24]: FlickrStyle10k contains 10,000 Flickr images with stylized captions, where only 7,000 images are public. Each image contains 5, 1, and 1 captions for factual (no specific style), humorous, and romantic styles, respectively. Following [30], we randomly select 6,000 and 1,000 of them to construct the training and testing sets.

**SketchyCOCO caption** [26]: SketchyCOCO is constructed by collecting instance freehand sketches covering 3 background and 14 foreground classes. However, SketchyCOCO is not for IC since it only has the classification label. To help us evaluate the performance of IC on sketch-style images quantitatively, we construct a small benchmark based on SketchyCOCO through a text prompt "A drawing of a [CLASS]", where [CLASS] is the class name. More details can be seen in Appendix C.

### 4.2. Implementation Details

As described in Sec. 3, all the experiments are executed based on frozen pre-trained models without any fine-tuning. Specifically, we choose CLIP-ViT-B/32 as our image-text matching network and BERT-Base as our LM. As for different controllable IC tasks, we select corresponding discriminators whose details are introduced in Sec. 4.4.

In Sec. 4.3, we first evaluate the performance on standard uncontrolled IC. Then in Sec. 4.4, we explore the controllability of our method on 4 controllable IC tasks including length, infilling, style, and parts-of-speech (POS). Finally, in Sec. 4.5, we study the speed of generation. $K, T, \alpha, \beta$ are set as 200, 15, 0.02, and 2 among all experiments. For MSCOCO captions and SketchyCOCO captions, we set sentence lengths $n$ as 12 and 5, respectively. As for stylized IC and POS controlled IC, we set $\gamma$ as 5. All experiments are conducted on a single RTX3090 GPU.

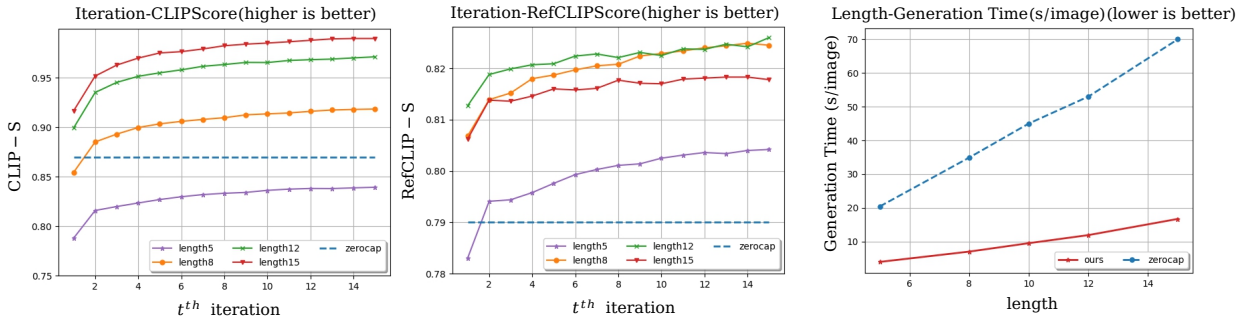
### 4.3. Evaluation on Accuracy and Diversity

We first evaluate the accuracy and diversity of our framework based on standard IC task (without control).

**Evaluation Metrics.** Prior methods usually evaluate their IC performance on accuracy-based metrics. Following [65], we use supervised metrics, *i.e.* metrics requiring human references, including BLEU-4 (B-4) [52], METEOR (M) [7], CIDEr (C) [66], SPICE (S) [3] and RefCLIPScore (RefCLIP-S) [32]. RefCLIPScore measures the semantic similarity between references and predictions. Besides, we also use an unsupervised metric, CLIPScore(CLIP-S) [32,65]. CLIPScore is a reference-free metric measuring the similarity between an image and the corresponding caption, which is the most critical metric for zero-shot IC [65]. Another important performance of IC is to evaluate the diversity of generated captions. We follow [12, 65] and use three metrics, Vocab [65], Self-CIDEr(S-C) [73] and Div-n [5]. Vocab is the vocabulary size of all generated captions on the testing set, reflecting the word richness of different

| Metrics | Accuracy | | | | | | Diversity | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Supervised | | | | | Unsupervised | | | | |
| | B-4(↑) | M(↑) | C(↑) | S(↑) | RefCLIP-S(↑) | CLIP-S(↑) | Vocab (↑) | S-C(↑) | Div-1(↑) | Div-2(↑) |
| Supervised Methods | | | | | | | | | | |
| ClipCap [49] | 32.15 | 27.1 | 108.35 | 20.12 | 0.81 | 0.77 | 1650 | - | - | - |
| MAGIC [64] | 12.90 | 17.22 | 48.33 | 10.92 | 0.77 | 0.74 | 1765 | - | - | - |
| CLIP-VL [61] | 40.2 | 29.7 | 134.2 | 23.8 | 0.82 | 0.77 | 2464 | - | - | - |
| ViTCAP [22] | 41.2 | 30.1 | 138.1 | 24.1 | 0.80 | 0.73 | 1173 | - | - | - |
| GRIT [50] | 42.4 | 30.6 | 144.2 | 24.3 | 0.82 | 0.77 | 1049 | - | - | - |
| VinVL [80] | 41.0 | 31.1 | 140.9 | 25.2 | **0.83** | 0.78 | 1125 | - | - | - |
| LEMON [33] | **42.6** | **31.4** | **145.5** | **25.5** | - | - | - | - | - | - |
| Supervised and Diversity-based Methods | | | | | | | | | | |
| Div-BS [72] | 32.5 | 25.5 | 103.4 | 18.7 | - | - | - | - | 0.20 | 0.25 |
| AG-CVAE [68] | 31.1 | 24.5 | 100.1 | 17.9 | - | - | - | - | 0.23 | 0.32 |
| POS [19] | 31.6 | 25.5 | 104.5 | 18.8 | - | - | - | - | 0.24 | 0.35 |
| ASG2Caption [14] | 31.6 | 25.5 | 104.5 | 18.8 | - | - | - | 0.76 | 0.43 | 0.56 |
| Zero Shot Methods | | | | | | | | | | |
| ZeroCap [65] | 2.60 | 11.50 | 14.60 | 5.50 | 0.79 | 0.87 | 8681 | 0.63 | 0.31 | 0.45 |
| **Ours (sequential)** | 1.31 | 11.54 | 12.84 | 5.17 | **0.83** | **1.01** | 9566 | 0.63 | 0.40 | 0.56 |
| **Ours (shuffle)** | 1.29 | 11.23 | 13.26 | 5.01 | **0.83** | 0.99 | **15462** | **0.95** | **0.62** | **0.87** |

Table 1. Performance compared with SOTA methods on MSCOCO dataset. The baselines can be divided into three parts, supervised, supervised diversity-based methods and zero-shot methods. Ours (sequential) is our framework with sequential generated order, while ours (shuffle) is randomly shuffled generated order. For accuracy metrics, we report the CLIP re-ranked best-1 performance among all iterations. To compute the diversity metrics which need multiple captions, Ours (sequential) is computed on 5 captions in last 5 iteration steps while Ours (shuffle) first randomly sample 5 generation orders, and then select the last-one caption.




(a) Comparision on CLIPScore. (b) Comparision on RefCLIPScore. (c) Comparision on Time-consuming.

Figure 3. (a) and (b): The change of CLIPScore and RefCLIPScore with the iteration steps under different lengths of captions. (c): The generation speed with different lengths. Our framework achieves better accuracy and faster generation speed compared with Zerocap.

| Methods | | B-1(↑) | M(↑) | C(↑) | CLIP-S(↑) |
|---|---|---|---|---|---|
| Supervised | MAGIC [64] | 21.88 | 11.77 | 13.00 | 0.66 |
| | ViTCAP [22] | 27.69 | 17.58 | 22.29 | 0.63 |
| | GRIT [50] | 17.84 | **26.62** | 17.84 | 0.68 |
| Zero Shot | ZeroCap [65] | 27.08 | 20.67 | 21.11 | 0.86 |
| | **Ours** | **39.61** | 20.71 | **34.43** | **0.88** |

Table 2. Performance on SketchyCOCO caption dataset.

methods. Self-CIDER and Div-n are popular diversity metrics based on pairwise similarities between captions.

**Quantitative Results** The quantitative results are reported in Table 1 based on the MSCOCO dataset. Obviously, for supervised metrics B-4, M, C, and S, zero-shot methods without any fine-tuning including ZeroCap and ours, lag behind other supervised methods. This makes sense because supervised methods trained on MSCOCO can benefit domain bias, which means training and testing sets are labeled by the same annotators and thus have similar caption style. However, as we discuss before, IC should not have the standard answers. On the other hand, training or fine-tuning on one dataset does help models produce captions similar to the training data, but limits the word vocabulary size and thus decreases the diversity of generated captions. Specifically, shown in the *Diversity* column, even compared with existing supervised methods, Div-BS [72], AG-CVAE [68], POS [19], and ASG2Caption [14] that focus on improving the diversity, our method surpasses them with a large margin. Furthermore, our framework gets comparable and superior performance compared with SOTA methods on semantic-related metrics, RefCLIPScore and CLIPScore, respectively, indicating that our method can generate high image-matching caption.

Moreover, Table. 2 reports the zero-shot performance on sketch-style images (SketchyCOCO caption dataset) compared with previous SOTA methods. Our framework outperforms supervised methods (trained on MSCOCO) on most of the metrics because of the domain gap between

| sentence length control | | | | | |
|---|---|---|---|---|---|
| 3~5 | A stuffed black bear. | A fruit dish. | A calm businessman concentrating hard. | A blond farm cow. | A central animal courtyard. |
| 7~9 | A bear toy named Cooper admiring himself. | A fruit dish in tin color offering sweet orange. | A financial administrator watching financial statements online. | A farm buffalo around metal enclosure and foliage. | A cornered cat shown against numerous pigeons. |
| 11~13 | A stuffed teddy dark bear smiling with yoga pose in a mirror. | A photo showing Osaka orange fruits appearing in a stainless steel pot. | A man distracted thinking business report with neatly trimmed white hair. | A village animal cow shows in tree ferns background and fences. | A mute cat meeting numerous birds and pigeons in a Greek square. |

Figure 4. Examples of **length controlling** by ConZIC. Given one image, we show the generated captions controlled by three different predefined lengths. Empirically, short captions are more global and generally describe the most salient object in images, while long captions talk about more visual details.

| Method | Noun | | | Verb | | |
|---|---|---|---|---|---|---|
| | B-1(↑) | WSim(↑) | BSim(↑) | B-1(↑) | WSim(↑) | BSim(↑) |
| ZeroCap [65] | 0.00 | 0.001 | 0.11 | 0.00 | 0.001 | 0.10 |
| **ConZIC** | **0.37** | **0.39** | **0.52** | **0.25** | **0.46** | **0.50** |

Table 3. Results of one word infilling task on MSCOCO dataset.

MSCOCO and SketchyCOCO. Meanwhile, we surpass another zero-shot method ZeroCap on all metrics.

**Qualitative Results.** As shown in Figs. 1 and 2, our framework can produce accurate and diverse captions with more words and abundant sentence patterns. More examples can be seen in Appendix D.

### 4.4. Evaluation on controllable IC tasks

We have considered 4 controllable tasks. The first two, *i.e.*, length and infilling, are classifier-free. The last two, *i.e.*, style and parts-of-speech, rely on an off-the-shelf classifier. We will detail each control task as follows.

**Length.** For our framework ConZIC doing length-control IC, we just need to set the initial length of the caption. We do this experiment on MSCOCO. Considering that the average sentence length of annotated captions in MSCOCO is about 10, we have tried 4 different lengths: 5, 8, 12, and 15. The qualitative results are illustrated in Fig. 4. Empirically, given a target length, ConZIC can generate accurate descriptions with a length within $\pm 2$ due to the WordPiece-based tokenizer of BERT [70]. To understand the effect of iteration steps and caption length in ConZIC, Figs. 3a and 3b show the change of CLIPScore and RefCLIPScore with the increase of iteration steps, with ZeroCap as a baseline. These two scores increase with iterative updates, demonstrating the effectiveness of polishing mechanism. Compared results from Fig. 4 and Fig. 3a, we observe that longer sentence generally contains more details of images, which facilitate a higher CLIPScore. For RefCLIPScore in Fig. 3b, we find that Length 12 and Length 8 have similar better results than Length 5 and Length 15. We attribute it to the fact that the average length of caption in MSCOCO is about 10, and RefCLIPScore evaluates the similarity between generated and reference captions.

| Metrics | | B-3(↑) | M(↑) | CLIP-S(↑) | Acc(↑) |
|---|---|---|---|---|---|
| Positive | | | | | |
| Supervised | StyleNet [24] | 12.1 | 12.1 | - | 45.2 |
| | MSCap [30] | 16.2 | 16.8 | - | 92.5 |
| | MemCap [81] | **17.0** | **16.6** | - | 96.1 |
| Zero Shot | **ConZIC** | 1.89 | 5.39 | **0.99** | **97.2** |
| Negative | | | | | |
| Supervised | StyleNet [24] | 10.6 | 10.9 | - | 56.6 |
| | MSCap [30] | 15.4 | 16.2 | - | 93.4 |
| | MemCap [81] | **18.1** | **15.7** | - | 98.9 |
| Zero Shot | **ConZIC** | 1.78 | 5.54 | **0.97** | **99.1** |

Table 4. Sentiment controlled image captioning (*i.e.* positive, negative) performance comparisons on the SentiCap dataset. Acc is style classification accuracy.

**Infilling.** Given human-annotated caption with parts of words absent, infilling task targets to infill suitable words conditioning on the image content. we consider this task as a special controllable IC task since we need to generate text conditioning not only on the image content but also on the fixed left and right context. Most existing IC models, such as ZeroCap, can not do this task since the autoregressive LM they used can generate words only based on the left context. On the contrary, ConZIC does this task without using a classifier since its LM, *i.e.* Gibbs-BERT, is modeled on bidirectional attention. Firstly, we conduct quantitative experiments based on the setting where only one word is absent. Specifically, we randomly mask one verb or noun in MSCOCO reference caption, and then require ConZIC to infill it. Regarding the original word as the ground truth, we choose BLEU-1 (B-1) as the metric to evaluate the accuracy. However, many other words are also suitable for this position (diversity). Therefore, we use two metrics to measure the semantic similarity between predicted and reference words: WordNet word similarity (WSim) [53], and BERT embedding cosine similarity (BSim) [70]. Results are shown in Table 3, where ConZIC outperforms ZeroCap by a large margin. We provide more quantitative and qualitative infilling results by ConZIC in Appendix E.

**Style.** Given an image with a specific linguistic style, *e.g.*, positive, negative, romantic, or humorous, style-

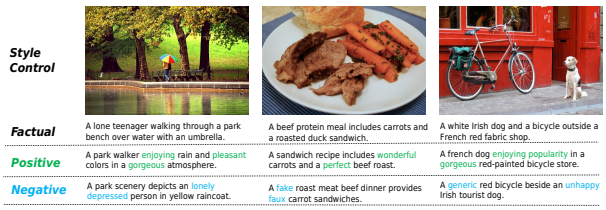



Figure 5. Examples of **sentiment controlled task** by ConZIC. Factual is the captions without controlling generated by our proposed framework. Positive and Negative are the stylized image captioning results of ConZIC, where sentiment-related words are highlighted in green and blue, respectively.

controlled task aims to generate corresponding descriptions. For this task, ConZIC needs a pre-trained classifier to discriminate the style of captions. Currently, we use SentiwordNet [6] for sentiment (positive or negative) controlling and use TextCNN [29] for romantic-humorous controlling. Firstly, we evaluate the sentiment-controlled capability of ConZIC on the SentiCap dataset. The quantitative results are shown in Table 4. As baselines, StyleNet [24], MSCap [30], and Memcap [81] achieve the SOTA performance on this task in *supervised* way, resulting in a higher B-3 and M. Following them, we test the accuracy (Acc) of ConZIC, which is obtained by feeding the generated captions into a sentiment classifier, where ConZIC is higher. Furthermore, we use CLIP-S to evaluate the correlation between image and caption whose results demonstrate that the captions generated by ConZIC are highly correlated to the images. Since three baselines do not provide public codes, we cannot evaluate them on CLIP-S but just report other three metrics from the paper. For better visualization of ConZIC for sentiment controlling, Figs. 1b, 2, and 5 show multiple results. Results and analysis on the FlickStyle10k about romantic-humorous styles are in Appendix E.

**Parts-of-speech (POS).** The task of POS-control IC is to generate captions which match given POS tags. For example, the POS of caption *a cat sitting in the bed* is *DET NOUN VERB ADP DET NOUN*. For this task, we use the POS classifier developed in [9]. Considering the fact that human generally describes an image under some common templates, like "somebody/something doing something at someplace", we design a POS tag sequence as *DET ADJ/NOUN NOUN VERB VERB ADV ADP DET ADJ/NOUN NOUN NOUN*. Under such a template, we report the results of ConZIC in Table 5. Clearly, Our method can achieve a high accuracy under this POS tag template, but M, C, CLIP-S get slightly lower because not all images are suitable for this POS tag. We also visualize some examples in Fig. 6. More analysis and results are discussed in Appendix E.

### 4.5. Evaluation on generation speed

The generation speed is significant for zero-shot IC. For ZeroCap and ConZIC, empirically, we find that the speed is mostly related to sentence length. Fig. 3c shows the generation speed of ConZIC and ZeroCap with different sentence lengths, evaluated on MSCOCO. Our method is around 5 times faster than ZeroCap with 15 iterations, which attributes to our proposed sampling-based Gibbs-BERT.

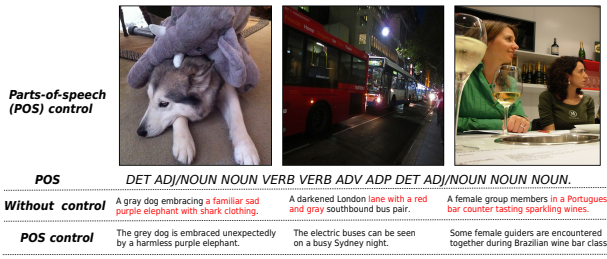



Figure 6. Examples of **parts-of-speech controlling** (POS) by ConZIC. *DET ... NOUN* is the predefined POS template. "Without POS" denotes uncontrolled results of proposed framework where words in color represent they do not satisfy the POS. "POS control" denotes the controllable results of ConZIC.

| Parts-of-speech | M(↑) | C(↑) | CLIP-S(↑) | Acc(↑) |
|---|---|---|---|---|
| without POS | 11.54 | 12.84 | 1.01 | 15.54 |
| with POS | 8.25 | 10.89 | 0.96 | 83.36 |

Table 5. Results of parts-of-speech control on MSCOCO.

## 5. Conclusion and future work

In this paper, we propose a flexible and efficient framework for controllable zero-shot IC, named ConZIC. Firstly, by discovering the relation between MLMs and Gibbs sampling, we develop a new sampling-based language model called Gibbs-BERT. Compared with widely used autoregressive models, Gibbs-BERT has flexible generation order, bringing the self-correct capability by bidirectional attention. To integrate Gibbs-BERT with the CLIP for image-text matching and the pre-trained discriminator for controlling, ConZIC can realize better zero-shot IC with/without control signals. However, as our analysis of failure examples on ConZIC and ZeroCap in Appendix F, the study of zero-shot IC is still in its infancy, leaving a large space to go further. For example, ConZIC and ZeroCap often ignore small targets in the image, which may be alleviated by a detector for capturing small targets in the image. Besides, developing more appropriate metrics, especially for controllable IC, rather than using those supervised metrics, is also important for the development of zero-shot IC.

## 6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2006; in part by Shaanxi Youth Innovation Team Project; in part by the 111 Project under Grant B18039; in part by the Fundamental Research Funds for the Central Universities QTZX22160.

## References


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

[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. 2

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 1, 5

[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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[5] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4261–4270, 2019. 5

[6] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010. 8

[7] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 1, 5

[8] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 2

[9] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006. 8

[10] 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. 2

[11] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992. 2

[12] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16856, 2021. 2, 5

[13] Qi Chen, Chaorui Deng, and Qi Wu. Learning distinct and representative modes for image captioning. *arXiv preprint arXiv:2209.08231*, 2022. 3, 4

[14] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9962–9971, 2020. 2, 3, 6

[15] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903, 2017. 3

[16] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019. 3

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

[18] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. In *European Conference on Computer Vision*, pages 712–729. Springer, 2020. 3

[19] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10695–10704, 2019. 2, 3, 6

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

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

[22] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18009–18019, 2022. 1, 2, 6

[23] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang. Cascaded revision network for novel object captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3413–3421, 2020. 2

[24] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017. 2, 3, 5, 7, 8

[25] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017. 3

[26] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5174–5183, 2020. 5

[27] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 2

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

[29] Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374, 2019. 8

[30] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4204–4213, 2019. 3, 5, 7, 8

[31] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016. 2

[32] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[33] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 1, 2, 6

[34] Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. In *proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1575–1583, 2021. 2

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

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

[37] 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. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2, 5

[38] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 5

[39] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2019. 3

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

[41] 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. 2

[42] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12497–12506, 2019. 2

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

[44] Annika Lindh, Robert J Ross, and John D Kelleher. Language-driven region pointer advancement for controllable image captioning. *arXiv preprint arXiv:2011.14901*, 2020. 3

[45] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[46] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 2

[47] Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 2, 5

[48] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600, 2018. 3

[49] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 1, 2, 6

[50] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022. 1, 2, 6

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

[52] 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. 1, 5

[53] Ted Pedersen, Siddharth Patwardhan, Jason Michelizzi, et al. Wordnet:: Similarity-measuring the relatedness of concepts. In *AAAI*, volume 4, pages 25–29, 2004. 7

[54] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 2

[55] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

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

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4

[58] 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. 2

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

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

[61] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 6

[62] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526, 2019. 3

[63] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022. 2

[64] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 2, 6

[65] 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, 5, 6, 7

[66] 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. 1, 5

[67] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2017. 2

[68] Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3, 6

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

[70] Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019. 2, 7

[71] Li Wang, Zechen Bai, Yonghua Zhang, and Hongtao Lu. Show, recall, and tell: Image captioning with recall mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12176–12183, 2020. 2

[72] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *Advances in Neural Information Processing Systems*, 30, 2017. 3, 6

[73] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4195–4203, 2019. 2, 3, 5

[74] Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. A survey on non-autoregressive generation for neural machine translation and beyond. *arXiv preprint arXiv:2204.09269*, 2022. 3, 4

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

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

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

[78] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

[79] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3008, 2021. 2

[80] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2, 6

[81] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12984–12992, 2020. 3, 7, 8

[82] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 2