Exploring Discrete Diffusion Models for Image Captioning

Zixin Zhu[1*] Yixuan Wei[2*] Jianfeng Wang[3] Zhe Gan[3] Zheng Zhang[3] Le Wang[1,†] Gang Hua[1]
Lijuan Wang[3] Zicheng Liu[3] Han Hu[3,†]
[1]Xi'an Jiaotong University [2]Tsinghua University
[3]Microsoft
zhuzixin@stu.xjtu.edu.cn
lewang@xjtu.edu.cn
{t-yixuanwei,jianfw,zhe.gan,zhez,lijuanw,zliu,hanhu}@microsoft.com

Abstract

The image captioning task is typically realized by an auto-regressive method that decodes the text tokens one by one. We present a diffusion-based captioning model, dubbed the name DDCap, to allow more decoding flexibility. Unlike image generation, where the output is continuous and redundant with a fixed length, texts in image captions are categorical and short with varied lengths. Therefore, naively applying the discrete diffusion model to text decoding does not work well, as shown in our experiments. To address the performance gap, we propose several key techniques including best-first inference, concentrated attention mask, text length prediction, and image-free training. On COCO without additional caption pre-training, it achieves a CIDEr score of 117.8, which is +5.0 higher than the auto-regressive baseline with the same architecture in the controlled setting. It also performs +26.8 higher CIDEr score than the auto-regressive baseline (230.3 v.s. 203.5) on a caption infilling task. With 4M vision-language pre-training images and the base-sized model, we reach a CIDEr score of 125.1 on COCO, which is competitive to the best well-developed auto-regressive frameworks. The code is available at https://github.com/buxiangzhiren/DDCap.

1. Introduction

Diffusion models [25, 61] have been successfully exploited for image generation [8, 48, 53, 56, 57], producing results with high fidelity. However, there is limited work, if at all, exploring diffusion models for text generation such as that in image captioning [31, 66], where natural language description is generated to describe an image. In view of this under-exploration, in this work, we aim to conduct a systematic study on how we may adopt diffusion models for accurate image captioning.

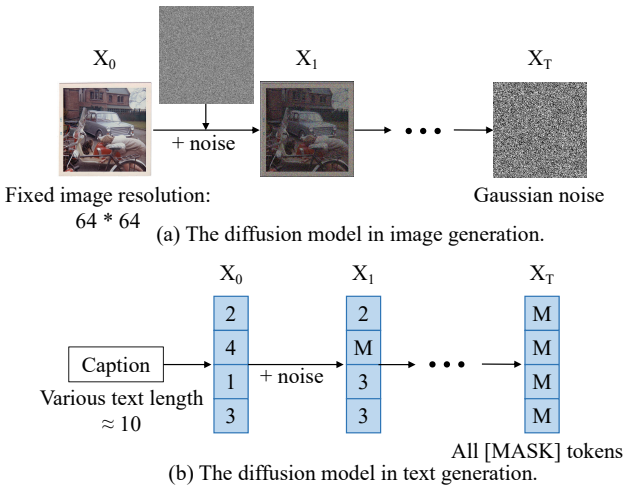


Figure 1. (a) Image generation: the noise is continuous and redundant with a fixed length (b) Text generation: the noise is categorical and short with varied lengths.

Our first attempt of naively applying either continuous or discrete diffusion models for image captioning did not work well, producing results that are much more inferior to state-of-the-art results produced from mainstream auto-regressive models pre-trained on millions of data and even billions of data, *i.e.*, OSCAR [40], VinVL [81], UFO [67], ViTCap [13], SimVLM [71], and GIT [69], to name a few. This motivated us to conduct a careful gap analysis.

As shown in Figure 1, the diffusion model exhibits different natures between the image generation and the text generation. On one hand, in image generation, the output images are *continuous and redundant* with a *fixed length*, and the information residing in the output is *redundant*. On

*Equal. †Corresponding authors. Zixin and Yixuan are interns at MSRA.

the other hand, in image captioning, the output texts are *discrete and concise* with a *variable length*. The diffused image resolution is typically $64 \times 64 = 4096$, while the length of the image caption tokens can be around 10. Besides, in the inference procedure of a diffusion model, some correct words generated at the current step may flip to a wrong words because of furtherly added noises.

Based on the above analysis and in observation that text tokens are discrete in nature, we focused our research exploration on making discrete diffusion model produce accurate image captions. Our proposed model, namely DDCap, is specifically designed to address the above gaps. First of all, we add a network branch to predict the total token length to flexibly accommodate variable lengths of the texts. Secondly, since texts are discrete and concise in nature, we design a *concentrated attention mask* module so that we may adaptively concentrate on more informative tokens.

Thirdly, we propose a first best inference strategy, where the top-K_t recovered tokens from each diffusion step unchanged in subsequent diffusion steps. In other words, in each diffusion step, we only add noises to those tokens that are not marked as fixed in previous diffusion steps. The top-K_t in the current diffusion step are selected only from these remaining tokens which had been added noises, and they will be fixed in subsequent steps. The number K_t is roughly set as the total number of tokens we need to recover divided by the number of diffusion steps.

Last but not least, we further design an image-free training technique, motivated by the classifier-free diffusion model [64]. In some training examples, we remove the image conditions and force the network to learn the text prior knowledge. This enables a trade-off to balance the dataset prior and the image condition. In inference, the text prior knowledge and image condition are appropriately combined to obtain better captions than just depend on images. With these orchestrations, we pre-train the DDCap model with a CLIP image encoder on datasets with 4M images (roughly 10M image-text pairs in total). We are able to achieve a CIDEr score of 125.1 on the COCO dataset, which is competitive when compared with state-of-the-art auto-regressive model. In further exploration of the capacity of our proposed DDCap model, we hypothesize that our DDCap model considers more holistic contexts to generate the texts than the auto-regressive baseline, when generating the text captions.

We further verified our hypothesis by conducting experiments in an infilling tasks, which is illustrated in Figure 6. In this task, we remove all adjectives in the ground truth, and require the model to predict the removed adjectives. As for evaluation metrics, we add the clip scores [23] for semantic matching. Our proposed DDCap model outperforms the auto-regressive baseline by a significant margin.

The contributions of this paper are hence summarized as follows.

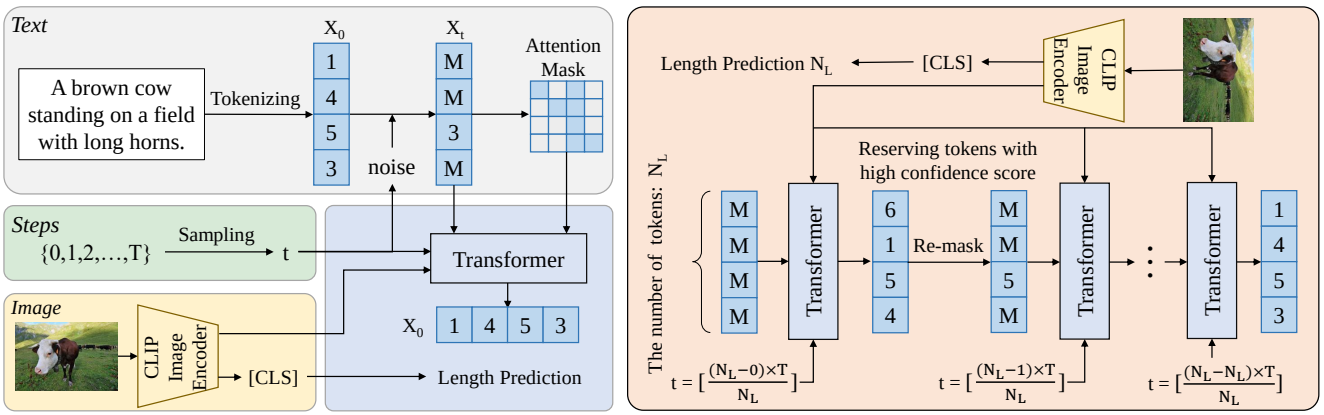
- We are the first to apply discrete diffusion models for image captioning, and provide the first evidence that diffusion models can achieve competitive performance to the best well-developed auto-regressive models.
- We propose four key designs: (i) length prediction is used to deal with the varied length issue in text generation; (ii) concentrated attention mask is used to extract compact text information without the interference of undesired noise; (iii) best-first inference is proposed to reduce the chance of contaminating correctly generated tokens; and $(iiii)$ image-free training to balance the information in the text prior and the image condition.
- We further show the advantages of discrete diffusion models in a caption infilling task.

The remainder of the paper is organized as follows. We summarize related work in Section 2. The details of the proposed DDCap model is proposed in Section 3. Extensive experimental results are reported in Section 4, with both quantitative and qualitative discussions. We finally conclude in Section 5.

2. Related Work

Image Captioning. Image captioning has been a long-standing task and recent years have witness great progress, especially with large-scale vision-language pretraining. Early models [2, 31, 55, 66] use a visual encoder to extract visual features and apply recurrent neural network as decoder for caption generation. Later on, attention-based methods have been proposed to capture multimodal alignment [29, 45, 74] and perform object relational reasoning [76, 77]. Besides, researchers have explored to leverage semantic attributes [18, 73, 78, 79] and scene graphs [75] for captioning. Enhanced attention-based models are further proposed to improve the performance, such as ORT [22], AoANet [29], M^2 Transformer [7], X-LAN [50], and RSTNet [83]. More recent efforts on image captioning are reflected from the perspective of novel model architecture design [13, 15, 47, 50] and the use of prior knowledge [24, 35, 42, 46, 72]. For example, Fang *et al.* [13] present a detector-free image captioning model with fully transformer architecture, and introduce a novel module to predict semantic concepts for high performance. Note that most previous works adopt a standard auto-regressive method to decode tokens one by one.

For non-regressive approaches, masked language modeling is leveraged in [19] and reinforcement learning is applied in [21]. Instead, we focus on exploring diffusion models for image captioning, and achieve better performance.



(a) The process of training.

(b) The process of inference.

Figure 2. (a) A caption is first tokenized. Then, these tokens are added noise to become other text tokens or `[MASK]` tokens. The strength of noise depends on the sampled step t. Lastly, with the help of image tokens, we take these text tokens carried noise as the input of transformer to predict the pure text tokens. Meanwhile, we use a simple MLP to predict the length of the caption based on the `[CLS]` token. (b) According to the predicted length which denotes the number of `[MASK]` tokens, we take all `[MASK]` tokens as the input. Moreover, each step t adjusted by the ratio of the predicted length and total noise step T is sent to the Adaptive LayerNorm layer of the transformer. Image tokens are used to conduct cross-attention with text tokens. Combining three inputs, we can obtain the results of each step. The tokens with high confidence score will be retained instead of re-masking.

Vision-Language Pre-training. Recent advances in image captioning can be largely attributed to vision-language pre-training (VLP). Early VLP models [17, 38, 62, 84] adopt a two-stage approach, where a pre-trained object detector is first used to extract regional features offline, and then a transformer is stacked on top for multimodal fusion. Prominent examples along this line include ViLBERT [44], LXMERT [63], UNITER [6], OSCAR [40], VinVL [81], and LEMON [27]. More recently, end-to-end VLP models that use convolutional networks and vision transformers as the image encoder are becoming the mainstream [5, 10, 11, 33, 36, 37, 70]. For example, SimVLM [71] is pre-trained with a single prefix language modeling objective on large-scale image-text data, GIT [69] simplifies the architecture as one image encoder and one text decoder under a single language modeling task, and CoCa [80] is pre-trained with a mix of contrastive and generative losses. In this work, we also perform VLP for image captioning, but distinct from all the above works, we are the first to pre-train a discrete diffusion model for image captioning (and more generally, for text generation).

Diffusion Models. Diffusion models can be roughly divided into two categories: (i) continuous diffusion, and (ii) discrete diffusion. Continuous diffusion models [8, 25, 61, 82] add noise on the continuous-valued input or latent features. Recently, this family of models have achieved great successes in text-to-image generation, such as GLIDE [48], DALLE-2 [53], Imagen [57], and Stable Diffusion [56]. On the other hand, discrete diffusion [60], especially its application to text generation, has been rarely studied, which is the focus of this work. Specifically, ImageBART [12] and VQ-Diffusion [20] propose to model the latent discrete code space of a VQ-VAE [54] by learning a parametric model using a conditional variant of DDPM [25], for the task of text-to-image generation. For text generation, discrete diffusion has only shown some initial success on relatively toy problems [3, 26]. Recent work [41] has also tried to use continuous diffusion for text generation, but more for fine-grained control tasks. We are the first to apply discrete diffusion models for image captioning, and provide the first evidence that diffusion models can achieve competitive performance to the best well-developed auto-regressive models.

3. Methods

We present DDCap, which is a discrete diffusion model dedicated for image captioning task. This allows more decoding flexibility, compared to the current dominant auto-regressive methods. For example, multiple tokens can be predicted simultaneously rather than one by one. The tokens from the left side can also depend on those on the right side, instead of a fixed left-to-right generation process. In Section 3.1, we first introduce the key idea of applying discrete diffusion models for the captioning task, and then in Section 3.2, we detail multiple techniques that play a crucial role to boost the performance.

3.1. Discrete diffusion model for image captioning

Unlike continuous diffusion models which add token-level continuous noises individually to the RGB values or

features of each token, discrete diffusion models perform a noising process at the sentence level, where mask tokens randomly replace regular tokens with noise levels represented by masking ratio.

Noising process We denote each text token in a caption as a discrete state, x. In the experiments, we use the BPE tokenizer [58] in GPT-2 to generate the token identities of the sentence.

The noising process can be thought of as the Markov process, which gradually adds noises step-by-step. We denote the corresponding noisy token at Step t as x_t. At each step, each token has a probability of γ_t transiting to a special absorbing state of `[MASK]`. If x_{t-1} is not `[MASK]`, the transition probability vector from Step $t-1$ to t is defined as

$$p(x_t|x_{t-1}) = \begin{cases} \alpha_t, & x_t = x_{t-1}; \\ \gamma_t, & x_t = \texttt{[MASK]}; \\ 1-\alpha_t-\gamma_t, & \text{otherwise.} \end{cases} \tag{1}$$

In other words, the each token of $\boldsymbol{x}_{t-1}$ has a probability of α_t to be unchanged and has a probability of γ_t to be replaced by the [MASK] token, leaving the probability of $\beta_t = 1-\alpha_t-\gamma_t$ to be replaced by other tokens except the [MASK] token in the vocabulary. If x_{t-1} is a `[MASK]` token, the transition probability vector from Step $t-1$ to t is defined

$$p(x_t|x_{t-1} = \texttt{[MASK]}) = \begin{cases} 1, & x_t = \texttt{[MASK]}; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

By these noising process, with a sufficient number of steps T, all tokens will become `[MASK]`.

Denoising process The denoising diffusion model starts with an all `[MASK]` sequence and gradually recovers the original signal step-by-step. A Transformer network is used to predict the reserve projection, i.e., $p_\theta(x_{t-1}|x_t, y)$, where y is the image feature extracted from an image encoder, i.e., initialized from CLIP [52] in our experiments. Cross-attention layers are employed to incorporate image features as conditions for the caption generation process. The time step t is encoded as a sinusoidal positional embedding to guide the adaptive layer norm [20] in the Transformer layers:

$$p = t * \text{step}_{\text{scale}}/T, \tag{3}$$

$$\text{PE}_i = \begin{cases} sin(p/10000^{2i/d_{\text{model}}}), i < d_{\text{model}}/2 \\ cos(p/10000^{2i/d_{\text{model}}}), i \geq d_{\text{model}}/2, \end{cases} \tag{4}$$

where $\text{step}_{\text{scale}}$ is the wavelength, i.e., 8000 in our experiments, and d_{model} is the hidden dimension.

Training Unlike continuous diffusion models, where denoising networks typically predict noises, discrete diffusion models perform the noising process at the sentence-level through a masking strategy, and are difficult to directly predict noises. Inspired by [3], we directly predict the original text tokens x_0 instead:

$$\mathcal{L}_{x_0} = -\log p_\theta(x_0|x_t, y). \tag{5}$$

The training process is illustrated as the left part of Figure 2.

Inference Starting from x_T, the model first predicts x_{T-1}, then predicts x_{T-2}, and so on. After T steps, the discrete diffusion model produces the final result of x_0.

The process of predicting x_{t-1} from x_t is computed as follows. First, $\hat{x}_0$ is estimated from $p_\theta(x_0|x_t, y)$ via the trained denoising network. Then, $\{t-1\}$-step noise is added on the predicted $\hat{x}_0$ through a Markov chain to get x_{t-1}.

3.2. Key Designs for Performance Improvement

Empirically, with naive implementations, we observe significant worse performance with the diffusion model than the auto-regressive counterpart. We hypothesize that this is due to that text generation is very different from image generation. For example, the generated text lengths are different, while the generated image resolutions are normally fixed. The information in the text is compact, while the image contains redundant information. The mask token provides almost no information, while the image is continuously noised into different levels. To address the above issues, we propose the following four techniques: (i) length prediction, (ii) concentrated attention mask, (iii) best-first inference and $(iiii)$ image-free training.

Length Prediction. In auto-regressive methods, the text tokens are predicted from left to right one by one until a special `[EOS]` token. However, diffusion model has no concept of left-to-right or right-to-left directions. The performance is also much worse if we respect the `[EOS]` token and treat only the left-side tokens as valid. Instead of relying on the special token, we propose to add a network branch to specifically predict the total token length. Specifically, we regard the `[CLS]` token from the CLIP image encoder as the feature for length prediction, implemented via a simple Multilayer Perceptron (MLP). Cross-entropy loss is used for training. Empirically, we also find it is beneficial to cut the gradient back to the image encoder. Without the length as a prior, we may have to use a maximum token length to generate the captions. Thus, another benefit is the speed-up with an appropriately predicted token length.

Concentrated Attention Mask (CAM). In our diffusion model, the absorbing `[MASK]` token carries almost no information, and it can take a majority in the sequence. For instance, the denoising process starts from a full `[MASK]` sequence. With the bidirectional attention, the `[MASK]` tokens may overwhelm the attention layers. To make the tokens only concentrate on informative tokens, we propose the following changes on the attention mask: (i) normal text tokens do not depend on the `[MASK]` tokens; (ii) the `[MASK]` token does not depend on the other `[MASK]` tokens. Empirically, we find this novel attention mask design significantly boosts the performance.

Best-first Inference. During inference, the text token can be recovered earlier than the last step. To reduce the chance of contaminating these correct tokens, we propose to keep the top-K^t recovered tokens unchanged at each step. Let N_L be the predicted token length. If $N_L \leq T$ (T is the total diffusion steps during training), we also reduce the inference diffusion steps to N_L, such that at each step, one extra token is recovered. Meanwhile, the step index t, which is used in the adaptive layer norm, is also shrunk proportionally. If the predicted token length is larger ($N_L > T$), at each step t (from T to 1), we keep $K^t = \lfloor N_L(T-t+1)/T \rfloor - \lfloor N_L(T-t)/T \rfloor$ tokens unchanged.

Image-free Training. Generating the text is the key of the image captioning task. To focus more on the text modeling part, we propose an image-free strategy to enhance the weight of text modeling, which is motivated from the classifier-free diffusion [64]. Specifically, with a probability of r (0.2 in experiments), the image features are replaced with a trainable embedding f, and the diffusion model calculates $p_\theta(x_0|x_t, f)$ without image conditions. The loss then becomes

$$\mathcal{L}'_{x_0} = -\log p_\theta(x_0|x_t, f). \tag{6}$$

With the loss, the network can focus more on learning how to generate text, rather than how to learn the image features. During inference, the probability likelihood is correspondingly computed as

$$\log p_\theta(x_0|x_t, y)' = \log p_\theta(x_0|x_t, f) + \tag{7}$$
$$s(\log p_\theta(x_0|x_t, y) - \log p_\theta(x_0|x_t, f)), \tag{8}$$

where s denotes the guidance scale ($s = 1.17$ in experiments). If $s = 1$, the image-free process is effectively removed. If $s > 1$, the network will respect more of the image signals.

4. Experiments

4.1. Setup

Dataset. We conduct our experiments on the popular COCO dataset [43]. Specifically, the COCO dataset contains 123,287 images labeled with 5 captions for each, including 82,783 training images, 40,504 validation images and 40,775 images as test set for online evaluation as well. Following the widely-used "Karpathy" split [32], we use 113,287 images for training, 5000 images for validation, and 5000 for testing. For our diffusion model, our vocabulary size is 50,257, and the max length of each sentence is 20. During training, the weight decay is set to 0.01, and the learning rate first linearly warms up to 2e-4, and then drops to 0 following cosine decay. The number of training epochs is 30, and the warm-up epoch is 5. The batch size is 512. Following the standard evaluation setup, we report the performances of our model and compare to other methods over five metrics: BLEU@4 [51], METEOR [4], ROUGE-L [30], CIDEr-D [65], and SPICE [1].

Network backbone and pre-training. ViT-B/16 [9] from the pre-trained CLIP model [52] is chosen as our image backbone, and the corresponding image patch size is 16. More specifically, ViT-B/16 has 12 transformer layers with 768 as the hidden size. For pre-training, following common practice [6, 11], we use a combined dataset including COCO [43], Conceptual Captions (CC) [59], SBU [49], and Visual Genome (VG) [34]. This results in roughly 4 million images with 10 million associated captions. We also pre-train the model based on ViT-L/16 with 15 epochs when comparing with state-of-the-art methods. For our diffusion model, the peak learning rate is 1e-4 with batch size 1024 during pre-training and 1e-5 with batch size 512 during fine-tuning. The diffusion model is randomly initialized and the learning rate of the well-initialized image encoder is reduced to 0.07 of the diffusion model.

4.2. Analysis

Due to the consideration of computation cost, we conduct ablation study based on the fixed image encoder and disable the image-free training unless explicitly specified. No pretraining is conducted, and all results are reported on the "Karpathy" validation set.

Ablation on the key designs. As discussed in Section 3.2, we have four key components in our model. Table 1 reports the ablation study with different combinations. Comparing row (c) and (a), we can see the Best-first Inference strategy can significantly boost the performance from 20.6 to 45.2 in CIDEr. Thus, it is crucial to keep the best predictions unchanged during the denoising process. By appropriately masking the attentions with CAM, the performance can be

#Row	Best-first inference	CAM	Length Prediction	Image-free training	C	B@4	M	R	S
a					20.6	7.4	18.8	34.6	12.3
b		✓			43.6	11.4	20.5	39.1	14.4
c	✓				45.2	20.3	26.9	47.3	21.3
d	✓		✓		92.6	27.3	25.4	51.8	18.7
e	✓	✓			97.5	28.2	28.1	54.0	**21.7**
f	✓	✓	✓		116.7	34.6	28.1	57.4	21.5
g	✓	✓	✓	✓	**117.8**	**35.0**	**28.2**	**57.4**	**21.7**

Table 1. Ablation study on the effectiveness of each component described in Sec. 3.2. CAM: concentrated attention mask.

Method	M2M	T2M	C	B@4	M	R	S
DDCap			92.6	27.3	25.4	51.8	18.7
DDCap		✓	94.1	27.2	25.7	52.1	18.9
DDCap	✓		115.6	34.2	27.9	57.2	21.4
DDCap	✓	✓	**116.7**	**34.6**	**28.1**	**57.4**	**21.5**
DDCap ($t < 10$)	✓	✓	98.8	29.8	25.2	52.8	19.4
DDCap ($t \geq 10$)	✓	✓	112.9	33.2	27.9	56.5	21.2

Table 2. Ablation study on our proposed concentrated attention mask (CAM). "M2M": the mask token does not depend on the other mask tokens. "T2M": the text token does not depend on mask tokens. "($t < 10$)": CAM is enabled at the second half of denoising steps. "($t \geq 10$)": CAM is enabled at the first half of denoising steps where most of the tokens are mask tokens.

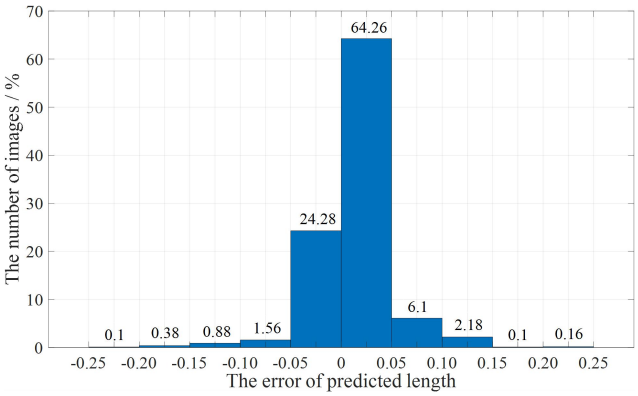


Figure 3. The accuracy of length prediction.

improved from 45.2 (row c) to 97.5 (row e). This suggests that it is necessary to prevent tokens from depending on `[MASK]` tokens. Otherwise, it may reduce the weights of other text tokens and lead to the reduction of the signal-to-noise ratio. The length prediction branch facilitates the model to find the end of the caption, and we can see the performance improved from 45.2 (row b) to 92.6 (row d). Lastly, image-free training also helps improve the performance (116.7 to 117.8 from row f to row g).

Concentrated attention mask. In our proposed concentrated attention mask (CAM), we have two modifications: (i) the text token does not depend on `[MASK]` tokens, denoted as T2M; (ii) the `[MASK]` token does not depend on other `[MASK]` tokens, denoted as M2M. Table 2 suggests that both T2M and M2M improves the performance and M2M provides much larger gain (from 92.6 to 115.6). The reason may be that `[MASK]` token's representation is the key to recover new tokens and it is beneficial not to focus on the other `[MASK]` tokens, which carry less information. We also experiment to enable CAM only at the first half or the second half denoising steps, as in the last two rows of Table 2. In the first half ($t \geq 10$), more tokens are `[MASK]` and thus CAM impacts more to the performance. The results also demonstrates CAM is an effective way to improve the performance.

Error rate of the length prediction. Figure 3 shows the error distribution of the length prediction branch. The error is calculated as $(N_L - GT)/GT$, where N_L denotes the predicted length and GT is the ground-truth length. As we can see, most of the predictions are exactly correct, the error rate of 88.54% images are less than 0.05. The max error rate is less than 0.25. This suggests the predicted length is reliable to determine the token length in decoding.

Embedding for time step t. We use the sinusoidal positional embedding as in Eqn. 4 to map the time step t to an embedding representation. Another way is to use a learnable embedding layer. Table 3 compares these two different methods. The results indicate the sinusoidal positional embedding with appropriate rescaling factor can achieve better performance.

Image-free training. The impact of hyper-parameters (*i.e.*, r and s) is shown in Figure 4. When the training ratio r is 0.2 and the inference scale is 1.17, our model achieve the best performance. It is worth noting that the best guidance scale range is different to the one which is normally set to 5 for image generation [64]

Method	C	B@4	M	R	S
DDCap w/o Adaptive T	115.5	34.2	28.0	57.3	21.3
DDCap w/ Adaln T	115.1	34.1	28.0	57.1	21.3
DDCap w/ sin/cos T (4000)	115.9	34.5	28.0	57.4	21.4
DDCap w/ sin/cos T (6000)	114.8	33.7	27.9	57.0	**21.5**
DDCap w/ sin/cos T (8000)	**116.7**	**34.6**	**28.1**	**57.6**	**21.5**
DDCap w/ sin/cos T (10000)	115.1	34.0	27.9	56.9	21.3

Table 3. The influences of different ways of Adaptive LayerNorm layers. "Adaln T" denotes the normal embedding layer, and " sin/cos T" denotes the "Sinusoidal PosEmb" layer.

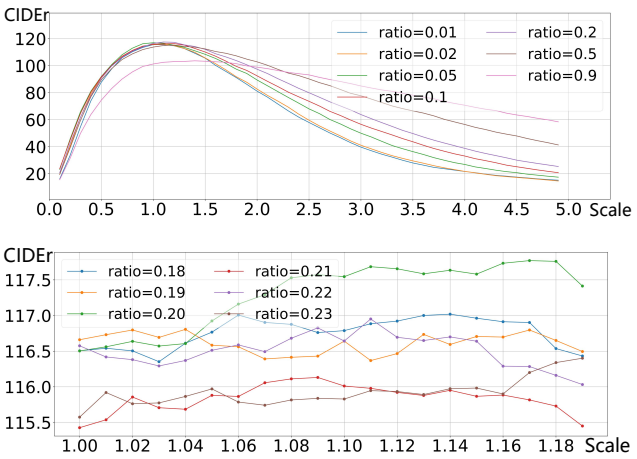


Figure 4. The guidance scale s and ratio r of image-free training. Top: The interval of the scale is 0.1. Bottom: The interval of the scale is 0.01.

Method	C	B@4	M	R	S
AR Basline	112.8	33.9	28.0	56.4	21.3
Continuous Diffusion	91.9	25.0	25.0	51.1	19.1
DDCap	**117.8**	**35.0**	**28.2**	**57.4**	**21.7**

Table 4. Comparison to an auto-regressive (AR) baseline and our implementation of a continuous diffusion model for captioning.

vs AR approach and continuous diffusion model. We compare our proposed method with an auto-regressive (AR) baseline and our implementation of a continuous diffusion model in a controlled way. For AR approach, we use the same image encoder to extract the image representation, and re-use our diffusion transformer network as the decoder, such that the number of model parameters are similar. For the continuous diffusion model, we map the discrete text tokens to a continuous reprentation through the pretrained embedding layer of GPT-2, and use the diffusion transformer to recover these embeddings. The number of total step T is set to 10,000, and the noise schedule is linear. We find it helps with the following modifications: 1) normalize the embedding layers with the mean and variance of all embedding vectors; 2) during inference, each embedding vector x_t is decoded to tokens and then re-embedded to vectors before estimating x_{t-1}. The training epochs are set to 30, 100 and 30, for AR model, continuous diffusion model, and our discrete diffusion model, respectively.

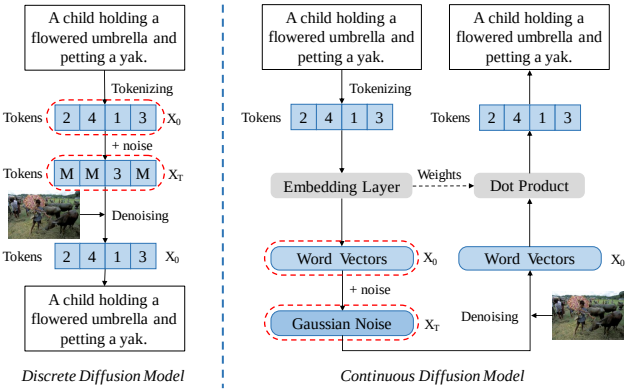


Figure 5. An overview of using *discrete* and *continuous* diffusion models for image captioning. Empirically, discrete diffusion achieves better performance, which is also the focus of this paper.

The results are shown in Table 4, which indicates that our discrete diffusion model demonstrates better performance than auto-regressive model and the carefully designed continuous diffusion model. Compared with auto-regressive model, the discrete diffusion model can utilize bidirectional context in text. Compared with continuous diffusion model, discrete diffusion model aligns better as the text is naturally discrete. Moreover, the differences between continuous and discrete diffusion models are shown in Figure 5.

4.3. Comparison with Prior Arts

Table 5 presents the comparison results on the COCO dataset. Our DDCap shows competitive performance. Specifically, compared with non-autoregressive methods, our model achieves the best performance, which suggests the effectiveness of our key designs. Compared with auto-regressive methods, our performance is better than many methods and is comparable to ViTCap [13]. Hopefully, our study on the diffusion model can motivate more efforts on this direction.

4.4. Caption Infilling Tasks

As the text tokens are not decoded from a fixed left-to-right order, our approach is naturally suitable for the fill-in task task, which targets to fill in the empty words in a sentence as shown in Figure 6. To evaluate the performance, we remove all the adjective words on the "Karpathy" validation dataset. The adjectives are detected by nltk[1]. The captions with no detected words are removed. Then, the to-

[1] https://www.nltk.org/

Method	#Param.	#Images	C	B@4	M	S
Auto-regressive models						
UVLP [84]	111.7M	4M	116.9	36.5	28.4	21.2
MiniVLM [68]	34.5M	14M	119.8	35.6	28.6	21.6
DistillVLM [14]	34.5M	7M	120.8	35.6	28.7	22.1
UFO_B [67]	0.1B	4M	122.8	36.0	28.9	22.2
$OSCAR_B$ [40]	0.1B+64M†	7M	123.7	36.5	30.3	23.1
$UNIMO_B$ [39]	-	9M	124.4	38.8	-	-
ViTCap [13]	0.2B	4M	125.2	36.3	29.3	22.6
$VinVL_B$ [81]	0.1B+0.2B†	6M	129.3	38.2	30.3	23.6
GIT_B [69]	129M	4M	131.4	40.4	30.0	23.0
$LEMON_B$ [28]	111.7M	0.2B	133.3	40.3	30.2	23.3
$SimVLM_B$ [71]	-	1.8B	134.8	39.0	32.9	24.0
Non-autoregressive models						
MNIC [19]	-	-	108.5	31.5	27.5	21.1
$NAIC_{B,KD}$ [21]	-	-	115.5	35.3	27.3	20.8
FNIC [16]	-	-	115.7	36.2	27.1	20.2
DDCap (Ours)	280.1M	4M	**125.1**	**37.1**	**29.1**	**22.7**

Table 5. Performance comparison on COCO captioning Karpathy [32] split with pretraining, where B@4, M, R, C denote BLEU@4, METEOR, ROUGE-L, CIDEr and SPICE scores. CIDEr optimization is not used for all models. (†) VinVL/OSCAR: the extra parameters are for object detector. All of the results do not contain CIDEr optimization.

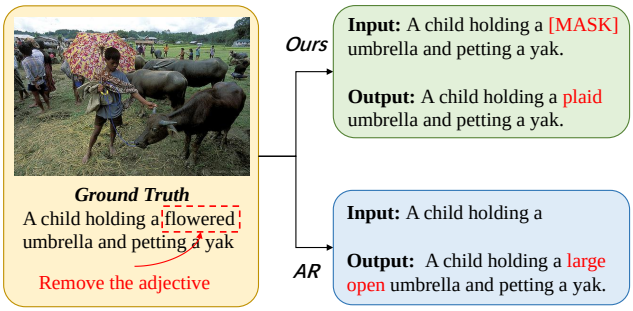


Figure 6. Visualization of the caption infilling task.

Method	C	B@4	M	R	S	CLIP-Score
AR baseline	203.5	76.3	49.3	89.1	36.5	75.7
DDCap	**230.3**	**85.1**	**56.3**	**93.1**	**39.9**	**76.4**

Table 6. Comparison between our model and an auto-regressive baseline for the caption infilling task.

be-filled positions are filled with the mask token to start a diffusion process to recover the full sentence. As for evaluation metrics, we add the CLIP scores [23] for semantic matching. The comparison are shown in Table 6, suggesting the bi-directional model can well solve the task.

4.5. Qualitative Analysis

Figure 8 shows the text tokens in the intermediate steps. Empirically, simple objects and articles are predicted first, followed by preposition and adjectives. Finally, DDCap will add nouns to make the sentence grammatically correct. In other words, the way of describing the image roughly follows the order of first to describe the objects and then the relationship between the objects.

Furthermore, Figure 7 shows prediction examples on the MSCOCO validation set, which shows reasonable prediction results. Compared with auto-regressive baseline, our DDCap has obvious advantages in recognizing objects. This advantages maybe come from our concentrated attention mask which provide more text context.

However, our DDCap sometimes generates some repeated words. This issue may be handled by two ways: (i) we first get the caption, and then find the these repeated words. these repeated words will be re-masked and re-generated. (ii) the repeated words in a caption will be deleted. Then we get a new length of this caption without repeated words. According to the new length, we re-generate the whole caption.

5. Conclusions

This paper introduces a novel discrete diffusion model, termed DDCap, for image captioning. DDCap proposes several novel designs: length prediction, concentrated attention mask, and best-first inference. Ablation experiments under controlled settings demonstrate the effectiveness of each component. Results on COCO datast show that DDCap is comparable with state-of-the-art auto-regressive methods. In addition, we have introduced a new caption infilling task to highlight our advantages. For future work, we plan to perform larger-scale pre-training, and we hope our work can inspire more efforts on using diffusion models for text generation.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 5

[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 *CVPR*, pages 6077–6086, 2018. 2

[3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021. 3, 4

DDCap: A woman riding a motorcycle down a street.

AR: A woman riding on the back of a motorcycle.

GT: A woman is riding a motorcycle down the street.

DDCap: A red fire hydrant spewing water onto a street.

AR: A red fire hydrant spewing water onto a street.

GT: A red fire hydrant is leaking onto a side walk.

DDCap: A bird is perched on top of a tree branch.

AR: A bird sitting on a tree branch with leaves.

GT: A bird with red eyes perched on top of a tree branch.

DDCap: A group of people playing frisbee by a lake.

AR: A herd of elephants wading in a lake.

GT: A picture of some people playing with a frisbee.

DDCap: A yellow and yellow toilet with a red helmet on it.

AR: A yellow fire hydrant with a red hat on top of it.

GT: A red helmet is on a yellow toilet in the dirt.

DDCap: A traffic light next to a street at night.

AR: A traffic light on a pole on a street.

GT: A traffic signal sitting next to a street at night.

DDCap: A cat laying on a chair with a stuffed animal.

AR: A cat is sitting on a stuffed animal.

GT: A cat lays down on a stuffed animal.

DDCap: An elephant is standing in a grassy field.

AR: A large elephant standing on a lush green field.

GT: An elephant is in some brown grass and some trees.

DDCap: A small airplane that is sitting on a runway.

AR: A small yellow and red plane on a runway.

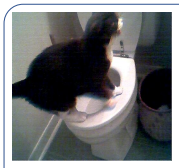
GT: A small airplane taking off from an airport runway.

DDCap: A large bus that is parked in a lot.

AR: A man standing in front of a large bus stop.

GT: A large campaign trailer parked in a parking lot.

Figure 7. Visualization of our model and auto-regressive baseline on validation images of COCO dataset. The wrong parts of a caption are highlighted.



t=19 **[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]** . . .
t=14 A cat **[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]** <|EOS|>
t=9 A cat **[MASK] [MASK] [MASK] [MASK] [MASK] [MASK]** a toilet. <|EOS|>
t=4 A cat **[MASK] [MASK]** on the **[MASK]** of a toilet. <|EOS|>
t=0 A cat is standing on the rim of a toilet. <|EOS|>
GT A cat on the lid of a toilet looking perturbed <|EOS|>

t=19 **[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]** . . .
t=14 A **[MASK]** fire **[MASK]** rant **[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]**
t=9 A blue fire hydrant **[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]** <|EOS|>
t=4 A blue fire hydrant on the **[MASK]** of **[MASK] [MASK]**. <|EOS|>
t=0 A blue fire hydrant on the side of a road. <|EOS|>
GT A yellow and blue fire hydrant sitting on a sidewalk.<| EOS |>

Figure 8. Visualization of the intermediate steps of the caption generation process with our model.

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005. 5

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020. 3, 5

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10578–10587, 2020. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1, 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021. 5

[10] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*, 2022. 3

[11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022. 3, 5

[12] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. In *NeurIPS*, 2021. 3

[13] 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 *CVPR*, 2022. 1, 2, 7, 8

[14] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-

linguistic model via knowledge distillation. In *ICCV*, pages 1428–1438, 2021. 8

[15] Zhengcong Fei, Xu Yan, Shuhui Wang, and Qi Tian. Deecap: Dynamic early exiting for efficient image captioning. In *CVPR*, pages 12216–12226, 2022. 2

[16] Zheng-cong Fei. Fast image caption generation with position alignment. *arXiv preprint arXiv:1912.06365*, 2019. 8

[17] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 3

[18] 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. 2

[19] Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. Masked non-autoregressive image captioning. *arXiv preprint arXiv:1906.00717*, 2019. 2, 8

[20] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 3, 4

[21] Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *arXiv preprint arXiv:2005.04690*, 2020. 2, 8

[22] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019. 2

[23] 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. 2, 8

[24] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, pages 13450–13459, 2022. 2

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 3

[26] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. 2021. 3

[27] 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 *CVPR*, 2022. 3

[28] 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 *CVPR*, pages 17980–17989, 2022. 8

[29] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019. 2

[30] Unnat Jain, Svetlana Lazebnik, and Alexander G. Schwing. Two can play this game: Visual dialog with discriminative question generation and answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2018. 5

[31] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2

[32] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017. 5, 8, 13

[33] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 3

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 5

[35] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *CVPR*, pages 17969–17979, 2022. 2

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[37] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 3

[38] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[39] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 8

[40] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 1, 3, 8

[41] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022. 3

[42] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *CVPR*, pages 17990–17999, 2022. 2

[43] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014. 5

[44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3

[45] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2

[46] Zihang Meng, David Yang, Xuefei Cao, Ashish Shah, and Ser-Nam Lim. Object-centric unsupervised image captioning. In *ECCV*, pages 219–235. Springer, 2022. 2

[47] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *ECCV*, pages 167–184. Springer, 2022. 2

[48] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 3

[49] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 5

[50] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, pages 10971–10980, 2020. 2

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[52] 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 *ICML*, pages 8748–8763. PMLR, 2021. 4, 5

[53] 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

[54] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 3

[55] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024, 2017. 2

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3

[57] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3

[58] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 4

[59] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 5

[60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 3

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1, 3

[62] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 3

[63] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 3

[64] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 2, 5, 6

[65] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 5

[66] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1, 2

[67] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021. 1, 8

[68] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Minivlm: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946*, 2020. 8

[69] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1, 3, 8

[70] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 3

[71] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 3, 8

[72] Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. Difnet: Boosting visual information flow for image captioning. In *CVPR*, pages 18020–18029, 2022. 2

[73] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212, 2016. 2

[74] 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 *ICML*, 2015. 2

[75] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019. 2

[76] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 2

[77] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *ICCV*, 2019. 2

[78] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017. 2

[79] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016. 2

[80] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 3

[81] 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 *CVPR*, 2021. 1, 3, 8

[82] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022. 3

[83] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *CVPR*, pages 15465–15474, 2021. 2

[84] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020. 3, 8

Appendices

Hyperparameters	Image encoder	Diffusion model
Layers	12	12
Hidden size	768	768
Attention heads	12	12
Patch size	16 * 16	-
Adaptive layer norm	-	sinusoidal
Noise schedule	-	linear schedule
The total step T	-	20
Training steps	30k	30k
Batch size	512	512
AdamW ϵ	1e-8	1e-8
AdamW β	(0.9, 0.999)	(0.9, 0.999)
Peak learning rate	2e-5	2e-4
Learning rate schedule	Cosine	Cosine
Warmup steps	6k	6k
Drop Path	0.1	-
Weight decay	0.01	0.01
Image resolution	224	-
Text length	-	20

Table 7. Hyperparameters for training DDCap on COCO captioning without pretraining.

Hyperparameters	Image encoder	Diffusion model
Layers	12	12
Hidden size	768	768
Attention heads	12	12
Patch size	16 * 16	-
Adaptive layer norm	-	sinusoidal
Noise schedule	-	linear schedule
The total step T	-	20
Training steps	142k	142k
Batch size	1024	1024
AdamW ϵ	1e-8	1e-8
AdamW β	(0.9, 0.999)	(0.9, 0.999)
Peak learning rate	1e-5	1e-4
Learning rate schedule	Cosine	Cosine
Warmup steps	47k	47k
Drop Path	0.1	-
Weight decay	0.01	0.01
Image resolution	224	-
Text length	-	20

Table 8. Hyperparameters for pretraining DDCap.

A. Ablation on the "Karpathy" test dataset

In the main paper, we report the ablation study on "Karpathy" [32] validation dataset. Here, we report the results on the test set in Table 10, in which we have consistent conclusions.

Hyperparameters	Image encoder	Diffusion model
Training steps	30k	30k
Batch size	512	512
Peak learning rate	7e-6	1e-5
Warmup steps	6k	6k

Table 9. Hyperparameters for fine-tuning DDCap on COCO captioning.

B. Hyperparameters

The hyperparameters for training DDCap on COCO captioning are shown in Table 7 with no pretraining. The hyperparameters for pretraining DDCap are shown in Table 8.

The hyperparameters for fine-tuning DDCap on COCO captioning are shown in Table 9.

#Row	Best-first inference	CAM	Length Prediction	Image-free training	C	B@4	M	R	S
a					20.4	7.1	18.8	34.3	12.2
b		✓			39.0	10.5	20.3	38.9	14.1
c	✓				45.4	20.4	26.9	47.2	21.4
d	✓		✓		95.3	27.8	25.8	52.1	19.3
e	✓	✓			97.2	27.7	28.0	53.8	**21.8**
f	✓	✓	✓		116.6	**34.4**	28.0	**57.2**	21.3
g	✓	✓	✓	✓	**117.9**	**34.4**	**28.1**	57.1	21.6

Table 10. Ablation study on the effectiveness of each component on the "Karpathy" test dataset. CAM: concentrated attention mask.