Prefix-diffusion: A Lightweight Diffusion Model for Diverse Image Captioning

Anonymous ACL submission

Abstract

While impressive performance has been achieved in image captioning, the limited diversity of the generated captions and the large parameter scale remain major barriers to the real-word application of these systems. In this work, we propose a lightweight image captioning network in combination with continuous diffusion, called Prefix-diffusion. To achieve diversity, we design an efficient method that injects prefix image embeddings into the denoising process of the diffusion model. In order to reduce trainable parameters, we employ a pre-trained model to extract image features and further design an extra mapping network. Prefix-diffusion is able to generate diverse captions with relatively less parameters, while maintaining the fluency and relevance of the captions benefiting from the generative capabilities of the diffusion model. Our work paves the way for scaling up diffusion models for image captioning, and achieves promising performance compared with recent approaches.[1]

1 Introduction

Image captioning, which combines computer vision (CV) and natural language processing (NLP), focuses mainly on producing a description of an image. Existing works on image captioning typically employ an encoder-decoder architecture (Vinyals et al., 2015; Anderson et al., 2018; Zhou et al., 2020) to generate captions word-by-word. However, such models require large trainable parameters to bridge the visual and textual representations. By utilizing the powerful representation capability of pre-trained models like CLIP(Radford et al., 2021), recent methods (Lovenia et al., 2022; Zhu et al., 2022; Mokady et al., 2021) map visual semantic information to language space for image captioning. Although autoregressive models have become the typical approach for image captioning, their left-to-right generative manner leads to cumulative errors. Moreover, human-like captions not only maintain fluency and relevance properties, but also contain diverse wordings and rich expressions.

Figure 1: The diverse captions generated by Prefix-diffusion. The model is trained on the COCO dataset. More examples will be given in the supplementary material.

Recently, the popular diffusion model (Sohl-Dickstein et al., 2015), which generates samples through an iterative denoising process, has provided a promising path to generate tokens in parallel and inherently increase the diversity of captions. Diffusion models (Sohl-Dickstein et al., 2015) have become an active area of research owing to their ability to generate comparable results with GANs (Goodfellow et al., 2020) on computer vision tasks. The strength of diffusion models trained on vast image databases has led to an almost ubiquitous fascination among researchers in producing highly typical content, such as image generation and editing (Nichol et al., 2021; Balaji et al., 2022; Kim et al., 2022; Gal et al., 2022). Nevertheless, the path is blocked by the discreteness of texts and the gap between different modals.

For the continuous diffusion models (Ho et al.,

[1]Code will be released upon publication.

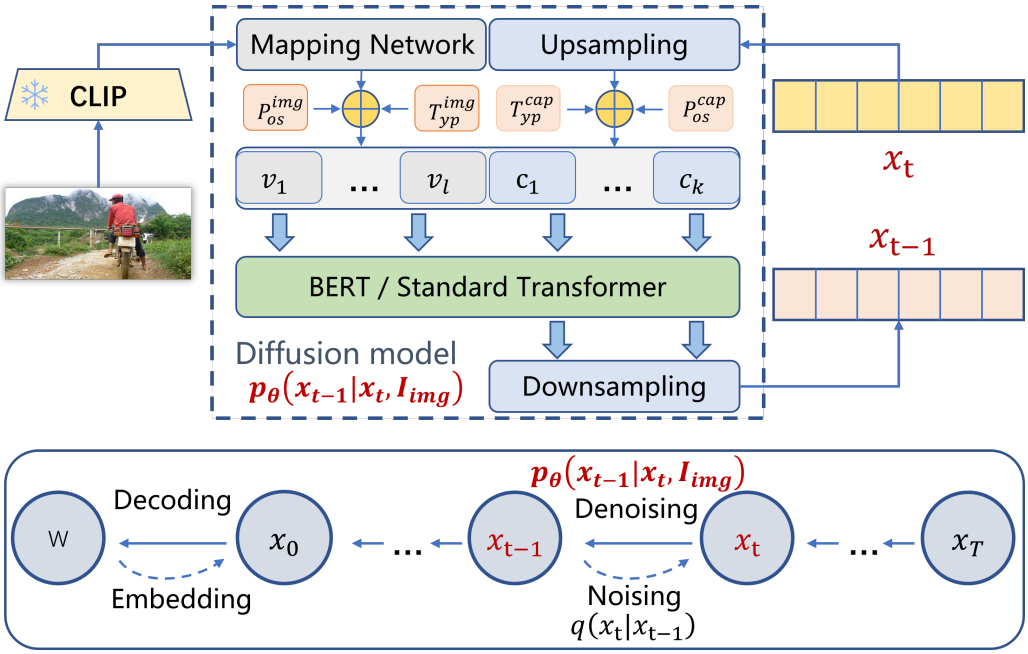

Figure 2: Illustration of Prefix-diffusion. The bottom lies the diffusion process. The reverse process is defined by $p_\theta\left(x_{t-1} \mid x_t, I_{img}\right)$ and the diffusion model is depicted in the upper dashed box. We use the frozen CLIP to extract image features and train a lightweight mapping network to connect the image space and the text space.

2020; Nichol and Dhariwal, 2021; Song et al., 2020), they only work on continuous data but yield inferior results in generating text and image captioning, especially compared to the results of the autoregressive models. To effectively benefit from continuous diffusion, Diffusion-LM (Li et al., 2022) extends the the standard diffusion process with an embedding step followed by a rounding step, generating the high-quality text under six control targets. The discreteness of texts has been overcome, whereas the gap between different modals stays unsolved. For image captioning with continuous diffusion, it is a more challenging task, which further requires the fusion of the image information.

In this paper, we propose a lightweight captioning model based on the continuous diffusion, namely Prefix-diffusion. The model tackles three key problems in image caption generation. Firstly, we utilize diffusion models to solve the limited diversity of the generated captions. Noticing that diffusion models have the powerful generative capabilities but few research applied them to image captioning. Secondly, different from image captioning models that have a large number of parameters and are computationally expensive, our framework saves computing resources with the pretrained CLIP model to extract image features. Last but not least, our method is able to generate more accurate captions in parallel, since it injects prefix image embeddings into the denoising process of the diffusion model. This essentially solves the problem of sequential error accumulation.

Figure 1 shows the captions generated by Prefix-diffusion, where the captions accurately describe the content of the image with fluency. Different from the method of beam search, our method can cover all distributions of the training datasets and generate diverse captions.

The overall contributions of our work are:

- We propose a lightweight method Prefix-diffusion to generate diverse captions. Our work tackles the multi-modal issue for the diffusion model and paves the way for scaling it up for image captioning.
- Prefix-diffusion generates diverse captions in a variety of forms, which is specifically reflected in the increase of Dist-3 and vocabulary usage by 6.3 and 3.1 compared with the baselines, respectively.
- Prefix-diffusion reduces more than 38% trainable parameters compared with existing CLIP-based methods(Nukrai et al., 2022; Mokady et al., 2021), while achieving comparable or even better results in newer metrics.

2 Related Work

2.1 Image Captioning

The autoregressive models achieve promising performance on image captioning. The next token of the caption is conditioned on the former tokens. To generate more neural captions, (Lu et al., 2018) predicts the slot locations that are explicitly tied to image regions. GET (Ji et al., 2021) captures a more comprehensive global representation by using a novel transformer architecture, to guide the caption generation. Similarly, (Li et al., 2019; Luo et al., 2021) use transformer to leverage the image information efficiently. Thanks to the powerful multimodal representation capability of CLIP (Radford et al., 2021), (Mokady et al., 2021; Galatolo et al., 2021) take an image embedding as the input which is encoded by the CLIP visual encoder. Then they use the GPT-2 (Radford et al., 2019) model to produce a sequence of words that describe the content of the input image. But autoregressive models suffer from the limitation of generation speed and the accumulation of errors.

Non-autoregressive models have recently attracted attention due to their fast inference speed and generation quality. (Gao et al., 2019) randomly masks the input sequences with certain ratios to train a masked language model, and generates captions parallelly during inference. Considering non-autoregressive image captioning as a cooperative multi-agent problem, (Guo et al., 2020) proposes a novel counterfactuals-critical multi-agent learning algorithm to improved the inference speed. (Fei, 2020) proposes a non-autoregressive image captioning approach based on the idea of iterative back modification, which refines the output in a limited number of steps. To determine the length of the image caption, (Deng et al., 2020) designs a non-autoregressive decoder for length-controllable image captioning.

2.2 Diffusion Model

Diffusion models (Sohl-Dickstein et al., 2015) have demonstrated impressive capabilities in creative applications. For text-to-image generation, a task of generating a corresponding image from a description, (Balaji et al., 2022; Nichol et al., 2021; Rombach et al., 2022; Gu et al., 2022) apply discrete diffusion models to produce high-resolution images conditioned on the text prompts. Diffsound (Yang et al., 2022) proposes a novel decoder based on the diffusion model to generate high-quality sound. Similarly, ProDiff (Huang et al., 2022) studies on diffusion parameterization for text-to-speech and achieves superior sample quality and diversity. In the text generation domain, Diffusion-LM (Li et al., 2022) starts with a sequence of Gaussian noise vectors and denoises them incrementally into vectors corresponding to words. Diffusion-LM enables efficient gradient-based methods for controllable generation, achieving promising results in the new forms of complex fine-grained control tasks. Moreover, (Gong et al., 2022; Strudel et al., 2022) extend vanilla diffusion models to learn conditional text generation. However, few research applies the diffusion model to image captioning, because of the cross-modal challenge and the discreteness of texts.

3 Methodology

As illustrated in Figure 2, we propose Prefix-diffusion for injecting image features to learn image captioning. Different from image generating, our method requires to map discrete texts to a continuous space by a word embedding. For the conditioned image, we first extract its features by the CLIP image encoder, and then input them to the mapping network to obtain the prefix image embeddings. We then concatenate the prefix image embeddings and the caption embeddings in the denoising process of the diffusion model. The concatenated vectors are fed into a deep neural network (e.g. BERT(Kenton and Toutanova, 2019) or the standard transformer). Since our work merely trains a mapping network and a neural network, the trainable parameter scale is reduced significantly.

Forward process. Following Diffusion-LM (Li et al., 2022), we adopt an embedding function $EMB(W)$ to map a discrete word into a continuous space. Define a caption W with k words. Through the embedding function, we have $EMB(W) = [EMB(\omega_1), ..., EMB(\omega_k)] \in \mathbb{R}^{k \times d_1}$, where d_1 is the dimension of the vector. In our experiments, we find that the value of d_1 works well at 48. Reducing the dimension will decrease the performance, while increasing the dimension will enlarges the computational burden.

For the forward process, diffusion models (Ho et al., 2020; Nichol and Dhariwal, 2021; Song et al., 2020) add noise progressively to training a sample according to a variance schedule $\beta_1, ..., \beta_T$. The forward process has no learnable parameters and

we get x_t by the following equation:

$$x_t = \sqrt{1-\beta_t} x_{t-1} + \sqrt{\beta_t}\epsilon \tag{1}$$

where $\epsilon \sim N(0,1)$ and $\beta_t : 0.01 \longrightarrow 0.03$ are hyperparameters representing the variance schedule across diffusion steps. We have tried different noise methods, with the truncation linear noise schedule method being the best. We validate this observation in section 4.3.3.

Reverse process. The reverse process generates new samples from $x_T \sim N(0, I)$. The data is sampled using the following reverse diffusion process:

$$p_\theta(x_{t-1} \mid x_t, I_{img}) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, I_{img}), \sigma(t)^2 I\right) \tag{2}$$

where I_{img} denotes the visual information from CLIP.

In order to learn the reverse process, neural networks are trained to predict μ_θ and $\sigma(t)^2$ is a fixed variance.

$$\mu_\theta(x_t, I_{img}) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 \tag{3}$$

$$\sigma(t)^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{4}$$

In order to get μ_θ, we compute x_0 with the following equation:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\tilde{z}\right) \tag{5}$$

where $\tilde{z}$ can be obtained by deep neural networks (e.g. transformer).

$$\tilde{z} = \Phi(x_t, I_{img}, t). \tag{6}$$

Here Φ denotes the neural network which is depicted in the dashed box in the Figure 2. Since the transformer architecture has been shown to outperform many other architectures on a wide range of text generation tasks, we explored two different transformer architectures as the neural network: BERT and the standard transformer. Different from other continuous diffusion approaches, we inject image features into the transformer architectures. This process changes the original mean in the caption space, as illustrated in Figure 3.

In the following, we will explain in detail how to inject the image information into the model. Firstly we use CLIP to encode image and receive its image features I'_{img}. Then we train a mapping network F on I'_{img} and obtain the visual prefix I^m_{img} of length l:

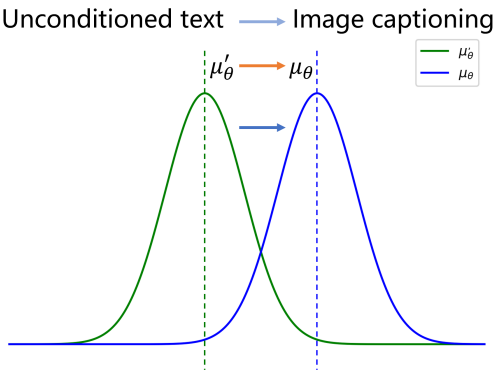

Figure 3: After we concatenate the image features in the reverse process, the original mean μ'_θ is changed to μ_θ in the caption space. Hence, the unconditioned text is converted to an image caption.

$$\begin{cases} I_{img} = CLIP(image) \\ I^m_{img} = \left\{v'_1, v'_2, ..., v'_l\right\} = F(I_{img}) \end{cases}. \tag{7}$$

We specifically formulate $I^m_{img} \in \mathbb{R}^{l \times d_2}$ as $\left\{v'_1, v'_2, ..., v'_l\right\}$ for the convenience of subsequent expression. To save the computation cost, we employ a simple Multi-Layer Perceptron (MLP) as the mapping network. Through an upsampling network, a sequence embedding x_t has the same dimension as I^m_{img}, denoted as $\left\{c'_1, c'_2, ..., c'_k\right\} \in \mathbb{R}^{k \times d_2}$. k is the length of the caption and d_2 is the dimension of the embedding.

Before concatenating the visual prefix embedding and the caption embedding, we add positional embedding P_{os} and type embedding T_{yp} to it:

$$\{c_1, c_2, ..., c_k\} = \{c'_1, c'_2, ..., c'_k\} + P^{cap}_{os} + T^{cap}_{yp} \tag{8}$$

$$\{v_1, v_2, ..., v_l\} = \{v'_1, v'_2, ..., v'_l\} + P^{img}_{os} + T^{img}_{yp}. \tag{9}$$

The positional embedding indicates the model where the feature is located, which is essential information. Similarly, the type embedding tells the model where the image features lie. Then the visual prefix and the caption embedding are concatenated into a sequence $\{v_1, ..., v_l, t_1, ..., t_k\}$, and processed by a standard transformer or BERT

Method		Common Metrics ↑						Similarity Score ↑			Diversity ↑		
		B@1	B@3	M	R-L	C	S	CLIP-S	Ref-CLIP	P-Bert	D@2	D@3	Voc-u
MTIC		80.8	50.9	29.2	58.6	131.2	22.6	60.3	68.6	94.0	7.9	16.3	8.3
DLCT		81.1	51.1	29.4	58.9	133.1	22.8	60.6	69.0	94.1	8.1	17.1	8.3
Frozen Clip Feature	CapDec	68.3	36.6	25.2	51.2	91.7	18.3	60.4	67.8	93.4	8.3	14.9	1.9
	ClipCap	73.6	42.3	**26.7**	54.4	105.8	19.8	60.8	68.6	**93.8**	11.3	21.7	2.6
	Ours(T)	77.7	43.4	25.8	55.8	106.3	19.4	63.4	70.9	93.2	11.2	25.9	4.7
	Ours(B)	**78.1**	**44.2**	26.6	**56.1**	**109.3**	**20.4**	**63.7**	**71.2**	93.7	**12.7**	**28.0**	**5.7**

Table 1: The results of image captioning on COCO. For all the metrics, the higher the better. We use boldface to indicate the best performance. The second best result is underlined. Ours(T) and Ours(B) use a standard transformer and BERT respectively. The values of vocabulary usage are reported at percentage (%).

Method	Common Metrics ↑						Similarity Score ↑			Diversity ↑		
	B@1	B@3	M	R-L	C	S	CLIP-S	Ref-CLIP	P-Bert	D@2	D@3	Voc-u
CapDec	57.6	27.9	20.0	44.5	42.0	14.3	58.0	61.4	92.8	15.5	25.2	1.3
ClipCap	67.0	35.2	**22.5**	49.0	60.8	**16.5**	60.9	65.0	**93.0**	20.9	34.5	1.77
Ours(T)	68.7	34.9	20.1	48.7	53.8	14.2	61.6	66.3	92.2	23.1	41.0	3.6
Ours(B)	**71.0**	**36.2**	21.1	**49.3**	**61.4**	15.2	**64.7**	**68.6**	92.0	**27.6**	**46.0**	**4.0**

Table 2: The results of image captioning on Flickr30k. For all the metrics, the higher the better. We use boldface to indicate the best performance. The second best result is underlined.

network:

$$\{y_1, y_2, ..., y_l, y_{l+1}, ..., y_{l+k}\} = Network(concat(v_1, ..., v_l, c_1, ..., c_k)). \quad (10)$$

We split y_i and use $\{y_{l+1}, ..., y_{l+k}\}$ as the input of the downsampling, yielding the output $x_{t-1} \in \mathbb{R}^{k \times d_1}$ of the diffusion model.

Decoding process. In the decoding process, we strengthen the similarity of images and captions with CLIP scores. The benefit of CLIP in the current work is that it can provide a cosine similarity score between numerous texts and an image. Utilizing the CLIP embedding of an image, we calculate the cosine similarity between the image and the n candidate captions. We then choose the most relevant captions. The similarity is computed as follows:

$$similarity\,(I_{img}, W^n_{txt}) = \frac{I_{img} \cdot W^n_{txt}}{|I_{img}| \cdot |W^n_{txt}|} \quad (11)$$

where I_{img} is the image features extracted by CLIP and W^n_{txt} is the features of the n candidate captions. This is a retrieval-base (Ramos et al., 2022; Zhao et al., 2020) technique that picks the best appropriate caption from a set of candidate captions. We use this approach based on the advantage of Prefix-diffusion: our model can generate diverse captions with different Gaussian noises. We verify the effectiveness of this retrieval-base method in section 4.3.3.

4 Experiment

In this section, we conduct quantitative and qualitative experiments to evaluate our approach. We first introduce the implementation details in subsection 4.1 and 4.2. Then we compare the performance of our approach with the others on various evaluation metrics (subsection 4.3.1 and 4.3.2). Finally, the ablation experiments (subsection 4.3.3) are also presented to analyze the significance of our design.

4.1 Dataset and Evaluation Metric

We use COCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015) as the datasets for image captioning. We split the datasets for training, validation, and test according to the Karpathy et al (Karpathy and Fei-Fei, 2015), where the test sets of the two datasets contain 5000 images and 1000 images respectively. To evaluate the generalization ability of our model, we train the model on one dataset while evaluating on the other.

In this paper, we adopt automatic evaluation to appraise the generated captions. In addition to the common metrics and similarity, we consider two metrics to evaluate the diversity of the generated captions.

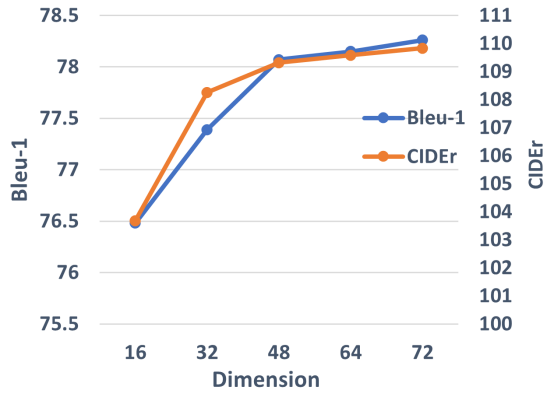

Figure 4: The performance effect of the word dimension on COCO. We report the metrics of Bleu-1 and CIDEr.

Method	Human Evaluation↑			Paras (M) ↓
	Fluency	Sim	Div	
MTIC	3.65	3.63	3.52	38.44
DLCT	3.70	3.25	3.43	63.04
Capdec	3.53	2.95	3.29	178.03
ClipCap	3.83	3.38	3.67	155.91
Ours(T)	3.79	3.84	3.95	**38.25**
Ours(B)	**4.07**	**3.95**	**4.12**	94.83

Table 3: Thr results of human evaluation and the number of trainable parameters for different methods.

- *Common Metrics*. Following the common practice in the literatures, we perform evaluation using BLEU(B@N)(Papineni et al., 2002), METEOR(M)(Denkowski and Lavie, 2014), ROUGE-L(R-L)(Lin and Och, 2004), CIDEr(C)(Vedantam et al., 2015), SPICE(S)(Anderson et al., 2016).

- *Similarity*. We evaluate the generation by newer metrics: CLIP-S and RefCLIPScore (Ref-CLIP)(Hessel et al., 2021) , BERTScore (P-Bert)(Zhang et al., 2020), which achieve higher correlation with human judgmens.

- *Diversity*. Diversity (Li et al., 2016) is a metric that evaluates the diversity of the generated captions. We report Dist-2(D@2) and Dist-3(D@3) by measuring the diversity of bigrams and trigrams in the generation.

- *Vocabulary usage*. To analyze the diversity of the generated captions, according to (Dai et al., 2018), we compute vocabulary usage(Voc-u), which accounts for the percentage of words in the vocabulary that are used in the generated captions.

4.2 Baseline

We adopt the previous competitive image captioning approaches to serve as the baseline models:

MTIC (Cornia et al., 2020): MITC is a transformer-based architecture for image captioning. Its image features extracted are by ResNet (denoted as grid-based features).

DLCT (Luo et al., 2021): DLCT achieves the complementarity of region and grid features for image captioning. To extract visual features, DLCT uses the pretrained Faster-RCNN (Ren et al., 2015).

CapDec (Nukrai et al., 2022): CapDec is a simple and intuitive approach to learning a captioning model based on CLIP.

ClipCap (Mokady et al., 2021): ClipCap leverages powerful vision-language pre-trained models (CLIP) to simplify the captioning process. And we utilize the MLP mapping network and fine-tunes the language model. All the hyper-parameters are set following its original paper.

Since CapDec and ClipCap use CLIP to extract the same image features and freeze CLIP as our model, we use these methods as the primary baselines. We train our model for 200000 steps, with a batch size of 128. The dimension of word embedding is set to 48 and the diffusion steps $T = 1000$. All the experiments are run on NVIDIA Tesla V100 GPUs. In the decoding process, we configure the value of the candidate sentences with $n = 5$. Specifically, during the evaluation, we set the denoising steps $T = 50$, which greatly reduces the generation time.

4.3 Results

4.3.1 Image Captioning

We compare Prefix-diffusion to several baselines with different evaluation metrics, as is shown in Table 1. Our model outperforms all baselines on CLIP-S and Ref-CLIP metrics, and achieves comparable results on P-Bert score, indicating that the effectiveness of the continuous diffusion on image captioning. Not only that, we have a significant improvement on some diversity metrics (such as the D@2 and D@3). Furthermore, Prefix-diffusion covers the largest percentage of words, observed from the vocabulary used to generate captions. It implies that captions generated by Prefix-

Method	Common Metrics ↑						Similarity Score ↑			Diversity ↑		
	B@1	B@3	M	R-L	C	S	CLIP-S	Ref-CLIP	P-Bert	D@2	D@3	Voc-u
COCO⟹ Flickr30k												
CapDec	57.2	23.9	17.1	40.3	30.3	10.8	54.4	58.7	92.1	18.5	29.4	1.2
ClipCap	64.6	29.3	18.9	44.3	44.4	12.5	56.5	61.2	**92.5**	**19.7**	32.7	1.3
Ours(B)	**69.5**	**31.2**	**19.3**	**46.6**	**46.8**	**13.0**	**61.2**	**65.3**	91.9	19.4	**37.0**	**3.0**
Flickr30k⟹ COCO												
CapDec	44.1	15.2	15.7	36.4	25.7	8.6	47.7	51.4	90.4	5.5	10.4	2.0
ClipCap	55.7	**23.5**	**19.2**	42.0	**51.3**	**12.2**	54.9	60.0	**91.1**	11.3	21.3	3.5
Ours(B)	**57.2**	22.4	17.5	**42.5**	49.3	11.3	**57.5**	**62.8**	90.4	**13.6**	**29.9**	**6.6**

Table 4: The results of cross-domain captioning. COCO⟹ Flickr30k means model trained on COCO while evaluated on Flickr30k, and so is Flickr30k⟹ COCO. We use boldface to indicate the best performance.

diffusion contain diverse wordings and rich expressions. Our model can generate high-quality captions compared with captioning approaches that extract image feature with CLIP. Prefix-diffusion performs worse than MTIC and DLCT (who not use freeze features for image captioning) on the common metrics, partially due to the proven limitations of word-overlapping-based metrics across various domains(Hessel et al., 2021; Zhang et al., 2020), and also because our generation is more diverse in expression and correctly describe the visual content, which can be observed from similarity score and diversity metrics.

We also conduct experiments on dataset of Flickr30k, as presented in Table 2, from which we can draw similar conclusions with the dataset of COCO. Our model achieves impressive performance in the image captioning task compared to the baseline models. In detail, from the results of diversity metrics, we notice that the metrics of Dist-3 and vocabulary usage increase by more than 6.0 and 3.0, respectively. Additionally, we also observe an improvement of 2.6 and 2.8 in CLIP-S and Ref-CLIP metrics, respectively. This indicates that the diffusion model can effectively improve the caption diversity while ensuring coherence and relevance in the generated captions. To generate diverse captions, existing methods tend to generate different captions via top-k sampling. Intuitively, such methods may ignore syntactic diversity and semantic diversity that humans are really interested in. Unlike existing methods, Prefix-diffusion seeks to generate multiple captions with rich expressions from different Gaussian noises.

Figure 1 shows the captions generated by Prefix-diffusion. It is observed that the generated captions are pretty consistent with the image as well as keeping the qualified fluency. Meanwhile, our model is able to generate diverse captions that are more like human-generated.

Furthermore, we conduct human evaluation and report the number of trainable parameters to validate the applicability of our method. As is shown in Table 3, our model only requires a small number of model parameters. It brings potential advantages of saving memory storage space and computing costs, and thus being much more useful in practice. For human evaluation, we randomly selected 20 samples and presented them in a shuffled manner to 20 annotators. The annotators rated the fluency, similarity(Sim), and diversity(Div) of the captions on a scale from 1 to 5, with higher scores indicating better quality. From the human evaluation results, We can draw similar conclusions with the automatic evaluation. Our model outperforms the baselines in diversity while holding better fluency and relevance.

The dimension of word embeddings is an important hyper-parameter. The higher dimension leads to more training time and memory usage. To further study the effect of embedding dimension in Prefix-diffusion, we conduct experiments by training with different dimensions. As is shown in Figure 4, the metrics of Bleu-1 and CIDEr are improved as the embedding dimension increases. The reason is that a word embedding becomes richer with semantic information due to the higher dimension. However, there is a performance bottleneck when we continue to increase the dimension of word embeddings. It is observed that the performance trends to be stable when the dimension goes beyond 48.

n	Common Metrics ↑						Similarity Score ↑			Diversity ↑		
	B@1	B@3	M	R-L	C	S	CLIP-S	Ref-CLIP	P-Bert	D@2	D@3	Voc-u
1	77.2	43.6	26.0	55.6	105.2	19.5	60.4	68.6	93.1	11.9	26.4	5.4
5	78.1	**44.2**	**26.6**	**56.1**	**109.3**	**20.4**	63.7	71.2	**93.7**	12.7	28.0	5.7
10	**78.3**	43.8	**26.6**	56.0	109.1	20.3	65.3	72.2	93.4	13.1	28.8	5.8
15	78.2	43.4	26.5	55.8	108.5	20.3	**66.0**	**72.6**	93.4	**13.4**	**29.3**	**5.9**

Table 5: The effect of different values of candidate captions. $n = 1$ means no cosine similarity calculation in the decoding process.

Noise Schedule	Metrics ↑			
	B@1	CLIP-S	Ref-CLIP	P-Bert
Square	70.5	**66.8**	72.2	92.6
Linear	70.4	65.9	71.6	92.3
Cosine	70.5	66.5	72.0	92.5
T-Cosine	72.5	66.5	**72.3**	92.9
T-Linear	**78.1**	63.7	71.2	**93.7**

Table 6: The analysis of different noise schedule in the forward process. T-Linear and T-Cosine means truncation linear noise schedule and truncation cosine noise schedule respectively.

4.3.2 Cross-domain Captioning

We also conduct experiments on cross-domain captioning to evaluate the generalization capability of Prefix-diffusion. The results of the cross-domain evaluation are shown in Table 4. We train the model on the dataset of a source domain while evaluating it on another dataset. From the results of COCO$\Longrightarrow$Flickr30k, Prefix-diffusion achieves excellent performance over all compared approaches, with the results on the common metrics being the best. In addition, it acquires significant improvements on both Dist-3 and vocabulary usage metrics. This is due to the powerful generative ability of the diffusion model. When we train on flickr30k while evaluating on COCO, the results also show that our approach has strong capability in the cross-domain scenario. By comparing the two results, we find that Prefix-diffusion works even better when trained on a larger dataset, implying the better generalization ability.

4.3.3 Ablation

We perform an ablation study on the dataset of COCO to quantify the contribution of each module in Prefix-diffusion.

Table 5 presents the effect on the number of candidate captions. From the two groups of experiments, $n = 1$ and $n = 5$, it can be seen that this selection strategy improves the performance of image captioning. We observe a significant increase in the CIDEr metric, which boosts the CIDEr score from 105.2 to 109.3. It confirms the function of calculating the similarity between the image and the candidate captions and choosing the highest. But too many candidate captions lead to a reduction in the performance of the caption fluency. This is because we use the CLIP score as the only similarity selection metric, which may neglect the fluency of captions.

As presented in Table 6, We investigate the performance of different noise schedules. Observing the results, we conclude that truncated linear noise schedule is able to generate more precise and descriptive captions. We also conclude that the semantic information is corrupted by the complicated noise schedule in the forward process, leading to a more difficult learning problem in the denoising process.

5 Conclusion and Future Work

In this paper, we propose a lightweight network for image captioning in combination with continuous diffusion, called Prefix-diffusion. Experiments and further analysis demonstrate that it can generate diverse captions while maintaining the fluency and relevance of the captions. By trained on one dataset but evaluated on the other, Prefix-diffusion presents remarkable generalization ability. Besides, our model requires a small number of training parameters, which is more applicable in reality. We also conduct ablation experiments to show the effect of the selection strategy and noise schedules. The empirical results verify that Prefix-diffusion has powerful generative ability for image captioning. For future work, we will continue to explore the potential impact of diffusion models on image captioning.

Limitations

As presented in Table 1 and Table 2, though Prefix-diffusion can generate diverse captions with relatively less parameters, it is inferior to MTIC and DLCT on the common metrics. But it performs well on newer metrics which have been shown higher correlation with human generation. The reason is that our generated captions have a rich expression that is inconsistent with the reference text, but still convey the same underlying semantics. The length is an important property as it reflects the amount of information carried by a caption. Since our model is a non-autoregressive model, we cannot control the length of the generated text, leading to a less accurate description of the image. We leave this part of exploration for future work.

Ethics Statement

Since the proposed Prefix-diffusion can be used to generate captions. With the advantages of being accurate, diverse and descriptive, its generation is more like human-generated. This would benefit image captioning applications on downstream tasks, such as chatting robots and automatic voice guide system. On the other hand, the large number of image captions will make it difficult to distinguish human-wrote from machine-generated. Hence, exploring adversarial attacks on image captioning is necessary. Moreover, excellent captions should involve a variety of words and rich expressions, which prevents them from being too dull or tedious. The diffusion model generates new samples from different noises. Therefore, Prefix-diffusion can be used to improve the diversity of the captions.

References

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

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

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

Bo Dai, Sanja Fidler, and Dahua Lin. 2018. A neural compositional paradigm for image captioning. *Advances in Neural Information Processing Systems*, 31:656–666.

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

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

Zhengcong Fei. 2020. Iterative back modification for faster image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3182–3190.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. 2021. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv preprint arXiv:2102.01645*.

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

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605.

Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1655–1663.

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

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8928–8937.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

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

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

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

Holy Lovenia, Bryan Wilie, Romain Barraud, Samuel Cahyawijaya, Willy Chung, and Pascale Fung. 2022. Every picture tells a story: Image-grounded controllable stylistic story generation. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 40–52.

J. Lu, J. Yang, D. Batra, and D. Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2286–2293.

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

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

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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

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

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2022. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. 2022. Self-conditioned embedding diffusion for text generation.

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

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

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2022. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang, and Jiaxuan Zhang. 2020. Image caption generation via unified retrieval and generation-based method. *Applied Sciences*, 10(18):6235.

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

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*.