

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

Guisheng Liu¹, Yi Li^{1*}, Zhengcong Fei³, Haiyan Fu¹, Xiangyang Luo², Yanqing Guo¹

¹Dalian University of Technology, ²Information Engineering University, ³Meituan
lgs0000@mail.dlut.edu.cn, {liyi, fuhy, guoyq}@dlut.edu.cn
xiangyangluo@126.com, feizhengcong@meituan.com

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.

Keywords: diversity, lightweight, diffusion models

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., 2022a; 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.

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 edit-



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.

ing (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., 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 standard diffusion process with an embedding step followed by a rounding step, generating the high-quality text under six control targets. The dis-

Tiền tố-khuếch tán: Một mô hình khuếch tán nhẹ cho Chú thích hình ảnh đa dạng

Guisheng Liu¹, Yi Li¹, Zhengcong Fei³, Haiyan Fu¹, Xiangyang Luo², Yanqing Guo¹

¹Đại học Công nghệ Đại Liên, ²Đại học Kỹ thuật Thông tin, 3Meituan
lgs0000@mail.dlut.edu.cn, {liyi, fuhy, guoyq}@dlut.edu.cn
xiangyangluo@126.com, feizhengcong@meituan.com

Tóm tắt Mục

dù hiệu suất ấn tượng đã đạt được trong chú thích hình ảnh, nhưng tính đa dạng hạn chế của các chú thích được tạo ra và quy mô tham số lớn vẫn là rào cản lớn đối với ứng dụng thực tế của các hệ thống này. Trong công trình này, chúng tôi đề xuất một mạng chú thích hình ảnh nhẹ kết hợp với khuếch tán liên tục, được gọi là khuếch tán tiền tố. Để đạt được tính đa dạng, chúng tôi thiết kế một phương pháp hiệu quả đưa những hình ảnh tiền tố vào quá trình khử nhiễu của mô hình khuếch tán. Để giảm các tham số có thể đào tạo, chúng tôi sử dụng một mô hình được đào tạo trước để trích xuất các đặc điểm hình ảnh và thiết kế thêm một mạng ánh xạ. Khuếch tán tiền tố có thể tạo ra các chú thích đa dạng với tương đồng ít tham số, đồng thời vẫn duy trì tính lưu loát và tính liên quan của các chú thích được hưởng lợi từ khả năng tạo ra của mô hình khuếch tán. Công trình của chúng tôi mở đường cho việc mở rộng quy mô các mô hình khuếch tán để chú thích hình ảnh và đạt được hiệu suất đầy hứa hẹn so với các phương pháp tiếp cận gần đây.

Từ khóa: đa dạng, nhẹ, mô hình khuếch tán

1. Giới thiệu

Chú thích hình ảnh, kết hợp thị giác máy tính (CV) và xử lý ngôn ngữ tự nhiên (NLP), chủ yếu tập trung vào việc tạo ra mô tả về hình ảnh. Các tác phẩm hiện có về chú thích hình ảnh thường sử dụng kiến trúc mã hóa-giải mã (Vinyals và cộng sự, 2015; Anderson và cộng sự, 2018; Zhou và cộng sự, 2020) để tạo chú thích từng từ. Tuy nhiên, các mô hình như vậy yêu cầu các tham số có thể đào tạo lớn để kết nối các biểu diễn trực quan và văn bản.

Bằng cách sử dụng khả năng biểu diễn mạnh mẽ của các mô hình được đào tạo trước như CLIP(Radford và cộng sự, 2021), các phương pháp gần đây (Lovenia và cộng sự, 2022; Zhu và cộng sự, 2022a; Mokady và cộng sự, 2021) ánh xạ thông tin ngữ nghĩa trực quan vào không gian ngôn ngữ để chú thích hình ảnh. Mặc dù các mô hình hồi quy tự động đã trở thành phương pháp điển hình để chú thích hình ảnh, nhưng cách tạo từ trái sang phải của chúng dẫn đến lỗi tích lũy. Hơn nữa, phụ đề giống con người không chỉ duy trì được tính lưu loát và tính liên quan mà còn chứa đựng nhiều từ ngữ đa dạng và cách diễn đạt phong phú.

Gần đây, mô hình khuếch tán phổ biến (Sohl-Dickstein và cộng sự, 2015), tạo ra các mẫu thông qua quá trình khử nhiễu lặp đi lặp lại, đã cung cấp một con đường đầy hứa hẹn để tạo ra các mã thông báo song song và tăng cường tính đa dạng của chú thích.

Các mô hình khuếch tán (Sohl-Dickstein và cộng sự, 2015) đã trở thành một lĩnh vực nghiên cứu tích cực do khả năng tạo ra kết quả tương đương với GAN (Goodfellow và cộng sự, 2020) trên các tác vụ thị giác máy. Sức mạnh của các mô hình khuếch tán được đào tạo trên cơ sở dữ liệu hình ảnh khổng lồ đã dẫn đến sự say mê gần như phổ biến trong số các nhà nghiên cứu trong việc tạo ra nội dung có tính diễn hình cao, chẳng hạn như tạo hình ảnh và chỉnh sửa.



Hình 1: Các chú thích đa dạng được tạo ra bởi Prefix-diffusion. Mô hình được đào tạo trên tập dữ liệu COCO. Nhiều ví dụ hơn sẽ được đưa ra trong tài liệu bổ sung.

(Nichol và cộng sự, 2021; Balaji và cộng sự, 2022; Kim và cộng sự, 2022; Gal và cộng sự, 2022). Tuy nhiên, con đường này bị chặn bởi tính rời rạc của các văn bản và khoảng cách giữa các phương thức khác nhau.

Đối với các mô hình khuếch tán liên tục (Ho et al., 2020; Nichol và Dhariwal, 2021; Song et al., 2020), chúng chỉ hoạt động trên dữ liệu liên tục nhưng mang lại kết quả kém hơn trong việc tạo chú thích văn bản và hình ảnh, đặc biệt là khi so sánh với kết quả của các mô hình tự hồi quy. Để hưởng lợi hiệu quả từ khuếch tán liên tục, Diffusion-LM (Li et al., 2022) mở rộng quy trình khuếch tán tiêu chuẩn với bùa ánh xạ tiếp theo là bùa ánh xạ làm tròn, tạo ra văn bản chất lượng cao theo tiêu chuẩn kiểm soát.

Tác giả liên hệ

12954

* Corresponding author

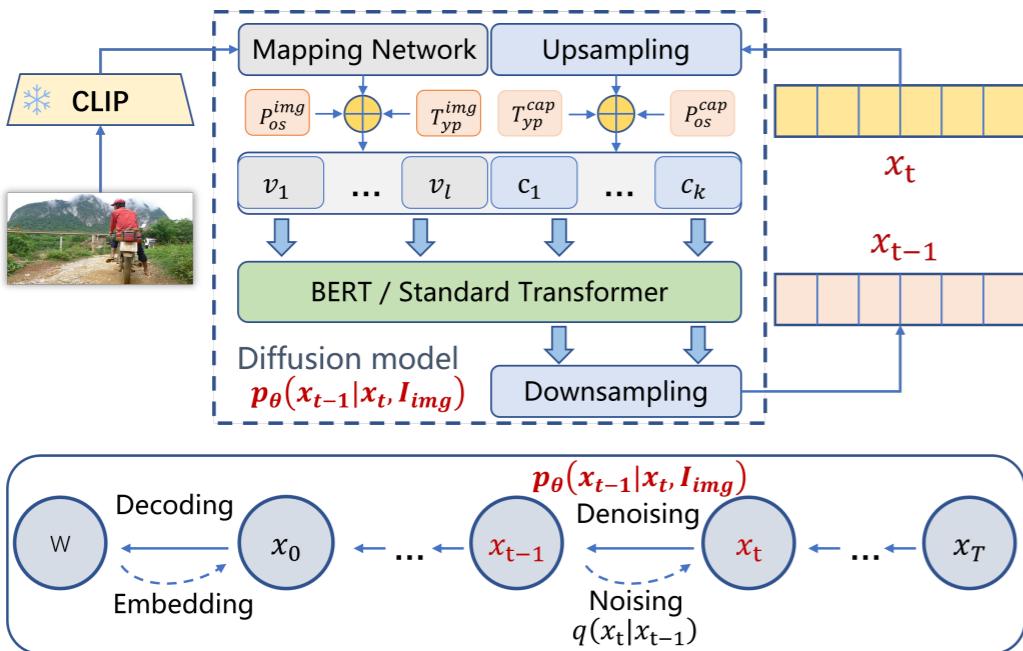


Figure 2: Illustration of Prefix-diffusion. The bottom lies the diffusion process. The reverse process is defined by $p_\theta(x_{t-1} | x_t, I_{img})$ and the diffusion model is depicted in the upper dashed box. We use the frozen CLIP to extract image features and train a lightweight mapping network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

creteness 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 pre-trained 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.¹ Our work tackles the multi-modal issue for the diffusion model and paves the way for scaling it up for image captioning.

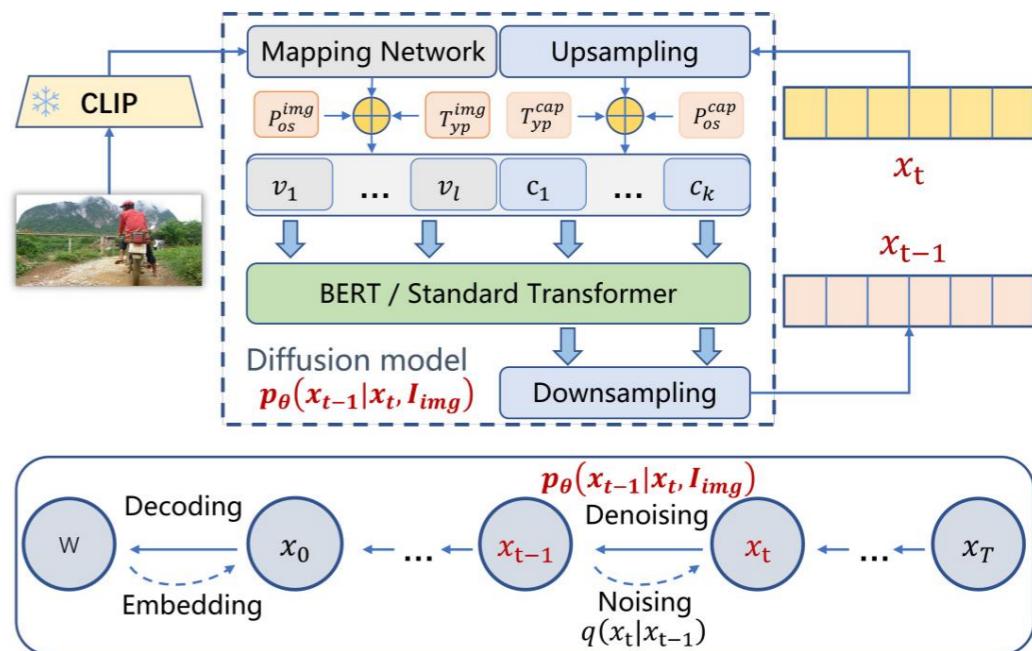
- Prefix-diffusion generates diverse captions in a variety of forms, which is specifically reflected in the increase of Dist-3 and vocabulary usage by 6.3 and 3.1 compared with the baselines, respectively.
- Prefix-diffusion reduces more than 38% trainable parameters compared with existing CLIP-based 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

¹<https://github.com/lgs00/Prefix-diffusion>



Hình 2: Minh họa về khuếch tán tiền tố. Phía dưới là quá trình khuếch tán. Quá trình ngược lại được định nghĩa bởi $p_\theta(x_{t-1} | x_t, I_{img})$ và mô hình khuếch tán được mô tả trong hộp nét đứt phía trên. Chúng tôi sử dụng CLIP đông lạnh để trích xuất các đặc điểm hình ảnh và đào tạo mạng ánh xạ nhẹ để kết nối không gian hình ảnh và không gian văn bản.

Sự thiếu chật chẽ của văn bản đã được khắc phục, trong khi khoảng cách giữa các phuơng thức khác nhau vẫn chưa được giải quyết. Đôi với việc chú thích hình ảnh bằng phuơng pháp khuếch tán liên tục, đây là một nhiệm vụ khó khăn hơn, đòi hỏi phải hợp nhất thông tin hình ảnh.

Trong bài báo này, chúng tôi đề xuất một mô hình chú thích nhẹ dựa trên sự khuếch tán liên tục, cụ thể là khuếch tán tiền tố. Mô hình này giải quyết ba vấn đề chính trong việc tạo chú thích hình ảnh. Đầu tiên, chúng tôi sử dụng các mô hình khuếch tán để giải quyết tính đa dạng hạn chế của các chú thích đã được tạo ra. Lưu ý rằng các mô hình khuếch tán có khả năng tạo mạnh mẽ nhưng ít nghiên cứu áp dụng chúng vào chú thích hình ảnh. Thứ hai, khác với các mô hình chú thích hình ảnh có nhiều tham số và tốn kém về mặt tính toán, khôn khổ của chúng tôi tiết kiệm tài nguyên tính toán với mô hình CLIP được đào tạo trước để trích xuất các đặc điểm hình ảnh. Cuối cùng như ng không kém phần quan trọng, phuơng pháp của chúng tôi có thể tạo ra các chú thích chính xác hơn song song, vì nó đưa nhung hình ảnh tiền tố vào quá trình khử nhiễu của mô hình khuếch tán.

về cơ bản, điều này giải quyết được vấn đề tích tụ lỗi tuần tự.

Hình 1 cho thấy các chú thích được tạo ra bởi Prefix-diffusion, trong đó các chú thích mô tả chính xác nội dung của hình ảnh một cách trôi chảy. Khác với phuơng pháp tìm kiếm chùm tia, phuơng pháp của chúng tôi có thể bao phủ tất cả các phân phối của tập dữ liệu đào tạo và tạo ra các chú thích đa dạng.

Những đóng góp chung của công trình của chúng tôi là:

- Chúng tôi đề xuất một phuơng pháp nhẹ

khuếch tán để tạo ra các chú thích đa dạng.1 Công trình của chúng tôi giải quyết vấn đề đa phuơng thức cho mô hình khuếch tán và mở rộng cho việc mở rộng quy mô để chú thích hình ảnh.

• Sự khuếch tán tiền tố tạo ra nhiều chú thích đa dạng ở nhiều dạng khác nhau, điều này được phản ánh cụ thể ở mức tăng của Dist-3 và mức sử dụng từ vựng lần lượt là 6,3 và 3,1 so với mức cơ sở .

• Phuơng pháp khuếch tán tiền tố làm giảm hơn 38% các tham số có thể đào tạo được so với các phuơng pháp dựa trên CLIP hiện có (Nukrai và cộng sự, 2022; Mokady và cộng sự, 2021), đồng thời đạt được kết quả tương đương hoặc thậm chí tốt hơn ở các số liệu mới hơn.

2. Công trình liên quan

2.1. Chú thích hình ảnh Các mô hình

hồi quy tự động đạt được hiệu suất dày hứa hẹn về chú thích hình ảnh. Mã thông báo tiếp theo của chú thích được điều kiện hóa dựa trên các mã thông báo trước đó. Để tạo ra nhiều chú thích thần kinh hơn, (Lu và cộng sự, 2018) dự đoán các vị trí khe cắm được liên kết rõ ràng với các vùng hình ảnh. GET (Ji và cộng sự, 2021) nắm bắt được biểu diễn toàn cầu toàn diện hơn bằng cách sử dụng kiến trúc biến áp mới để hướng dẫn việc tạo chú thích. Tương tự như vậy, (Li và cộng sự, 2019; Luo

¹<https://github.com/lgs00/Prefix-diffusion>

et al., 2021) use transformer to leverage the image information efficiently. Thanks to the powerful multi-modal 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. DD-Cap (Zhu et al., 2022b) adds a network branch to specifically predict the total token length and design

a concentrated attention mask module to concentrate on more informative tokens. To generate more specific captions, (Kornblith et al., 2023) explore strategies to guide the image captioning model by modifying the decoding distribution. Bit Diffusion (Chen et al., 2022) enables continuous state diffusion models to generate discrete data by utilizing analog bits and a simple thresholding operation for decoding. These methods (Xu et al., 2023; Tang et al., 2024) can be directly modeled by continuous state diffusion models and use the features of CLIP as a guide. Different from existing methods, we extend the line of diverse image description by proposing a lightweight continuous diffusion model, which is essential but has received little attention previously.

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 $EMB(W) = [EMB(\omega_1), \dots, EMB(\omega_k)] \in \mathbb{R}^{k \times d_1}$, 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 train a sample according to a variance schedule β_1, \dots, β_T . The forward process has no learnable parameters and we get x_t by the following equation:

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon \quad (1)$$

where $\epsilon \sim N(0, 1)$ and $\beta_t : 0.01 \rightarrow 0.03$ are hyperparameters representing the variance schedule across diffusion steps. We have tried different noise methods, with the truncation linear noise schedule

(et al., 2021) sử dụng máy biến áp để tận dụng hình ảnh thông tin hiệu quả. Nhờ khả năng biểu diễn đa phuơng thức mạnh mẽ của CLIP (Radford và cộng sự, 2021), (Mokady và cộng sự, 2021; Galatolo và cộng sự, 2021) lấy hình ảnh nhúng làm đầu vào để tạo ra một chuỗi các từ mô tả nội dung của hình ảnh đầu vào. Như các mô hình hồi quy tự động gặp phải hạn chế về tốc độ tạo và sự tích tụ lỗi.

Các mô hình không tự hồi quy gần đây đã thu hút sự chú ý do tốc độ suy luận nhanh của chúng và chất lượng thế hệ. (Gao et al., 2019) ngẫu nhiên che dấu các chuỗi đầu vào với tỷ lệ nhất định để đào tạo một mô hình ngôn ngữ được che giấu và tạo ra chú thích song song trong quá trình suy luận. Xem xét chú thích hình ảnh không tự hồi quy như một vấn đề tác nhân hợp tác, (Guo et al., 2020) để xuất một vấn đề tác nhân phân biện mới thuật toán học tập để cải thiện tốc độ suy luận. (Fei, 2020) để xuất một hình ảnh không tự hồi quy phuơng pháp chú thích dựa trên ý tưởng lặp đi lặp lại sửa đổi ngữ cảnh, tinh chỉnh đầu ra trong một số bước giới hạn. Để xác định độ dài của chú thích hình ảnh, (Deng et al., 2020) thiết kế một bộ giải mã không tự hồi quy cho chiều dài có thể kiểm soát chú thích hình ảnh.

2.2. Mô hình khuếch tán

Các mô hình khuếch tán (Sohl-Dickstein và cộng sự, 2015) có đã chứng minh khả năng ẩn tuồng trong các ứng dụng sáng tạo. Đối với việc tạo văn bản thành hình ảnh, một nhiệm vụ tạo ra một hình ảnh tư duy ứng từ một mô tả, (Balaji et al., 2022; Nichol et al., 2021; Rombach et al., 2022; Gu et al., 2022) áp dụng rời rạc mô hình khuếch tán để tạo ra hình ảnh có độ phân giải cao có điều kiện trên các lời nhắc văn bản. Diffsound (Yang et al., 2022) để xuất một bộ giải mã mới dựa trên mô hình khuếch tán để tạo ra âm thanh chất lượng cao. Tương tự như vậy, ProDiff (Huang và cộng sự, 2022) nghiên cứu về tham số hóa khuếch tán cho văn bản thành giọng nói và đạt được chất lượng mẫu và tính đa dạng cao hơn. Trong miền tạo văn bản, Diffusion-LM (Li et al., 2022) bắt đầu bằng một chuỗi nhiều Gaussian các vectơ và khử nhiễu chúng theo từng bước thành các vectơ tư duy ứng với các từ. Diffusion-LM cho phép phuơng pháp dựa trên gradient hiệu quả để kiểm soát thế hệ, đạt được những kết quả đầy hứa hẹn trong các dạng nhiệm vụ kiểm soát chi tiết phức tạp. Hơn nữa, (Gong et al., 2022; Strudel et al., 2022) mở rộng các mô hình khuếch tán vani để học có điều kiện.

Tuy nhiên, ít nghiên cứu áp dụng sự khuếch tán mô hình để chú thích hình ảnh, vì thách thức liên phuơng thức và tính rời rạc của văn bản. DD-Cap (Zhu et al., 2022b) thêm một nhánh mạng vào cu thể dự đoán tổng chiều dài và thiết kế của mã thông báo

một mô-đun mặt nạ tập trung sự chú ý để tập trung vào các mă thông tin nhiều hơn. Để tạo ra nhiều hơn chú thích cụ thể, (Kornblith et al., 2023) khám phá chiến lược hứa ứng dẫn mô hình chú thích hình ảnh bằng sửa đổi phân phối giải mã. Phân tán bit (Chen et al., 2022) cho phép các mô hình khuếch tán trạng thái liên tục tạo ra dữ liệu rời rạc bằng cách sử dụng bit tư tưởng tự và một hoạt động ngẫu nhiên đơn giản cho giải mã. Những phuơng pháp này (Xu et al., 2023; Tang et al., 2024) có thể được mô hình hóa trực tiếp bằng các mô hình khuếch tán trạng thái liên tục và sử dụng các tính năng của CLIP như một hứa ứng dẫn. Khác với các phuơng pháp hiện có, chúng tôi mở rộng dòng mô tả hình ảnh đa dạng bằng cách để xuất một mô hình khuếch tán liên tục nhẹ, điều này là cần thiết nhưng ít được quan tâm trước đây.

3. Phuơng pháp luận

Như minh họa trong Hình 2, chúng tôi đề xuất Tiền tố khuếch tán để đưa các đặc điểm hình ảnh vào để học hình ảnh chú thích. Khác với việc tạo hình ảnh, phuơng pháp này yêu cầu ánh xạ các văn bản rời rạc vào một không gian liên tục bằng cách nhúng từ. Đối với hình ảnh có điều kiện, trước tiên chúng tôi trích xuất các đặc điểm của nó bằng Bộ mã hóa hình ảnh CLIP, sau đó nhập chúng vào lập bản đồ mạng để có được các nhúng hình ảnh tiền tố. Sau đó, chúng tôi nối các hình ảnh tiền tố nhúng và nhúng chú thích trong

quá trình khử nhiễu của mô hình khuếch tán. Các vectơ nối tiếp được đưa vào mạng nơ-ron sâu (ví dụ BERT(Kenton và Toutanova, 2019) hoặc máy biến áp tiêu chuẩn). Vì công việc của chúng tôi chỉ đơn thuần đào tạo một mạng lưới lập bản đồ và một mạng lưới nơ-ron, thang tham số có thể đào tạo được giảm đáng kể.

Tiến trình chuyển tiếp. Theo sau Diffusion-LM (Li et al., 2022), chúng tôi áp dụng một hàm nhúng $EMB(W)$ để ánh xạ một từ rời rạc thành một từ liên tục không gian. Định nghĩa một chú thích W với k từ. Thông qua hàm nhúng, chúng ta có $EMB(W) = [EMB(\omega_1), \dots, EMB(\omega_k)] \in \mathbb{R}^{k \times d_1}$, nơi d_1 là chiều của vectơ. Trong các thí nghiệm của chúng tôi, chúng tôi thấy rằng giá trị của d_1 hoạt động tốt ở mức 48. Giảm kích thước sẽ làm giảm hiệu suất, trong khi tăng kích thước sẽ làm cho nó lớn hơn gánh nặng tính toán.

Đối với quá trình chuyển tiếp, các mô hình khuếch tán (Ho và cộng sự, 2020; Nichol và Dhariwal, 2021; Song và cộng sự, 2020) thêm tiếng ồn dần dần để đào tạo một mẫu theo một lịch trình phuơng sai β_1, \dots, β_T . Quá trình chuyển tiếp không có tham số có thể học được và chúng ta có x_t theo phuơng trình sau:

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon \quad (1)$$

trong đó $\epsilon \sim N(0, 1)$ và $\beta_t : 0.01 \rightarrow 0.03$ là các siêu tham số biểu diễn lịch trình phuơng sai qua các bước khuếch tán. Chúng tôi đã thử các nhiễu khác nhau phuơng pháp, với lịch trình tiếng ồn tuyển tính cắt bớt

Method	Common Metrics ↑							Similarity Score ↑			Diversity ↑			
	B@1	B@3	B@4	M	R-L	C	S	CLIP-S	Ref-CLIP	P-Bert	D@2	D@3	Voc-u	
LLaMA-Adapter	\	\	36.2	\	\	122.2	\	\	\	\	\	\	\	
VisionLLM	\	\	32.1	\	\	114.2	\	\	\	\	\	\	\	
BLIP2	\	\	43.7	\	\	145.8	\	\	\	\	\	\	\	
MTIC	80.8	50.9	39.1	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	39.2	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	26.6	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	31.1	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	30.8	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	31.8	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	<u>92.8</u>	15.5	25.2	1.3	
ClipCap	67.0	<u>35.2</u>	22.5	<u>49.0</u>	<u>60.8</u>	16.5	60.9	65.0	93.0	20.9	34.5	1.77	
Ours(T)	<u>68.7</u>	34.9	20.1	48.7	53.8	14.2	<u>61.6</u>	<u>66.3</u>	92.2	<u>23.1</u>	<u>41.0</u>	<u>3.6</u>	
Ours(B)	71.0	36.2	<u>21.1</u>	49.3	61.4	<u>15.2</u>	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.

relevant captions. The similarity is computed as follows:

$$\text{similarity}(I_{\text{img}}, W_{\text{txt}}^n) = \frac{I_{\text{img}} \cdot W_{\text{txt}}^n}{|I_{\text{img}}| \cdot |W_{\text{txt}}^n|} \quad (11)$$

where I_{img} is the image features extracted by CLIP and W_{txt}^n 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 caption-

ing. 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 score, we consider two metrics to evaluate the diversity of the generated captions.

- **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 judgments.

- **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.

Phương pháp	Số liệu chung							Điểm tương đồng					Sự đa dạng	
	B@1	B@3	B@4	M	R-L	C	S	CLIP-S	Tham khảo	CLIP	P-Bert	D@2	D@3	
Bộ chuyển đổi LLaMA	\	\	36,2	\	\	122,2	\	\	\	\	\	\	\	\
Tầm nhìn LLM	\	\	32,1	\	\	114,2	\	\	\	\	\	\	\	\
BLIP2	\	\	43,7	\	\	145,8	\	\	\	\	\	\	\	\
MTIC	80,8	50,9	39,1	29,2	58,6	131,2	22,6	60,3	81,1	51,1	39,2	29,4	58,9	133,1
DLCT	22,8	60,6												
Dòng lạnh	CapDec	68,3	36,6	26,6	25,2	51,2	91,7	18,3	60,4	ClipCap	73,6	42,3	31,1	26,7
Đoàn trinh		35,4	105,8	19,8	60,8	Của chúng tôi (T)	77,7	43,4	30,8	25,8	55,8	106,3	93,7	1,9
Tính năng		63,4	Của chúng tôi (B)	78,1	44,2	31,8	26,6	56,1	109,3	3,20,4	63,7		70,9	12,7
														28,0
														4,7
														5,7

Bảng 1: Kết quả chú thích hình ảnh trên COCO. Đối với tất cả các số liệu, càng cao càng tốt. Chúng tôi sử dụng in đậm để chỉ ra hiệu suất tốt nhất. Kết quả tốt thứ hai được gạch chân. Của chúng tôi(T) và Của chúng tôi(B) sử dụng một máy biến áp chuẩn và BERT tương ứng. Các giá trị sử dụng từ vựng được báo cáo tại phần trăm (%) .

Phương pháp	Số liệu chung							Điểm tương đồng					Sự đa dạng			
	B@1	B@3	M	R-L	C	S	CLIP-S	Tham khảo	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	<u>92,8</u>	58,0	61,4	92,8	15,5	25,2	1,3	
	35,2	<u>22,5</u>	49,0	<u>60,8</u>	<u>16,5</u>	60,9	Của chúng tôi (T)	68,7	34,9	65,0	93,0	20,9	34,5	1,77		
Ours(T)	20,1	48,7	53,8	14,2	61,6	66,3	Của chúng tôi (B)	71,0	36,2	21,1	66,3	92,2	23,1	<u>41,0</u>	3,6	
Ours(B)	49,3	61,4	15,2	64,7	68,6	92,0						68,6	92,0	27,6	46,0	4,0

Bảng 2: Kết quả chú thích hình ảnh trên Flickr30k. Đối với tất cả các số liệu, càng cao càng tốt. Chúng tôi sử dụng in đậm để chỉ hiệu suất tốt nhất. Kết quả tốt thứ hai được gạch chân.

chú thích có liên quan. Sự tương đồng được tính toán như sau đây:

$$\text{sự tương đồng}(I_{\text{img}}, W_{\text{txt}}) = \frac{I_{\text{img}} \cdot W_{\text{txt}}}{|I_{\text{img}}| \cdot |W_{\text{txt}}|} \quad (11)$$

trong đó I_{img} là các đặc điểm hình ảnh được trích xuất bởi CLIP và W_{txt} là các đặc điểm của n tiêu đề ứng viên. Đây là cơ sở truy xuất (Ramos et al., 2022;

Kỹ thuật (Zhao et al., 2020) chọn chú thích phù hợp nhất từ một tập hợp các chú thích ứng viên.

Chúng tôi sử dụng cách tiếp cận này dựa trên lợi thế của Tiêu tố-khuếch tán: mô hình của chúng tôi có thể tạo ra sự đa dạng chú thích với các tiếng ồn Gaussian khác nhau. Chúng tôi xác minh hiệu quả của phương pháp truy xuất cơ sở

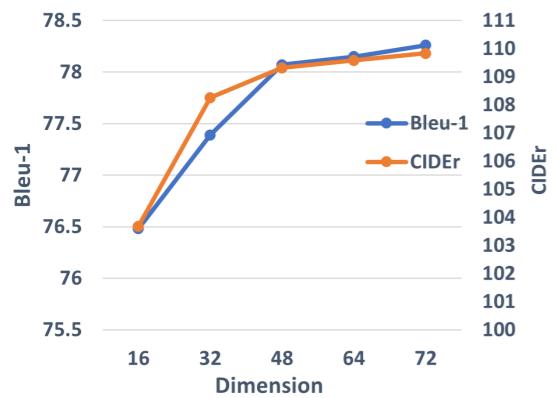


Figure 4: The performance effect of the word dimension on COCO. We report the metrics of Bleu-1 and CIDEr.

- 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:

LLM : In order to provide a thorough evaluation, we incorporate benchmarking against LLM, including LLaMA-Adapter(Zhang et al., 2023), VisionLLM(Wang et al., 2024) and BLIP2(Li et al., 2023).

MTIC (Cornia et al., 2020): MTIC 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.

In the experimental setup, the length of the text is set in advance as k due to the fact that our model is non-autoregressive. We choose $k = 24$ based on the specific characteristics of the Coco dataset. To denote the end of each sample, we use the symbol 'ENDS'. In cases where the length of a sample falls

Method	Human Evaluation↑			Parameters(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.

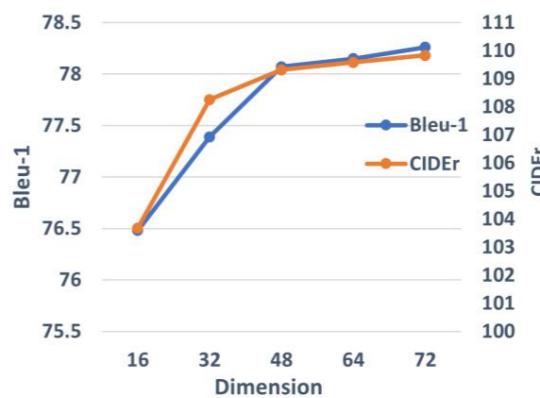
short, we utilize padding with the symbol 'PAD' to ensure consistency in the datasets.

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-diffusion contain diverse wordings and rich expressions. Our model can generate high-quality captions compared with captioning approaches that extract image feature with CLIP. Our method exhibits competitive performance on various aspects of the COCO dataset, yet it is crucial to acknowledge the limitations when contrasted with LLMs benefiting from extensive textual and visual training data. 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



Hình 4: Hiệu ứng hiệu suất của chiều từ trên COCO. Chúng tôi báo cáo các số liệu của Bleu-1 và CIDEr.

- Sử dụng từ vựng. Để phân tích sự đa dạng của các chủ thích được tạo ra, theo (Dai et al., 2018), chúng tôi tính toán cách sử dụng từ vựng (Voc-u), chiếm tỷ lệ phần trăm các từ trong từ vựng được sử dụng trong tạo ra chủ thích.

4.2. Đa ờng cơ sở

Chúng tôi áp dụng các phương pháp chú thích hình ảnh cạnh tranh trước đây để làm mô hình cơ sở:

LLM : Để cung cấp một đánh giá toàn diện, chúng tôi kết hợp chuẩn mực so với LLM, bao gồm LLaMA-Adapter(Zhang và cộng sự, 2023), VisionLLM (Wang và cộng sự, 2024) và BLIP2(Li và cộng sự, 2023).

MTIC (Cornia và cộng sự, 2020): MTIC là một kiến trúc dựa trên bộ chuyển đổi để chú thích hình ảnh. Hình ảnh của nó các đặc điểm được trích xuất bằng ResNet (được biểu thị là các đặc điểm dựa trên lối).

DLCT (Luo et al., 2021): DLCT đạt được tính bổ sung của các đặc điểm vùng và lối cho hình ảnh chủ thích. Để trích xuất các tính năng trực quan, DLCT sử dụng Faster-RCNN được đào tạo trước (Ren et al., 2015).

CapDec (Nukrai và cộng sự, 2022): CapDec là một phương pháp đơn giản và trực quan để học cách viết phụ đề nhiều cách diễn đạt khác nhau và cách diễn đạt phong phú. Mô hình của chúng tôi có thể tạo ra các chủ thích chất lượng cao so với

ClipCap (Mokady et al., 2021): ClipCap đơn giản - các mô hình được đào tạo trước về ngôn ngữ thi giác mạnh mẽ (CLIP) để đơn giản hóa quá trình chú thích. Và chúng tôi sử dụng mạng lối lập bản đồ MLP và tinh chỉnh mô hình ngôn ngữ. Tất cả các siêu tham số đều được thiết lập theo bản gốc của nó.

Trong thiết lập thử nghiệm, độ dài của văn bản là được đặt trước là k do thực tế là mô hình của chúng tôi là không hồi quy tự động. Chúng tôi chọn $k = 24$ dựa trên các đặc điểm cụ thể của tập dữ liệu Coco. Để để biểu thị phần cuối của mỗi mẫu, chúng ta sử dụng ký hiệu 'KẾT THÚC'. Trong trường hợp chiều dài của mẫu giảm

Phương pháp	Đánh giá của con người			Tham số(M)
	Sim	lưu	loát	
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

Bảng 3: Kết quả đánh giá của con người và số lượng tham số có thể đào tạo cho các phương pháp khác nhau.

Nói tóm lại, chúng tôi sử dụng phần mềm có ký hiệu 'PAD' để đảm bảo tính nhất quán trong các tập dữ liệu.

Vì CapDec và ClipCap sử dụng CLIP để trích xuất cùng một hình ảnh có tính năng và đóng băng CLIP như của chúng tôi mô hình, chúng tôi sử dụng các phương pháp này làm đường cơ sở chính. Chúng tôi đào tạo mô hình của mình trong 20000 bước, với kích thước lõi là 128. Kích thước lõi nhúng từ được đặt thành 48 và các bước khuếch tán $T = 1000$. Tất cả các thí nghiệm đều được chạy trên NVIDIA Tesla V100 GPU. Trong quá trình giải mã, chúng tôi cấu hình giá trị của các câu ứng cử viên có $n = 5$. Cụ thể, trong quá trình đánh giá, chúng tôi thiết lập các bước khử nhiễu $T = 50$, giúp giảm đáng kể thời gian thế hệ.

4.3. Kết quả

4.3.1. Chủ thích hình ảnh

Chúng tôi so sánh tiền tố khuếch tán với một số đường cơ sở với các số liệu đánh giá khác nhau, như được thể hiện trong

Bảng 1. Mô hình của chúng tôi vượt trội hơn tất cả các đường cơ sở trên Các số liệu CLIP-S và Ref-CLIP, và đạt được kết quả tương đương về điểm P-Bert, cho thấy rằng hiệu quả của sự khuếch tán liên tục trên hình ảnh chủ thích. Không chỉ vậy, chúng tôi có sự cải thiện đáng kể về một số số liệu đa dạng (chẳng hạn như D@2 và D@3). Hơn nữa, sự khuếch tán tiền tố bao phủ phần trăm lớn nhất của các từ, được quan sát từ từ vựng được sử dụng để tạo chủ thích. Nó lưu ý

rằng các chủ thích được tạo ra bởi Prefix-diffusion chứa các đặc điểm riêng và lối cho hình ảnh chủ thích. Để trích xuất các tính năng trực quan, DLCT sử dụng Faster-RCNN được đào tạo trước (Ren et al., 2015).

CapDec (Nukrai và cộng sự, 2022): CapDec là một phương pháp đơn giản và trực quan để học cách viết phụ đề nhiều cách diễn đạt khác nhau và cách diễn đạt phong phú. Mô hình của chúng tôi có thể tạo ra các chủ thích chất lượng cao so với

phương pháp chú thích trích xuất đặc điểm hình ảnh với CLIP. Phương pháp của chúng tôi thể hiện hiệu suất cạnh tranh trên nhiều khía cạnh của tập dữ liệu COCO, tuy nhiên điều quan trọng là phải thừa nhận những hạn chế khi trái ngược với LLM được hưởng lợi từ dữ liệu đào tạo trực quan và văn bản mở rộng. Tiền tố khuếch tán thực hiện tệ hơn MTIC và DLCT (những người không sử dụng đóng băng các tính năng để chủ thích hình ảnh) trên các số liệu chung, một phần là do những hạn chế đã được chứng minh của các số liệu dựa trên sự chồng chéo từ trên nhiều tên miền(Hessel và cộng sự, 2021; Zhang và cộng sự, 2020), và cũng bởi vì thế hệ của chúng ta đa dạng hơn

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
<i>COCO</i> → <i>Flickr30k</i>												
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
<i>Flickr30k</i> → <i>COCO</i>												
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.

in expression and correctly describe the visual content, which can be observed from similarity score and diversity metrics.

We also conduct experiments on Flickr30k dataset, as presented in Table 2, from which we can draw similar conclusions with the COCO dataset. 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. In the process of gradually predicting noise, we speculate that the diffusion model introduces small perturbations, resulting in different directions due to the removal of noise over time. This achieves the goal of text diversity.

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. The diversity in generated text expands the model’s application scope, making it more widely applicable across various fields and industries.

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 num-

ber of model parameters. It brings potential advantages of saving memory storage space and computing costs, and thus being much more useful in practice. Unlike the slow generation speed commonly observed in image generation, our non-autoregressive approach enables parallel generation of all tokens instead of a sequential token-by-token generation method. Consequently, our method exhibits faster generation speeds. 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.

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

Phương pháp	Số liệu chung						Điểm tự đồng				Sự đa dạng	
	B@1	B@3	M	R-L	CS	CLIP-S	Tham khảo	CLIP-S	P-Bert	D@2	D@3	Voc-u
<i>COCO</i> = <i>Flickr30k</i>												
CapDec	57,2	23,9	17,1	40,3	30,3	10,8	54,4	58,7	92,1	18,5	29,4	92,5
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
<i>Flickr30k</i> = <i>COCO</i>												
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	90,4
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

Bảng 4: Kết quả của chủ thích xuyên miền. *COCO*=*Flickr30k* có nghĩa là mô hình được đào tạo trên COCO trong khi được đánh giá trên Flickr30k, và *Flickr30k*=*COCO* cũng vậy. Chúng tôi sử dụng chữ in đậm để chỉ hiệu suất tốt nhất.

trong cách diễn đạt và mô tả chính xác nội dung trực quan, có thể quan sát được từ điểm tự đồng và số liệu về sự đa dạng.

Chúng tôi cũng tiến hành thử nghiệm trên Flickr30k

tập dữ liệu, như đã trình bày trong Bảng 2, từ đó chúng ta có thể rút ra kết luận tương tự với tập dữ liệu COCO.

Mô hình của chúng tôi đạt được hiệu suất án tượng trong nhiệm vụ chủ thích hình ảnh so với các mô hình cơ sở. Chi tiết, từ kết quả của sự đa dạng

số liệu, chúng tôi nhận thấy rằng số liệu của Dist-3 và tăng cường sử dụng vốn từ vựng hơn 6.0 và 3.0, tương ứng. Ngoài ra, chúng tôi cũng quan sát một cải thiện 2.6 và 2.8 trong số liệu CLIP-S và Ref-CLIP,

tương ứng. Điều này chỉ ra rằng mô hình khuếch tán có thể cải thiện chủ thích một cách hiệu quả

sự đa dạng trong khi vẫn đảm bảo tính thống nhất và liên quan

trong các chủ thích được tạo ra. Để tạo ra sự đa dạng chủ thích, các phương pháp hiện có có xu hướng tạo ra các chủ thích khác nhau thông qua lấy mẫu top-k. Theo trực giác, chẳng hạn các phương pháp có thể bỏ qua sự đa dạng của phông và sự đa dạng ngữ nghĩa mà con người thực sự quan tâm.

Không giống như các phương pháp hiện có, tiền tố khuếch tán tìm cách tạo nhiều chủ thích với biểu thức phong phú

từ các tiếng ồn Gaussian khác nhau. Trong quá trình dần dần dự đoán tiếng ồn, chúng tôi suy đoán rằng

mô hình khuếch tán đưa ra những nhiễu loạn nhỏ, dẫn đến các huy ứng khác nhau do việc loại bỏ

của tiếng ồn theo thời gian. Điều này đạt được mục tiêu của văn bản đa dạng.

Hình 1 cho thấy các chủ thích được tạo ra bởi Prefix-diffusion. Người ta quan sát thấy rằng các chủ thích được tạo ra

khá phù hợp với hình ảnh cũng như giữ được sự trôi chảy có trình độ. Trong khi đó, mô hình của chúng tôi là

có thể tạo ra nhiều chủ thích đa dạng giống như do con người tạo ra. Sự đa dạng trong văn bản được tạo ra

của các tham số mô hình. Nó mang lại lợi thế tiềm năng trong việc tiết kiệm không gian lưu trữ bộ nhớ và chi phí tính toán, và do đó hữu ích hơn nhiều trong thực tế.

Không giống như tốc độ tạo ra chậm

thường được quan sát thấy trong quá trình tạo hình ảnh, phương pháp tiếp cận không hồi quy tự động của chúng tôi

cho phép tạo song song tất cả các mã thông báo thay vì

phương pháp tạo mã thông báo theo từng mã thông báo tuần tự. Do đó, phương pháp này thể hiện tốc độ tạo ra nhanh hơn. Đối với đánh giá của con người, chúng tôi đã chọn ngẫu nhiên 20 mẫu

và trình bày chúng theo cách xáo trộn cho 20

người chủ thích. Người chủ thích đánh giá mức độ lưu loát, tính tự đồng (Sim) và tính đa dạng (Div) của các chủ thích trên

thang điểm từ 1 đến 5, với điểm số cao hơn cho biết chất lượng tốt hơn. Từ kết quả đánh giá của con người,

Chúng ta có thể rút ra những kết luận tương tự với sự đa dạng

Kích thước của những từ là một siêu tham số quan trọng. Kích thước cao hơn

dẫn đến nhiều thời gian đào tạo và sử dụng bộ nhớ hơn. Để nghiên cứu sâu hơn về tác động của việc nhúng chiều trong khuếch tán tiền tố, chúng tôi tiến hành các thí nghiệm

bằng cách đào tạo với các kích thước khác nhau. Như được hiển thị trong Hình 4, các số liệu của Bleu-1 và CIDEr là

được cải thiện khi kích thước nhúng tăng lên.

Lý do là một từ nhúng trở thành giàu thông tin ngữ nghĩa hơn do cao hơn

kích thước. Tuy nhiên, có một nút thắt hiệu suất khi chúng ta tiếp tục tăng kích thước

của nhúng từ. Người ta quan sát thấy xu hướng hiệu suất ổn định khi kích thước

vượt quá 48.

4.3.2. Chủ thích xuyên miền

Chúng tôi cũng tiến hành các thí nghiệm về chủ thích xuyên để đánh giá khả năng khai thác của

Tiền tố-khuếch tán. Kết quả của miền chéo

đánh giá được thể hiện trong Bảng 4. Chúng tôi đào tạo mô hình trên tập dữ liệu của một miền nguồn trong khi đánh giá nó trên một tập dữ liệu khác. Từ kết quả của

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.

COCO \Rightarrow 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. We hypothesize that this is due to the fact that diffusion models can effectively capture key features in text and learn the distribution patterns of textual data.

4.3.3. Ablation

We perform the ablation study on the COCO dataset 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. Because of the remarkable results of the diffusion model on image editing, we will continue to investigate how image features guide step-by-step text generation, thereby enabling controlled image captioning. Additionally, an exploration of how different types of noise affect the model’s output would be a valuable and interesting topic for further research.

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. We speculate that in the process of gradually predicting noise, the diffusion model results in the diversity of the generated text due to small perturbations. 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. Because of the remarkable results of the diffusion model on image editing, we will continue to investigate how image features guide step-by-step text generation, thereby enabling controlled image captioning. Additionally, an exploration of how different types of noise affect the model’s output would be a valuable and interesting topic for further research.

N	Số liệu chung					Điểm tương đồng				Sự đa dạng	
	B@1	B@3	M	R-L	C	CLIP-S	Tham khảo-CLIP	P-Bert	D@2	D@3	Voc-u
1	77,2	43,6	26,0	55,6	105,2	19,5	60,4	5	78,1	44,2	26,6
5	56,1	109,3	20,4	63,7	10	78,3	43,8	26,6	56,0	109,1	20,3
10	65,3	15	78,2	43,4	26,5	55,8	108,5	20,3	66,0	72,2	28,8
15	72,6	63,7	71,2	93,7	66,0	63,7	71,2	93,4	13,4	29,3	5,9

Bảng 5: Tác động của các giá trị khác nhau của chú thích ứng viên. $n = 1$ có nghĩa là không có tính toán độ tương đồng cosin trong quá trình giải mã.

Lịch trình	Tiếng ồn			
	B@1	CLIP-S	Tham khảo-CLIP	P-Bert
Cosin tuyến	70,5	66,8	72,2	92,6
tính vuông	70,4	65,9	71,6	92,3
T-Tuyến	66,5	72,0	92,5	
Cosin T-Tuyến	66,5	72,3	92,9	
tính	63,7	71,2	93,7	

Bảng 6: Phân tích các lịch trình tiếng ồn khác nhau trong quá trình chuyển tiếp. T-Linear và T-Cosine có nghĩa là lịch trình tiếng ồn tuyến tính cắt cụt và lịch trình tiếng ồn cosin cắt cụt tương ứng.

COCO= Flickr30k, Tiền tố khuếch tán đạt được hiệu suất tuyệt vời trên tất cả các phương pháp được so sánh, với kết quả trên các số liệu chung là tốt nhất. Ngoài ra, nó còn có được những cải tiến đáng kể về cả số liệu sử dụng từ vựng và Dist-3. Điều này là do khả năng sinh sản mạnh mẽ của mô hình khuếch tán. Khi chúng ta đào tạo trên flickr30k trong khi đánh giá trên COCO, kết quả cũng cho thấy rằng cách tiếp cận có khả năng mạnh mẽ trong lĩnh vực chéo kịch bản. Bằng cách so sánh hai kết quả, chúng tôi thấy rằng Tiền tố khuếch tán hoạt động thậm chí còn tốt hơn khi được đào tạo trên một tập dữ liệu lớn hơn, ngay cả khi khai thác tốt hơn khả năng. Chúng tôi đưa ra giả thuyết rằng điều này là do thực tế rằng các mô hình khuếch tán có thể nắm bắt hiệu quả các chìa khóa các tính năng trong văn bản và tìm hiểu các mẫu phân phối của dữ liệu văn bản.

4.3.3. Phá hủy

Chúng tôi thực hiện nghiên cứu cắt bỏ trên COCO tập dữ liệu để định lượng sự đóng góp của từng mô-đun trong tiền tố khuếch tán.

Bảng 5 trình bày tác động đến số lượng chú thích ứng viên. Từ hai nhóm thí nghiệm, $n = 1$ và $n = 5$, có thể thấy rằng chiến lược lựa chọn cải thiện hiệu suất của chú thích hình ảnh. Chúng tôi quan sát thấy sự gia tăng đáng kể trong số liệu CIDEr, giúp tăng điểm CIDEr từ 105,2 đến 109,3. Nó xác nhận chức năng tính toán độ tương đồng giữa hình ảnh và ứng cử viên chú thích và chọn cao nhất. Như những chú thích ứng viên đã dẫn đến việc giảm

hiệu suất của độ trôi chảy của phụ đề. Điều này là do chúng tôi sử dụng điểm CLIP làm điểm tương đồng duy nhất thư ờc do lựa chọn, có thể bỏ qua sự lưu loát của chú thích.

Như được trình bày trong Bảng 6, chúng tôi điều tra hiệu suất của các lịch trình tiếng ồn khác nhau. Quan sát kết quả, chúng tôi kết luận rằng tiếng ồn tuyến tính bị cắt cụt lịch trình có thể tạo ra các chú thích chính xác và mô tả hơn. Chúng tôi cũng kết luận rằng thông tin ngữ nghĩa bị làm hỏng bởi sự phức tạp lịch trình tiếng ồn trong quá trình chuyển tiếp, dẫn đến một vấn đề học tập khó khăn hơn trong việc khử nhiễu quá trình.

5. Kết luận và công việc tương lai

Trong bài báo này, chúng tôi đề xuất một mạng lưới ánh kinh kết hợp với khuếch tán liên tục, được gọi là khuếch tán tiền tố. Thí nghiệm và phân tích sâu hơn chứng minh rằng nó có thể tạo ra nhiều phụ đề khác nhau trong khi vẫn duy trì tính lưu loát và tính liên quan của phụ đề. Bằng cách đào tạo trên một tập dữ liệu như được đánh giá trên tập dữ liệu khác, Tiền tố khuếch tán thể hiện khả năng khai thác hóa đáng chú ý. Bên cạnh đó, mô hình của chúng tôi yêu cầu một số lưu ý trong cách áp dụng nhiều hơn trong thực tế. Chúng tôi cũng tiến hành các thí nghiệm cắt bỏ để hiển thị hiệu ứng của chiến lược lựa chọn và tiếng ồn lịch trình. Chúng tôi suy đoán rằng trong quá trình dần dần dự đoán tiếng ồn, mô hình khuếch tán dẫn đến sự đa dạng của văn bản được tạo ra do nhiễu loạn nhỏ. Kết quả thực nghiệm xác minh rằng tiền tố khuếch tán có khả năng sinh sản mạnh mẽ để chủ thích cho hình ảnh.

Đối với công việc tương lai, chúng tôi sẽ tiếp tục khám phá tác động tiềm tàng của các mô hình khuếch tán lên chủ thích hình ảnh. Bởi vì những kết quả đáng chú ý của mô hình khuếch tán trên chỉnh sửa hình ảnh, chúng tôi sẽ tiếp tục để điều tra cách các tính năng hình ảnh hưng dẫn tạo văn bản từng bước, do đó cho phép kiểm soát chủ thích hình ảnh. Ngoài ra, một cuộc khám phá các loại tiếng ồn khác nhau ảnh hưởng đến đầu ra của mô hình sẽ là một chủ đề có giá trị và thú vị cho nghiên cứu sâu hơn.

6. 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.

7. 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.

8. Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 62106037, No. 62076052), in part by the Major Program of the National Social Science Foundation of China (No.19ZDA127), and in part by the Fundamental Research Funds for the Central Universities (No. DUT22YG205).

9. Bibliographical 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*.
- Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. 2022. *Analog bits: Generating discrete data using diffusion models with self-conditioning*. *arXiv preprint arXiv:2208.04202*.
- 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*.

6. Hạn chế

Như được trình bày trong Bảng 1 và Bảng 2, mặc dù khuếch tán tiền tố có thể tạo ra nhiều chủ thích khác nhau với ít tham số hơn như nó kém hơn MTIC và DLCT trên các số liệu chung. Nhưng nó hoạt động tốt trên các số liệu mới hơn đã được hiển thị cao hơn mối tương quan với thế hệ con người. Lý do là rằng các chủ thích do chúng tôi tạo ra có cách diễn đạt phong phú điều đó không nhất quán với văn bản thám khảo, nhưng vẫn truyền đạt cùng một ngữ nghĩa cơ bản. Chiều dài là một thuộc tính quan trọng vì nó phản ánh số lượng thông tin được mang theo bởi một chủ thích. Vì mô hình của chúng tôi là một mô hình không tự hồi quy, chúng ta không thể kiểm soát độ dài của văn bản được tạo ra, dẫn đến ít hơn mô tả chính xác về hình ảnh.

7. Tuyên bố về đạo đức

Vì tiền tố khuếch tán được đề xuất có thể được sử dụng để tạo phụ đề. Với những lợi thế của việc chính xác, đa dạng và mô tả, thế hệ của nó là giống như do con người tạo ra hơn. Điều này sẽ có lợi cho các ứng dụng chủ thích hình ảnh trên các tác vụ hạ nguồn, chẳng hạn như robot trò chuyện và hứa hẹn dẫn bằng giọng nói tự động hệ thống. Một khác, số lượng lớn chủ thích hình ảnh sẽ làm cho việc phân biệt trở nên khó khăn do con người viết từ máy tạo ra. Do đó, việc khám phá các cuộc tấn công đôi đầu vào chủ thích hình ảnh là cần thiết. Hơn nữa, phụ đề tuyệt vời nên bao gồm nhiều từ ngữ và cách diễn đạt phong phú, giúp chúng không trở nên quá nhàm chán hoặc tệ hại. Mô hình khuếch tán tạo ra các mẫu mới từ tiếng ồn khác nhau. Do đó, tiền tố khuếch tán có thể được sử dụng để cải thiện tính đa dạng của chủ thích.

8. Lời cảm ơn

Công trình này được hỗ trợ một phần bởi Quỹ Khoa học Tự nhiên Quốc gia Trung Quốc (Số 62106037, Số 62076052), một phần của Chương trình chính của Quỹ Khoa học xã hội quốc gia Trung Quốc (Số 19ZDA127), và một phần của Cơ bản Quỹ nghiên cứu cho các trường đại học trung ương (Số DUT22YG205).

9. Tài liệu tham khảo

- Peter Anderson, Basura Fernando, Mark Johnson, và Stephen Gould. 2016. Spice: Ngữ nghĩa đánh giá chủ thích hình ảnh theo đề xuất. Tại hội nghị Châu Âu về thị giác máy tính, các trang 382–398. Mùa xuân.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, và Lei Zhang. 2018. Sự chú ý từ dưới lên và từ trên xuống đối với chủ thích hình ảnh và câu hỏi trực quan trả lời. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 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: Mô hình khuếch tán văn bản thành hình ảnh với một nhóm các bộ khử nhiễu chuyên gia*. *arXiv* bản in trước *arXiv:2211.01324*.
- Ting Chen, Ruixiang Zhang và Geoffrey Hinton. 2022. *Bit tương tự: Tạo dữ liệu rời rạc bằng cách sử dụng mô hình khuếch tán có tự điều kiện*. *arXiv* bản in trước *arXiv:2208.04202*.
- Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, và Rita Cucchiara. 2020. *Bộ nhớ dạng lư ơng máy biến áp để thêm chủ thích cho hình ảnh*. Trong Biên bản của hội nghị IEEE/CVF về tầm nhìn máy tính và nhận dạng mẫu, trang 10578–10587.
- Bo Dai, Sanja Fidler và Dahua Lin. 2018. *Một mô hình bối cảnh trung gian cho chủ thích hình ảnh*. Nhũng tiến bộ trong hệ thống xử lý thông tin thần kinh , 31:656–666.
- Chaorui Deng, Ning Ding, Mingkui Tan và Qi Wu. 2020. *Chủ thích hình ảnh có thể kiểm soát độ dài*. Trong Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 712–729. Springer.
- Michael Denkowski và Alon Lavie. 2014. *Sao băng phổ biến: Đánh giá bản dịch theo ngôn ngữ cụ thể cho bất kỳ ngôn ngữ đích nào*. Trong Biên bản của hội thảo thử chín về dịch máy thông kê , trang 376–380.
- Trịnh Công Phi. 2020. *Sửa đổi ngữ cảnh lặp lại để thêm chủ thích cho hình ảnh nhanh hơn*. Trong Biên bản của Hội nghị quốc tế ACM lần thứ 28 về đa phương tiện, trang 3182–3190.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Hoặc Patashnik, Amit H Bermano, Gal Chechik và Daniel Cohen-Or. 2022. *Một hình ảnh có giá trị bằng một từ: Cá nhân hóa việc tạo văn bản thành hình ảnh bằng cách đảo ngữ của văn bản*. bản in trước *arXiv arXiv:2208.01618*.
- Federico A Galatolo, Mario GCA Cimino và Gigliola Vaglini. 2021. *Tạo hình ảnh từ chủ thích và ngữ cảnh lại thông qua tìm kiếm không gian tiềm ẩn tạo ra được hứa hẹn dẫn bằng clip*. bản in trước *arXiv arXiv:2102.01645*.
- Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma và Wen Gao. 2019. *Chủ thích hình ảnh không tự động hồi quy có che dấu*. bản in trước *arXiv 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.
 Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. 2023. *Guiding image captioning models toward more specific captions*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15259–15269.
 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.
 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*. In *International conference on machine learning*, pages 19730–19742. PMLR.
 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.
 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 SIGDIAL 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.
- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu và LingPeng Kong. 2022. *Diffuseq: Tạo văn bản theo trình tự với các mô hình khuếch tán*.
 Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville và Yoshua Bengio. 2020. *Mạng đối kháng sinh sản*. *Truyền thông của ACM*, 63(11):139–144.
 Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, và Baining Guo. 2022. *Mô hình khuếch tán lượng tử hóa vector để tổng hợp văn bản thành hình ảnh*. Trong Biên bản của Hội nghị IEEE/CVF về *Tầm nhìn máy tính* và Nhận dạng Mẫu, trang 10696–10706.
 Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian Anh ấy, Jie Jiang và Hanqing Lu. 2020. *Chú thích hình ảnh không tự động hồi quy với học tập đa tác nhân mang tính phản biện - phản biện*. *bản in tru ớc arXiv arXiv:2005.04690*.
 Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, và Yejin Choi. 2021. *Clipscore: A thư ớc đo đánh giá không tham chiếu cho chú thích hình ảnh*. Trong Biên bản Hội nghị năm 2021 về *Phương pháp thực nghiệm trong Xử lý ngôn ngữ tự nhiên*, trang 7514–7528.
 Jonathan Ho, Ajay Jain và Pieter Abbeel. 2020. *Mô hình xác suất khuếch tán khứ nhiều*. *Những tiến bộ trong hệ thống xử lý thông tin thần kinh*, 33:6840–6851.
 Rongjie Huang, Chu Zhao, Huadai Liu, Jinglin Liu, Chenye Cui và Yi Ren. 2022. *Prodiff: Mô hình khuếch tán nhanh triển cho chất lượng cao chuyển văn bản thành giọng nói*. Trong Biên bản của ACM lần thứ 30 Hội nghị quốc tế về *đa phương tiện*, trang 2595–2605.
 Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Tú ớng Luo, Wu Yongjian, Yue Gao và Rongrong Ji. 2021. *Cải thiện chú thích hình ảnh bằng cách tận dụng biểu diễn toàn cầu trong và giữa các lớp trong mạng lưới máy biến áp*. Trong Biên bản của AAAI hội nghị về *trí tuệ nhân tạo*, tập 35, trang 1655–1663.
 Andrej Karpathy và Li Fei-Fei. 2015. *Sự liên kết ngữ nghĩa a trực quan sâu sắc để tạo ra mô tả hình ảnh*. Trong Biên bản báo cáo của hội nghị IEEE về *thi giác máy tính và nhận dạng mẫu*, trang 3128–3137.
 Jacob Devlin Ming-Wei Chang Kenton và Lee Kristina Toutanova. 2019. *Bert: Đào tạo tru ớc của máy biến áp hai chiều sâu cho ngôn ngữ sự hiểu biết*. Trong Biên bản của NAACL-HLT, trang 4171–4186.
 Gwanghyun Kim, Taesung Kwon và Jong Chul Ye. 2022. *Mô hình khuếch tán có hứ ứng dẫn bằng văn bản để chỉnh sửa hình ảnh hiệu quả*. Trong Biên bản của Hội nghị IEEE/CVF về *Tầm nhìn máy tính* và Nhận dạng Mẫu, trang 2426–2435.
 Simon Kornblith, Lala Li, Zirui Wang và Thảo Nguyên. 2023. *Hứ ứng dẫn mô hình chủ thích hình ảnh hứ ứng tới những chủ thích cụ thể hơn*. Trong Biên bản của Hội nghị quốc tế IEEE/CVF về *Tầm nhìn máy tính*, trang 15259–15269.
 Quảng Lý, Lâm Siêu Chu, Bình Lưu, Dịch Dương. 2019. *Biến áp rối dùng để chú thích hình ảnh*. TRONG Biên bản hội nghị quốc tế IEEE/CVF về *thi giác máy tính*, trang 8928–8937.
 Jiwei Li, Michel Galley, Chris Brockett, Kiến Phong Gao và Bill Dolan. 2016. *Một sự thúc đẩy đa dạng hàm mục tiêu cho mô hình hội thoại thần kinh*. Trong Biên bản báo cáo của NAACL-HLT, trang 110–119.
 Junnan Li, Dongxu Li, Silvio Savarese và Steven Xin chào. 2023. *Blip-2: Khởi động ngôn ngữ - đào tạo tru ớc hình ảnh với bộ mã hóa hình ảnh đóng băng và các mô hình ngôn ngữ lớn*. Trong quốc tế hội nghị về *máy học*, trang 19730–19742. PMLR.
 Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang và Tatsunori B Hashimoto. 2022. *Diffusion-lm cải thiện khả năng tạo văn bản có thể kiểm soát đư ợc*. *bản in tru ớc arXiv arXiv:2205.14217*.
 Chin-Yew Lin và Franz Josef Och. 2004. *Đánh giá tự động chất lượng dịch máy*. *sử dụng dãy con chung dài nhất và thống kê bỏ qua bigram*. Trong Biên bản của lần thứ 42 Cuộc họp thường niên của Hiệp hội Ngôn ngữ học tính toán (ACL-04), trang 605–612.
 Thánh Lovenia, Bryan Wilie, Romain Barraud, Samuel Cahyawijaya, Willy Chung và Pascale Fung. 2022. *Mỗi bức tranh kể một câu chuyện: Tao ra câu chuyện theo phong cách có thể kiểm soát dựa trên hình ảnh*. Trong Biên bản Hội thảo chung SIGDIAL lần thứ 6 về Ngôn ngữ học tính toán cho Văn hóa Di sản, Khoa học xã hội, Nhân văn và Văn học, trang 40–52.
 J. Lu, J. Yang, D. Batra và D. Parikh. 2018. *Nói chuyện một cách trung lập*. Trong Biên bản báo cáo của hội nghị IEEE về *thi giác máy tính và nhận dạng mẫu*, trang 7219–7228.
 Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, và Rongrong Ji. 2021. *Chuyển đổi cộng tác hai cấp để thêm chú thích cho hình ảnh*. Trong Biên bản của Hội nghị AAAI về *Trí tuệ nhân tạo*, tập 35, trang 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.
- 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*.
- Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2024. *Any-to-any generation via composable diffusion*. Advances in Neural Information Processing Systems, 36.
- 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.
- Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. *Visionllm: Large language model is also an open-ended decoder for vision-centric tasks*. Advances in Neural Information Processing Systems, 36.
- Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. 2023. *Versatile diffusion: Text, images and variations all in one diffusion model*. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7754–7765.
- 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.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. *Llama-adapter: Efficient fine-tuning of language models with zero-init attention*. arXiv preprint arXiv:2303.16199.
- Ron Mokady, Amir Hertz và Amit H Bermano. 2021. *Clipcap: Tiền tố clip để thêm chú thích cho hình ảnh*. bản in tru ớc arXiv arXiv:2111.09734.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever và Mark Chen. 2021. *Glide: Hướng tới thế hệ hình ảnh chân thực và chính sửa bằng mô hình khuếch tán có hướng dẫn bằng văn bản*. bản in tru ớc arXiv arXiv:2112.10741.
- Alexander Quinn Nichol và Prafulla Dhariwal. 2021. *Cải thiện xác suất khuếch tán khử nhiễu mô hình*. Trong Hội nghị quốc tế về máy học, trang 8162–8171. PMLR.
- David Nukrai, Ron Mokady và Amir Globerson. 2022. *Đào tạo chỉ có văn bản để thêm chú thích cho hình ảnh bằng cách sử dụng clip có chèn nhiễu*. bản in tru ớc arXiv arXiv:2211.00575.
- Kishore Papineni, Salim Roukos, Todd Ward và Wei-Jing Zhu. 2002. *Bleu: một phương pháp đánh giá tự động bản dịch máy*. Trong Biên bản cuộc họp thường niên lần thứ 40 của Hiệp hội Ngôn ngữ học tính toán, trang 311–318.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. *Học tập có thể chuyển giao mô hình trực quan từ sự giám sát ngôn ngữ tự nhiên*. Trong Hội nghị quốc tế về học máy, trang 8748–8763. PMLR.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever và cộng sự. 2019. *Mô hình ngôn ngữ là đa nhiệm không giám sát ngược*. Blog OpenAI, 1(8):9.
- Rita Ramos, Bruno Martins, Desmond Elliott và Yova Kementchedjhieva. 2022. *Vốn hóa nhỏ: Chú thích hình ảnh nhẹ được nhắc đến bằng cách tăng cường truy xuất*. Trong Biên bản Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 2840–2849.
- Shaoqing Ren, Kaiming He, Ross Girshick và Jian Sun. 2015. *Faster r-cnn: Hướng tới thời gian thực phát hiện đối tượng bằng mạng đẽo xuất vùng*. Nhữn tiến bộ trong hệ thống xử lý thông tin thần kinh, 28.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser và Björn Ommer. 2022. *Tổng hợp hình ảnh có độ phân giải cao với mô hình khuếch tán tiềm ẩn*. Trong Biên bản Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 10684–10695.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan và Surya Ganguli. 2015. *Sâu học không giám sát sử dụng nhiệt động lực học không cân bằng*. Trong Hội nghị quốc tế về máy học, trang 2256–2265. PMLR.
- Bài hát Gia Minh, Chenlin Meng và Stefano Ermon. 2020. *Mô hình ản khuếch tán khử nhiễu*. Trong Hội nghị quốc tế về Biểu diễn học tập.
- Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre và Rémi Leblond. 2022. *Khuếch tán nhúng tự điều kiện để tạo văn bản*.
- Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng và Mohit Bansal. 2024. *Thể hệ bất kỳ tới bất kỳ thể hệ nào thông qua sự phô biến có thể kết hợp*. Nhữn tiến bộ trong hệ thống xử lý thông tin thần kinh, 36.
- Ramakrishna Vedantam, C Lawrence Zitnick và Devi Parikh. 2015. *Cider: Đánh giá mô tả hình ảnh dựa trên sự đồng thuận*. Trong Biên bản hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio và Dumitru Erhan. 2015. *Hiển thị và kể: Một hệ thần kinh trình tạo chú thích hình ảnh*. Trong Biên bản hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 3156–3164.
- Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. *Visionllm: Mô hình ngôn ngữ lớn cũng là một bộ giải mã mở cho các nhiệm vụ lấy tầm nhìn làm trung tâm*. Nhữn tiến bộ trong hệ thống xử lý thông tin thần kinh, 36.
- Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang và Humphrey Shi. 2023. *Khuếch tán đa năng : Văn bản, hình ảnh và các biến thể đều có trong một mô hình khuếch tán*. Trong Biên bản Hội nghị quốc tế IEEE/CVF về Thị giác máy tính, trang 7754–7765.
- Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou và Dong Yu. 2022. *Diffsound: Mô hình khuếch tán rời rạc để tạo văn bản thành âm thanh*. bản in tru ớc arXiv arXiv:2207.09983.
- Renrui Zhang, Jiaming Han, Aojun Chu, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao và Yu Qiao. 2023. *Llama-adapter: Tinh chỉnh hiệu quả các mô hình ngôn ngữ mà không cần chú ý đến lần khởi tạo*. bản in tru ớc arXiv arXiv:2303.16199.

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.

Luwei 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. 2022a. *Visualize before you write: Imagination-guided open-ended text generation*. *arXiv preprint arXiv:2210.03765*.

Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. 2022b. *Exploring discrete diffusion models for image captioning*. *arXiv preprint arXiv:2211.11694*.

10. Language Resource References

Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C Lawrence. 2014. *Microsoft coco: Common objects in context*. Springer.

Plummer, Bryan A and Wang, Liwei and Cervantes, Chris M and Caicedo, Juan C and Hockenmaier, Julia and Lazebnik, Svetlana. 2015. *Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger và Yoav Artzi. 2020. *Điểm Bert: Đánh giá việc tạo văn bản bằng bert*. Trong Hội nghị quốc tế về biểu diễn học tập.

Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang và Jiaxuan Zhang. 2020. *Chú thích hình ảnh tạo ra thông qua phương pháp truy xuất thống nhất và dựa trên thế hệ*. Khoa học ứng dụng, 10(18):6235.

Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso và Jianfeng Gao. 2020. *Đào tạo trứ ớc ngôn ngữ thị giác thống nhất* cho chú thích hình ảnh và vqa. Trong Biên bản Hội nghị về Trí tuệ nhân tạo AAAI , tập 34, trang 13041-13049.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein và William Yang Wang. 2022a. *Hình dung trứ ớc khi viết: Tạo văn bản mở theo hướng tư ớng tư ợng*. bản in trứ ớc arXiv arXiv:2210.03765.

Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zichen Liu và Han Hu. 2022b. *Khám phá các mô hình khuếch tán riêng biệt cho chú thích hình ảnh*. bản in trứ ớc arXiv arXiv:2211.11694.

10. Tài liệu tham khảo về ngôn ngữ

Lin, Tsung-Yi và Maire, Michael và Belongie, Serge và Hays, James và Perona, Pietro và Ramanan, Deva và Dollár, Piotr và Zitnick, C Lawrence. 2014. *Microsoft coco: Common các đối tượng trong ngữ cảnh*. Mùa xuân.

Plummer, Bryan A và Wang, Liwei và Cervantes, Chris M và Caicedo, Juan C và Hockenmaier, Julia và Lazebnik, Svetlana. 2015. *Flickr30k thực thể: Thu thập các từ ớng ứng giữa vùng và cụm từ để có các mô hình hình ảnh và câu phong phú hơn*.