

Controllable Image Captioning via Prompting

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, Linlin Li

Huawei Inc.

wn6149@mail.ustc.edu.cn, jh_xie@tongji.edu.cn, {wujihao, jiamingbo, lynn.lilinlin}@huawei.com

Abstract

Despite the remarkable progress of image captioning, existing captioners typically lack the controllable capability to generate desired image captions, e.g., describing the image in a rough or detailed manner, in a factual or emotional view, etc. In this paper, we show that a unified model is qualified to perform well in diverse domains and freely switch among multiple styles. Such a controllable capability is achieved by embedding the prompt learning into the image captioning framework. To be specific, we design a set of prompts to fine-tune the pre-trained image captioner. These prompts allow the model to absorb stylized data from different domains for joint training, without performance degradation in each domain. Furthermore, we optimize the prompts with learnable vectors in the continuous word embedding space, avoiding the heuristic prompt engineering and meanwhile exhibiting superior performance. In the inference stage, our model is able to generate desired stylized captions by choosing the corresponding prompts. Extensive experiments verify the controllable capability of the proposed method. Notably, we achieve outstanding performance on two diverse image captioning benchmarks including COCO Karpathy split and TextCaps using a unified model.

1 Introduction

Image captioning is one of the fundamental tasks in computer vision, which aims to automatically generate natural and readable sentences to describe the image contents. The last decade has witnessed the rapid progress of image captioning, thanks to the development of sophisticated visual representation learning (Zhang et al. 2021; Fang et al. 2021), cross-modal fusion (Pan et al. 2020; Huang et al. 2019; Li et al. 2020), vision-language pre-training (Hu et al. 2021; Li et al. 2022; Wang et al. 2021b), etc. Image captioning is a challenging task that requires the captioners to recognize the objects and attributes, understand their relationships, and properly organize them in the sentence.

Despite the remarkable advances, current image captioning algorithms generally lack the controllable capability to generate desired captions. In other words, once the captioning model is trained, the caption generation process can hardly be influenced. Typical cases include the control of

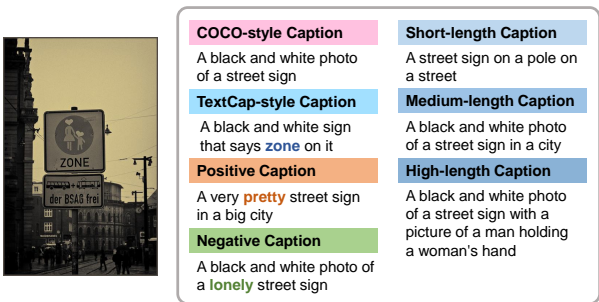


Figure 1: Leveraging a unified model, the proposed method is able to generate diverse captions such as COCO-style [COCO], TexCap-style [TextCap], Positive [Positive], Negative [Negative], and different caption lengths including Short-length [Short-length], Medium-length [Medium-length], and High-length [High-length]. Best view in color.

caption length and description style. (1) *Length controllable capability*. Sometimes, a brief description is required to get an overview of the image, while in other circumstances, a detailed caption is preferred to acquire more information. This can be roughly reflected by the controllable capability of the caption length, which is a basic demand in practical applications, but has been largely overlooked in existing methods. (2) *Style controllable capability*. An image can be described in quite different views. For example, given an image with textual contents (e.g., a poster or sign), some people care about the objects, but some may pay more attention to the textual words. Besides, people may generate non-factual captions, e.g., emotional descriptions that contain positive or negative expressions. It is of vital importance to insert different styles in the captioning model to enhance its expressibility. How to simultaneously maintain multiple styles and freely switch among them is an open problem. Existing captioning approaches typically separately handle each scenario, e.g., train a captioner on the COCO dataset (Lin et al. 2014) and train another model on the TextCaps dataset (Sidorov et al. 2020). As a result, these captioners are domain-specific, without style controllability.

In this paper, we show that a unified model is able to generate captions with different lengths and styles. As shown in Figure 1, our approach describes an image semantically

Có thể kiểm soát chú thích hình ảnh thông qua nhắc nhở

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, Linlin Li

Công ty Huawei

wn6149@mail.ustc.edu.cn, jh_xie@tongji.edu.cn, {wujihao, jiamingbo, lynn.lilinlin}@huawei.com



Hình 1: Tận dụng mô hình thống nhất, phương pháp đề xuất có thể tạo ra nhiều chú thích khác nhau như kiểu COCO [COCO], Kiểu TexCap [TextCap], Tích cực [Positive], Tiêu cực [Negative] và khác nhau độ dài chú thích bao gồm Độ dài ngắn [Short-length], Độ dài trung bình [Medium-length] và Độ dài cao [High-length]. Xem tốt nhất bằng màu.

độ dài chú thích và kiểu mô tả. (1) Có thể kiểm soát độ dài khả năng. Đôi khi, cần phải có một mô tả ngắn gọn để có được tổng quan về hình ảnh, trong khi ở những trường hợp khác, một nên có chú thích chi tiết để có thêm thông tin. Điều này có thể được phản ánh sơ bộ bằng khả năng kiểm soát độ dài chú thích, đây là yêu cầu cơ bản trong các ứng dụng thực tế như phân lớn đã bị bỏ qua trong các phương pháp hiện có. (2) Khả năng kiểm soát kiểu đáng. Một hình ảnh có thể được mô tả theo những góc nhìn khá khác nhau. Ví dụ, đưa ra một hình ảnh có nội dung văn bản (ví dụ, áp phích hoặc biển báo), một số mọi người chỉ quan tâm đến các đối tượng, nhưng một số có thể chú ý nhiều hơn đến các từ ngữ văn bản. Bên cạnh đó, mọi người có thể tạo ra chú thích không có thật, ví dụ, mô tả cảm xúc có chứa biểu hiện tích cực hoặc tiêu cực. Điều này có tầm quan trọng sống còn để chèn các kiểu khác nhau vào mô hình chú thích để nâng cao khả năng biểu đạt của nó. Làm thế nào để đồng thời duy trì nhiều phong cách và tự do chuyển đổi giữa chúng là một vấn đề mở. Các phương pháp chú thích hiện có thường xử lý riêng biệt mỗi kịch bản, ví dụ, đào tạo một người viết chú thích trên tập dữ liệu COCO (Lin et al. 2014) và đào tạo một mô hình khác trên TextCaps (Sidorov et al. 2020). Do đó, những người viết chú thích này mang tính đặc thù của từng miền, không thể kiểm soát được phong cách.

Trong bài báo này, chúng tôi trình bày rằng một mô hình thống nhất có thể tạo ra các chú thích có độ dài và kiểu khác nhau. Như đã trình bày trong Hình 1, cách tiếp cận của chúng tôi mô tả hình ảnh theo ngữ nghĩa

Chú thích hình ảnh là một trong những nhiệm vụ cơ bản trong tầm nhìn máy tính, nhằm mục đích tự động tạo ra hình ảnh tự nhiên và các câu dễ đọc để mô tả nội dung hình ảnh. thập kỷ qua đã chứng kiến sự tiến bộ nhanh chóng của việc chú thích hình ảnh, nhờ vào sự phát triển của công nghệ hình ảnh tính vi học biểu diễn (Zhang et al. 2021; Fang et al. 2021), sự kết hợp đa phương thức (Pan et al. 2020; Huang et al. 2019; Li et al. 2020), đào tạo trước ngôn ngữ thị giác (Hu et al. 2021; Li et al. 2022; Wang et al. 2021b), v.v. Chú thích hình ảnh là một nhiệm vụ đầy thách thức đòi hỏi người viết phụ đề phải nhận ra các đối tượng và thuộc tính, hiểu mối quan hệ của chúng và sắp xếp chúng một cách hợp lý trong câu.

Mặc dù có những tiến bộ đáng kể, các thuật toán chú thích hình ảnh hiện tại thường thiếu khả năng kiểm soát để tạo ra các chú thích mong muốn. Nói cách khác, sau khi mô hình chú thích được đào tạo, quá trình tạo chú thích có thể khó có thể bị ảnh hưởng. Các trường hợp điển hình bao gồm việc kiểm soát

accurately in diverse views. This captioning controllable capability is achieved by designing prompts within the cross-modal language model. After large-scale pre-training, the image captioner has already gained the ability to generate diverse captions, but is largely overwhelmed in the downstream fine-tuning, e.g., on a certain stylized dataset such as COCO (Lin et al. 2014). In this work, we aim to unveil the potential hidden in the pre-trained model to flexibly switch captioning styles. Our approach is motivated by the recent advance in prompt learning techniques (Liu et al. 2021) in natural language processing (NLP). In the proposed framework, prompts serve as the anchor points to gather data from different domains, facilitating the multi-domain joint training. By virtue of prompt engineering, captions with different lengths, different styles, and different emotions can be properly separated within a unified model. The prompts, together with the image-text pair, jointly serve as the training corpus to optimize the captioning model. Furthermore, instead of manually designing prompts, we encourage the captioner to automatically learn the prompt embeddings in an end-to-end manner. This continuous auto-prompt learning searches the suitable prompt representations in the entire word embedding space, which not only avoids the heuristic prompt design but also exhibits superior performance.

In the inference stage, different prompts serve as the prediction hints to guide the caption generation. By automatically learning multiple prompt embeddings, the proposed approach has the following merits. Our approach (i) is free of manual prompt engineering, which requires domain expertise and careful word tuning; (ii) is able to generate diverse stylized captions via a single model, which is infeasible for most existing state-of-the-art captioners such as BLIP (Li et al. 2022), LEMON (Hu et al. 2021), and SimVLM (Wang et al. 2021b); (iii) does not degrade the performance on different domains such as COCO (Lin et al. 2014) and TextCaps (Sidorov et al. 2020), and outperforms the traditional training strategy using a prefixed prompt; (iv) is simple and general, which is ready to perform on more domains by incorporating other stylized data.

In summary, the contributions of this work are three-fold:

- To our knowledge, we are the first to propose the prompt-based image captioning framework, which provides a simple yet effective manner to control the caption style.
- We validate the manually designed prompts. We further introduce auto-prompt learning to avoid the heuristic prompt design and achieve superior results.
- Qualitative and quantitative results verify the controllable capability of the proposed framework. Leveraging a unified model, we achieve outstanding performance on several benchmarks including COCO Karpathy set (Lin et al. 2014), NoCaps (Agrawal et al. 2019), and TextCaps (Sidorov et al. 2020).

2 Related Work

General Image Captioning. Image captioning aims to generate a textual description of the image contents (Vinyals et al. 2015), which typically contain a visual encoder to extract the image features and a multi-modal fusion model

such as LSTM and Transformer for text generation. To represent the visual contents, previous methods (Huang et al. 2019; Anderson et al. 2018; Deng et al. 2020; Cornia et al. 2020; Fei 2022; Ji et al. 2021) utilize the Region-of-Interest (RoI) features from object detectors (Ren et al. 2016). Recent captioning algorithms (Fang et al. 2021; Xu et al. 2021; Wang et al. 2021b) shed light on the grid features for high efficiency and potentially better performance due to end-to-end training. As for the cross-modal model, classic captioners (Anderson et al. 2018; Huang et al. 2019; Pan et al. 2020; Song et al. 2021) typically utilize the LSTM, while the recent approaches (Li et al. 2020; Zhang et al. 2021; Li et al. 2022; Wang et al. 2021b; Wang, Xu, and Sun 2022; Luo et al. 2021) leverage the attention-based models to fuse vision-language representations and predict the captions.

Controllable Image Captioning. Despite the impressive progress, fewer efforts have been made to control the caption generation. Cornia *et al.* (Cornia, Baraldi, and Cucchiara 2019) utilize image regions to generate region-specific captions. Chen *et al.* (Chen et al. 2020a) propose the abstract scene graph to represent user intention and control the generated image captions. Length-controllable captioning approach is proposed in (Deng et al. 2020), which learns length level embeddings to control the caption length. Shuster *et al.* (Shuster et al. 2019) release an image captioning dataset with personality traits as well as a baseline approach. Zhang *et al.* (Zhang et al. 2022) propose a multi-modal relational graph adversarial inference (MAGIC) framework for diverse text caption. SentiCap (Mathews, Xie, and He 2016) utilizes a switching recurrent neural network with word-level regularization to generate emotional captions. Chen *et al.* (Chen et al. 2018) present a style-factual LSTM to generate captions with diverse styles such as humorous and romantic. However, some of the aforementioned methods (Cornia, Baraldi, and Cucchiara 2019; Chen et al. 2020a, 2018) rely on additional tools or expensive annotations for supervision. In (Kobus, Crego, and Senellart 2016), domain/tag embeddings are involved to control the style, and thus the model architecture is tag-related. Some methods (Mathews, Xie, and He 2016; Chen et al. 2018) can be regarded as the ensemble framework, which include two groups of parameters for factual and stylized branches, increasing the model complexity.

In this work, we control the image captioning style from a different view, i.e., prompt learning. The proposed framework merely involves lightweight learnable prompt embeddings while keeping the baseline architecture unchanged, which is conceptually simple and easy to implement.

Vision-language Pre-training. Vision-language (VL) pre-training is a popular manner to bridge vision and language representations (Dou et al. 2021). CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) use the cross-modal contrastive learning to align the VL representations. Recent VL pre-training approaches (Zhou et al. 2020; Chen et al. 2020b; Huang et al. 2021) generally adopt the attention mechanism (Vaswani et al. 2017) to fuse the VL representations. After large-scale pre-training on the image-text corpus, these models are further fine-tuned on the downstream datasets to conduct a variety of VL tasks such as image captioning. SOHO (Huang et al. 2021) extracts compact image features via a

chính xác trong nhiều góc nhìn khác nhau. Khả năng kiểm soát chủ thích này đạt đư ợc bằng cách thiết kế lời nhắc trong mô hình ngôn ngữ đa phư ơng thức. Sau khi đào tạo trư ớc trên quy mô lớn, ngư ời chủ thích hình ảnh đã đạt đư ợc khả năng tạo ra nhiều chủ thích khác nhau, nhưng phần lớn bị choáng ngợp trong quá trình tinh chỉnh hạ lư u, ví dụ, trên một tập dữ liệu cách điệu nhất định như COCO (Lin et al. 2014). Trong công trình này, chúng tôi muốn tiết lộ tiềm năng ẩn trong mô hình đư ợc đào tạo trư ớc để chuyển đổi linh hoạt phong cách chủ thích. Cách tiếp cận của chúng tôi đư ợc thúc đẩy bởi gần đây tiến bộ trong các kỹ thuật học nhanh (Liu et al. 2021) trong xử lý ngôn ngữ tự nhiên (NLP). Trong khuôn khổ đư ợc đề xuất, các lời nhắc đóng vai trò là điểm neo để thu thập dữ liệu từ các miền khác nhau, tạo điều kiện thuận lợi cho việc đào tạo chung đa miền. Nhờ kỹ thuật nhanh chóng, các chủ thích có nhiều độ dài, phong cách khác nhau và cảm xúc khác nhau có thể đư ợc tách biệt hợp lý trong một mô hình thống nhất. Các lời nhắc, cùng nhau với cặp hình ảnh-văn bản, cùng nhau đóng vai trò là tập dữ liệu đào tạo để tối ư u hóa mô hình chủ thích. Hơ n nữa, thay vào đó của việc thiết kế lời nhắc thủ công, chúng tôi khuyến khích ngư ời viết chủ thích để tự động học các nhúng nhắc nhở theo cách từ đầu đến cuối. Việc học tự động nhắc nhở liên tục này tìm kiếm các biểu diễn nhắc nhở phù hợp trong toàn bộ không gian nhúng từ, không chỉ tránh đư ợc nhắc nhở theo phư ơng pháp trực quan thiết kế nhưng cũng thể hiện hiệu suất vư ợt trội.

Trong giai đoạn suy luận, các lời nhắc khác nhau đóng vai trò là gợi ý dự đoán để hướng dẫn việc tạo chủ thích. Bằng cách tự động học nhiều nhúng lời nhắc, đề xuất cách tiếp cận có những ưu điểm sau. Cách tiếp cận của chúng tôi (i) là miễn phí của kỹ thuật nhắc nhở thủ công, đòi hỏi chuyên môn về lĩnh vực này và điều chỉnh từ ngữ cẩn thận; (ii) có thể tạo ra các chủ thích cách điệu đa dạng thông qua một mô hình duy nhất, điều này không khả thi đối với hầu hết các trình chủ thích hiện đại như BLIP (Li và cộng sự 2022), LEMON (Hu và cộng sự 2021) và SimVLM (Wang et al. 2021b); (iii) không làm giảm hiệu suất trên các miền khác nhau như COCO (Lin et al. 2014) và TextCaps (Sidorov và cộng sự 2020) và vư ợt trội hơ n chiến lược đào tạo truyền thống bằng cách sử dụng lời nhắc có tiên tổ; (iv) đơn giản và chung chung, có thể thực hiện trên nhiều miền hơ n bằng cách kết hợp các dữ liệu cách điệu khác.

Tóm lại, tác phẩm này có ba đóng góp:

- Theo hiểu biết của chúng tôi, chúng tôi là ngư ời đầu tiên đề xuất khuôn khổ chủ thích hình ảnh dựa trên lời nhắc, cung cấp cách đơn giản nhưng hiệu quả để kiểm soát kiểu chủ thích.
- Chúng tôi xác thực các lời nhắc đư ợc thiết kế thủ công. Chúng tôi tiếp tục giới thiệu việc học lời nhắc tự động để tránh phư ơng pháp tìm kiếm thiết kế nhanh chóng và đạt đư ợc kết quả vư ợt trội.
- Kết quả định tính và định lư ợng xác minh khả năng kiểm soát của khuôn khổ đư ợc đề xuất. Tận dụng một mô hình thống nhất, chúng tôi đạt đư ợc hiệu suất vư ợt trội trên một số chuẩn mực bao gồm bộ COCO Karpathy (Lin et al. 2014), NoCaps (Agrawal et al. 2019) và TextCaps (Sidorov và cộng sự, 2020).

2 Công trình liên quan

Chủ thích hình ảnh chung. Chủ thích hình ảnh nhằm mục đích tạo ra mô tả văn bản về nội dung hình ảnh (Vinyals et al. 2015), thư ờng chứa bộ mã hóa hình ảnh để trích xuất các đặc điểm hình ảnh và mô hình hợp nhất đa phư ơng thức

chẳng hạn như LSTM và Transformer để tạo văn bản. Để biểu diễn nội dung trực quan, các phư ơng pháp trư ớc đây (Huang et al. 2019; Anderson và cộng sự. 2018; Đặng và cộng sự. 2020; Cornia và cộng sự. 2020; Fei 2022; Ji et al. 2021) sử dụng Khu vực quan tâm (RoI) các tính năng từ các máy dò đối tượng (Ren et al. 2016). Các thuật toán chủ thích gần đây (Fang et al. 2021; Xu et al. 2021; Wang et al. 2021b) làm sáng tỏ các tính năng lư ợi cho hiệu quả và hiệu suất có khả năng tốt hơn do đào tạo toàn diện. Đối với mô hình đa phư ơng thức, các chủ thích cổ điển (Anderson và cộng sự, 2018; Huang và cộng sự, 2019; Pan và cộng sự, 2020; Song et al. 2021) thư ờng sử dụng LSTM, trong khi các cách tiếp cận gần đây (Li et al. 2020; Zhang et al. 2021; Li et al. 2022; Wang et al. 2021b; Wang, Xu và Sun 2022; Luo et al. 2021) tận dụng các mô hình dựa trên sự chú ý để hợp nhất biểu diễn ngôn ngữ thị giác và dự đoán chủ thích.

Chủ thích hình ảnh có thể kiểm soát. Mặc dù ẩn tượng tiến bộ, ít nỗ lực hơn đã đư ợc thực hiện để kiểm soát chủ thích thể hệ. Cornia và cộng sự. (Cornia, Baraldi và Cucchiara 2019) sử dụng các vùng hình ảnh để tạo ra các chủ thích cụ thể cho từng vùng. Chen et al. (Chen et al. 2020a) đề xuất tóm tắt đồ thị cảnh để biểu diễn ý định của ngư ời dùng và kiểm soát các chủ thích hình ảnh đư ợc tạo ra. Phư ơng pháp chủ thích có thể kiểm soát độ dài đư ợc đề xuất trong (Deng et al. 2020), trong đó học đư ợc độ dài những mức độ để kiểm soát độ dài chủ thích. Shuster et al. (Shuster et al. 2019) phát hành một tập dữ liệu chủ thích hình ảnh với các đặc điểm tính cách cũng như cách tiếp cận cơ bản. Zhang et al. (Zhang et al. 2022) đề xuất một quan hệ đa phư ơng thức khuôn khổ suy luận đối nghịch đồ thị (MAGIC) cho đa dạng chủ thích văn bản. SentiCap (Mathews, Xie và He 2016) sử dụng mạng nơ -ron hồi quy chuyển mạch với cấp độ từ chính quy hóa để tạo ra chủ thích cảm xúc. Chen et al. (Chen et al. 2018) trình bày một LSTM theo phong cách thực tế để tạo ra chủ thích với nhiều phong cách khác nhau như hài hước và lãng mạn. Tuy nhiên, một số phư ơng pháp đã đề cập ở trên (Cornia, Baraldi và Cucchiara 2019; Chen và cộng sự. 2020a, 2018) dựa vào về các công cụ bổ sung hoặc chủ thích tồn kém để giám sát. Trong (Kobus, Crego và Senellart 2016), các nhúng miền/thể đư ợc sử dụng để kiểm soát kiểu và do đó kiến trúc mô hình có liên quan đến thể. Một số phư ơng pháp (Mathews, Xie và He 2016; Chen et al. 2018) có thể đư ợc coi là tập hợp khung, bao gồm hai nhóm tham số cho các nhánh thực tế và cách điệu, làm tăng độ phức tạp của mô hình.

Trong công việc này, chúng tôi kiểm soát phong cách chủ thích hình ảnh từ một góc nhìn khác, tức là, học nhanh. Khung đề xuất chỉ liên quan đến các nhúng nhanh có thể học đư ợc nhẹ trong khi vẫn giữ nguyên kiến trúc cơ sở,

về mặt khái niệm thì đơn giản và dễ thực hiện. Tiền đào tạo ngôn ngữ thị giác. Tiền đào tạo ngôn ngữ thị giác (VL) là một cách phổ biến để kết nối thị giác và ngôn ngữ biểu diễn (Dou et al. 2021). CLIP (Radford et al. 2021) và ALIGN (Jia et al. 2021) sử dụng phư ơng pháp tư ơng phản chéo học cách căn chỉnh các biểu diễn VL. Các phư ơng pháp đào tạo trư ớc VL gần đây (Zhou et al. 2020; Chen et al. 2020b; Huang et al. 2021) thư ờng áp dụng cơ chế chú ý (Vaswani et al. 2017) để hợp nhất các biểu diễn VL. Sau đào tạo trư ớc quy mô lớn trên ngữ liệu hình ảnh-văn bản, các mô hình này đư ợc tinh chỉnh thêm trên các tập dữ liệu hạ lư u để thực hiện nhiều tác vụ VL khác nhau như chủ thích hình ảnh. SOHO (Huang et al. 2021) trích xuất các đặc điểm hình ảnh nhỏ gọn thông qua

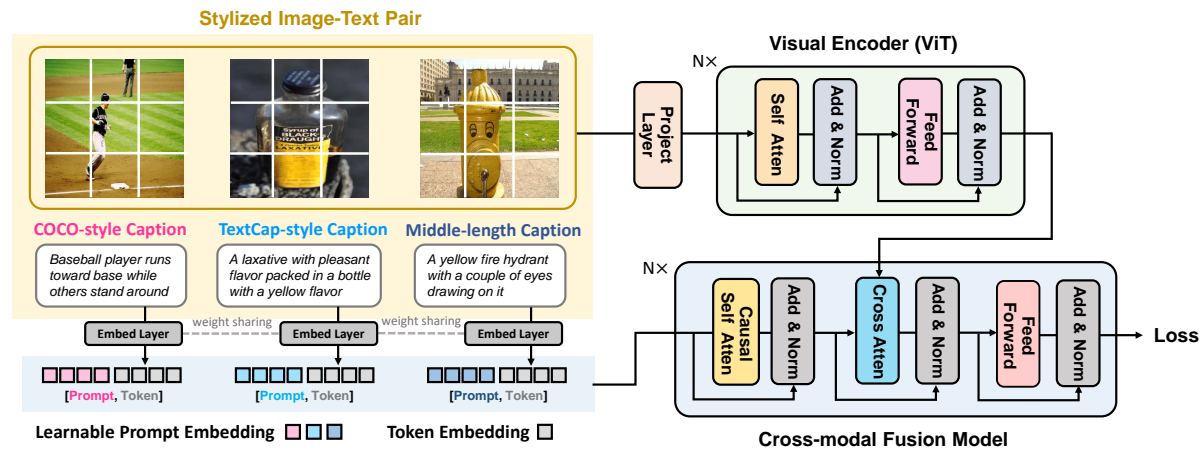


Figure 2: An overview of the proposed prompt-based image captioning framework. Our model optimizes multiple learnable prompt embeddings to absorb stylized data from different domains to jointly train the image captioner. In the inference stage, the model is able to generate diverse captions by feeding different prompts.

learned visual dictionary and trains the whole framework in an end-to-end manner. ALBEF (Li et al. 2021) conducts the cross-modal alignment using contrastive learning technique (Radford et al. 2021) before representation fusion. SimVLM (Wang et al. 2021b) utilizes prefix language modeling for model optimization on the large-scale VL corpus. Inspired by previous arts, we also involve VL pre-training to improve the captioning quality.

Prompt Learning. Prompt learning has gained increasing popularity in natural language processing (NLP) (Liu et al. 2021). Prompt learning allows the language model to be pre-trained on the large-scale corpus, and is able to perform downstream tasks by defining a proper prompting function. Jiang *et al.* (Jiang et al. 2020) propose mining-based and paraphrasing-based approaches to automatically generate high-quality prompts. Shin *et al.* (Shin et al. 2020) search for the proper prompts via a gradient-based approach. Recently, continuous prompt learning has been explored, which directly optimize prompt vectors in the continuous word embedding space (Zhong, Friedman, and Chen 2021; Li and Liang 2021; Lester, Al-Rfou, and Constant 2021; Zhou et al. 2021). It is worth mentioning that prompt learning has been rarely touched in the image captioning. Different from the traditional usage of prompt learning that aims to elicit knowledge for higher performance, we focus on the controllable capability of the captioning algorithm. In the proposed framework, except for the superior performance, the more attractive characteristic is that we can freely switch diverse styles via prompting, which greatly enhances the controllability and expressibility of the image captioner.

3 Approach

In this section, we introduce the method details of the proposed controllable image captioner. First, in Section 3.1, we revisit autoregressive image captioning, which serves as the baseline of our approach. Then, in Section 3.2, we elaborate the manual prompt engineering for image captioning.

Finally, we exhibit how to optimize the learnable prompts in Section 3.3 and the inference details in Section 3.4.

3.1 Revisiting Autoregressive Image Captioning

In our method, we adopt the unidirectional language modeling (LM) based image captioning framework as the baseline. Such a framework typically utilizes a transformer block to fuse the image v and text sequence $x = \{x_1, x_2, \dots, x_n\}$. The token x_t is generated in an autoregressive manner based on the previous tokens $x_{<t}$. The training objective of the cross-modal LM loss is as follows:

$$\mathcal{L}_{\text{LM}} = -\mathbb{E}_{(v, x) \in \mathcal{D}} \left[\sum_t \log P(x_t | g(v), f(x_{<t})) \right], \quad (1)$$

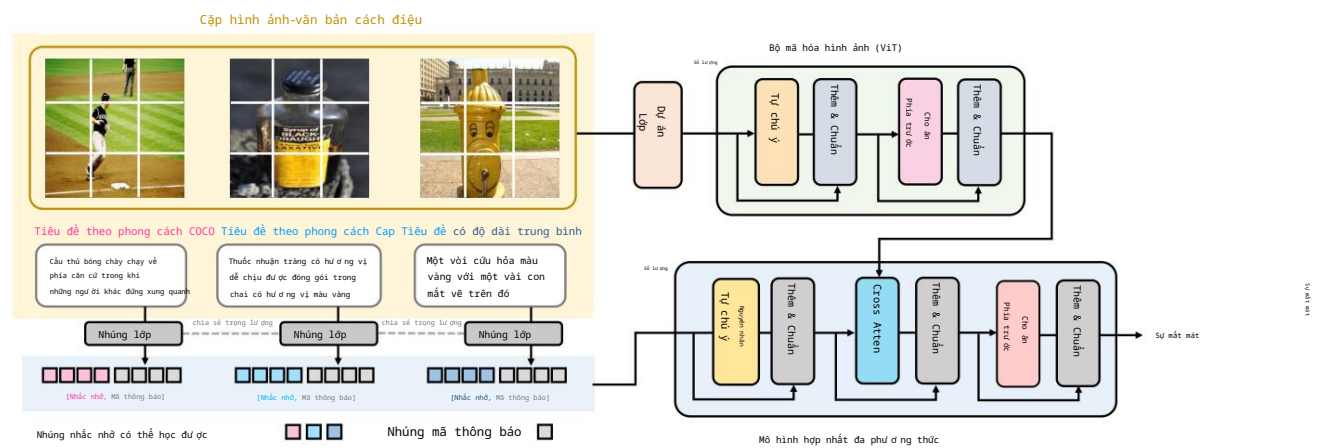
where $g(\cdot)$ denotes the visual encoder, $f(\cdot)$ represents the word embedding layer, $P(\cdot|\cdot)$ can be regarded as the cross-modal fusion model (e.g., transformer decoder in Figure 2), which receives the visual features $g(v)$ and previous token embeddings $f(x_{<t})$ to predict the next word token x_t .

During inference, the autoregressive models take a special token [BOS] as input to predict the first token x_1 , then x_1 is fed into the model to obtain the next token x_2 . This autoregressive prediction process is continued until the special token [EOS] is predicted.

3.2 Prompt-based Image Captioning

Model Pre-training. Following previous works (Zhang et al. 2021; Hu et al. 2021; Li et al. 2022; Wang et al. 2021b), we also adopt the large-scale pre-training on the noisy image-text corpus to improve the downstream captioning task. Besides the language modeling (LM) loss, we also adopt the image-text contrastive loss (Radford et al. 2021; Jia et al. 2021) and image-text matching loss (Chen et al. 2020b; Li et al. 2021, 2022) to jointly optimize the visual encoder and cross-modal fusion model, as follows:

$$\mathcal{L}_{\text{Pre-train}} = \mathcal{L}_{\text{Contrast}} + \mathcal{L}_{\text{Match}} + \mathcal{L}_{\text{LM}}. \quad (2)$$



Hình 2: Tổng quan về khuôn khổ chú thích hình ảnh dựa trên lời nhắc được đề xuất. Mô hình của chúng tôi tối ưu hóa nhiều khả năng học được những nhanh để hấp thụ dữ liệu cách điệu từ các miền khác nhau để cùng nhau đào tạo người chú thích hình ảnh. Trong giai đoạn suy luận, mô hình có thể tạo ra nhiều chú thích khác nhau bằng cách đưa ra nhiều lời nhắc khác nhau.

đã học từ điển hình ảnh và đào tạo toàn bộ khuôn khổ trong theo cách từ đầu đến cuối. ALBEF (Li et al. 2021) tiến hành căn chỉnh chéo phụ trợ sử dụng kỹ thuật học tự động phân (Radford et al. 2021) trước khi hợp nhất biểu diễn. SimVLM (Wang et al. 2021b) sử dụng mô hình ngôn ngữ tiền tố cho tối ưu hóa mô hình trên kho dữ liệu VL quy mô lớn. Lấy cảm hứng theo các nghệ thuật trước đó, chúng tôi cũng liên quan đến việc đào tạo trước VL để cải thiện chất lượng phụ đề. Học nhanh. Học nhanh đã đạt được sự gia tăng sự phổ biến trong xử lý ngôn ngữ tự nhiên (NLP) (Liu et al. 2021). Việc học nhanh cho phép mô hình ngôn ngữ được được đào tạo trước trên kho dữ liệu quy mô lớn và có thể thực hiện các tác vụ hạ nguồn bằng cách xác định chức năng nhắc nhở thích hợp. Jiang et al. (Jiang et al. 2020) đề xuất khai thác dựa trên và các phụ trợ pháp tiếp cận dựa trên diễn giải để tự động tạo ra các lời nhắc chất lượng cao. Shin et al. (Shin et al. 2020) tìm kiếm cho các lời nhắc thích hợp thông qua phụ trợ pháp tiếp cận dựa trên gradient. Gần đây, việc học lời nhắc liên tục đã được khám phá, trực tiếp tối ưu hóa các vectơ nhắc trong từ liên tục những không gian (Zhong, Friedman và Chen 2021; Li và Liang 2021; Lester, Al-Rfou và Constant 2021; Zhou et al. 2021). Điều đáng nói là việc học nhanh chống đã hiếm khi được đề cập đến trong chú thích hình ảnh. Khác với cách sử dụng truyền thống của việc học nhanh nhằm mục đích gợi ra kiến thức để có hiệu suất cao hơn, chúng tôi tập trung vào khả năng kiểm soát nhân của thuật toán chú thích. Trong đề xuất khung, ngoại trừ hiệu suất vượt trội, càng nhiều đặc điểm hấp dẫn là chúng ta có thể tự do chuyển đổi đa dạng tạo kiểu thông qua lời nhắc, giúp tăng cường đáng kể khả năng kiểm soát và khả năng diễn đạt của người chú thích hình ảnh.

3 Cách tiếp cận

Trong phần này, chúng tôi giới thiệu chi tiết phụ trợ pháp của trình chú thích hình ảnh có thể điều khiển được đề xuất. Đầu tiên, trong Phần 3.1, chúng tôi xem lại chú thích hình ảnh tự hồi quy, đóng vai trò như cơ sở của cách tiếp cận của chúng tôi. Sau đó, trong Phần 3.2, chúng tôi sẽ trình bày chi tiết về kỹ thuật nhắc nhở thủ công cho chú thích hình ảnh.

Cuối cùng, chúng tôi trình bày cách tối ưu hóa các lời nhắc có thể học được trong Mục 3.3 và các chi tiết suy luận trong Mục 3.4.

3.1 Xem lại chú thích hình ảnh tự động hồi quy

Trong phụ trợ pháp của mình, chúng tôi áp dụng khuôn khổ chú thích hình ảnh dựa trên mô hình ngôn ngữ đơn hướng (LM) làm cơ sở. Một khuôn khổ như vậy thường sử dụng một khối biến áp để hợp nhất hình ảnh v và chuỗi văn bản $x = \{x_1, x_2, \dots, x_n\}$. Mã thông báo xt được tạo theo cách tự hồi quy dựa trên trên các mã thông báo trước đó $x_{<t}$. Mục tiêu đào tạo của Mất mát LM liên phụ trợ thức như sau:

$$\text{LLM} = -\mathbb{E}_{(v, x) \in \mathcal{D}} \left[\sum_t \log P(x_t | g(v), f(x_{<t})) \right], \quad (1)$$

trong đó $g(\cdot)$ biểu thị bộ mã hóa trực quan, $f(\cdot)$ biểu thị lớp những từ, $P(\cdot|\cdot)$ có thể được coi là mô hình hợp nhất đa phụ trợ thức (ví dụ, bộ giải mã biến áp trong Hình 2), nhận được các tính năng trực quan $g(v)$ và mã thông báo trước đó những $f(x_{<t})$ để dự đoán mã thông báo tiếp theo x_t .

Trong quá trình suy luận, các mô hình hồi quy tự động thực hiện một cách đặc biệt token [BOS] làm đầu vào để dự đoán token đầu tiên x_1 , sau đó là x_1 được đưa vào mô hình để có được mã thông báo tiếp theo x_2 . Quá trình dự đoán hồi quy tự động này được tiếp tục cho đến khi đặc biệt token [EOS] được dự đoán.

3.2 Chú thích hình ảnh dựa trên lời nhắc

Mô hình đào tạo trước. Theo các tác phẩm trước đó (Zhang và cộng sự. 2021; Hu và cộng sự. 2021; Li và cộng sự. 2022; Vũ và cộng sự. 2021b), chúng tôi cũng áp dụng chương trình đào tạo trước quy mô lớn về ngữ liệu hình ảnh-văn bản nhiều để cải thiện nhiệm vụ chú thích hạ lưu. Bên cạnh việc mất mô hình ngôn ngữ (LM), chúng tôi cũng áp dụng sự mất mát tự động phản hình ảnh-văn bản (Radford et al. 2021; Jia et al. 2021) và mất kết nối hình ảnh-văn bản (Chen et al. 2020b; Li et al. 2021, 2022) để cùng nhau tối ưu hóa hình ảnh bộ mã hóa và mô hình hợp nhất đa phụ trợ thức, như sau:

$$\text{LPre-train} = \text{LContrast} + \text{LMatch} + \text{LLM}. \quad (2)$$

The contrastive loss measures the similarity of the image-text pairs via a light fusion manner such as dot-product, while the matching loss measures the image-text similarity via a heavy fusion manner such as cross-attention. It has been widely recognized that both of them can facilitate cross-modal alignment (Li et al. 2022). Therefore, although we focus on the image captioning, we additionally include the $\mathcal{L}_{\text{Contrast}}$ and $\mathcal{L}_{\text{Match}}$ in the pre-training stage. As for more details, please refer to BLIP (Li et al. 2022).

Prompt Engineering. After pre-training, the model already acquires zero-shot captioning capability thanks to the language modeling loss \mathcal{L}_{LM} . Therefore, previous LM-based image captioners such as SimVLM (Wang et al. 2021b) and BLIP (Li et al. 2022) leverage a pre-defined prompt such as “a picture of” or “a photo of” to facilitate the image captioning. In this work, we aim to unveil the model potential of generating diverse captions via prompting.

In contrast to single prompt engineering, in the fine-tuning stage, we design multiple prompts as the anchors to distinguish the training data from different domains. In this way, different stylized captions do not disturb their counterparts and together contribute to a stronger model. The manually designed prompts are illustrated in Table 1. (i) For the cross-domain scenario, e.g., evaluating a model on both COCO (Lin et al. 2014) and TextCaps (Sidorov et al. 2020), it is straightforward to assign different prompts for these datasets to learn domain-specific descriptions. (ii) As for the caption length control, we divide the image captions from COCO and TextCaps into three levels depending on the caption length. Captions whose length is in the range [0, 10), [10, 16), and [16, $+\infty$) are divided. Each of these subsets is assigned with a specific prompt, as shown in Table 1. (iii) Finally, current image captions are typically factual. Nevertheless, each image in the COCO dataset is labeled by five annotators, inevitably containing emotional descriptions. To this end, we collect the positive and negative captions in the COCO dataset to form the non-factual subsets, which contain the pre-defined positive words such as “great, nice, cute” and negative words such as “ugly, terrible, disgusting”. Despite these non-factual captions being rare, our method still learns satisfying styles using limited samples, justifying the few-shot learning ability of prompt engineering. The entire positive and negative words, and other potentially effective manual prompts are presented in the *supplementary material*.

Model Fine-tuning. In our framework, multiple training sets are mixed together to train a unified model. Compared to Eq. (1), we predict token x_t based on the visual features $g(\mathbf{v})$, prompt token embeddings $f(\mathbf{p})$, and previous token embeddings $f(\mathbf{x}_{<t})$. Different stylized data is assigned with a specific prompt as illustrated in Table 1. During training, we prepend these hand-crafted prompts to caption tokens as the textual description of the image. We assemble different stylized datasets to jointly train the captioning model using a prompt-based LM loss $\mathcal{L}_{\text{ProLM}}$ as follows:

$$\mathcal{L}_{\text{ProLM}} = - \sum_i \left[\mathbb{E}_{(\mathbf{v}, \mathbf{p}, \mathbf{x}) \in \mathcal{D}_i} \left[\sum_t \log P(x_t | g(\mathbf{v}), f(\mathbf{p}_i), f(\mathbf{x}_{<t})) \right] \right], \quad (3)$$

Caption Style	Manual Prompt \mathbf{p}
COCO-style	a normal picture that shows
TextCap-style	a textual picture that shows
Short-length	a picture with a short caption that shows
Medium-length	a picture with a medium caption that shows
High-length	a picture with a long caption that shows
Positive	a positive picture that shows
Negative	a negative picture that shows

Table 1: Illustration of the manual prompts.

where \mathbf{p}_i denotes the manual prompt for i -th dataset \mathcal{D}_i . Note that the prompt tokens \mathbf{p}_i and caption tokens $\mathbf{x}_{<t}$ share the same embedding mapping layer $f(\cdot)$. In this framework, we keep the baseline model architecture unchanged without additional learnable blocks, which is parameter-efficient.

3.3 Auto-prompt Learning

To avoid the laborious manual prompt engineering in Section 3.2, we further encourage the network to automatically learn the prompts in an end-to-end manner, as shown in Figure 2. Given a sequence of the manual prompt tokens such as “a textual picture that shows”, the model first maps each token to a unique numeric ID using WordPiece technique. Then, for a BERT-base model, the token IDs are projected to 768-dim word embeddings via the token embedding layer $f(\cdot)$ as the input of the vision-language fusion model, i.e., $f(\mathbf{p}) \in \mathbb{R}^{N \times 768}$, where N represents the prompt length. Instead of the manual prompt engineering, we propose to learn the caption prompt embeddings \mathbf{P} as follows:

$$\mathbf{P} = [\mathbf{P}]_1 [\mathbf{P}]_2 \cdots [\mathbf{P}]_N, \quad (4)$$

where each embedding vector $[\mathbf{P}]_k$ ($k \in 1, \cdots, N$) has the same dimension as the word embedding. In other words, $\mathbf{P} \in \mathbb{R}^{N \times 768}$ serves as an alternative of the manual prompt embedding $f(\mathbf{p})$. In the training stage, prompt embeddings \mathbf{P} are jointly optimized with the captioning network as follows:

$$\mathcal{L}_{\text{AutoProLM}} = - \sum_i \left[\mathbb{E}_{(\mathbf{v}, \mathbf{x}) \in \mathcal{D}_i} \left[\sum_t \log P(x_t | g(\mathbf{v}), \mathbf{P}_i, f(\mathbf{x}_{<t})) \right] \right]. \quad (5)$$

The proposed framework learns specific prompt embeddings \mathbf{P}_i for each domain-specific dataset \mathcal{D}_i . During the end-to-end training, the gradients can be effectively back-propagated to optimize the prompt embeddings. To this end, the captioner is able to fully explore the suitable prompt representations in the continuous word embedding space.

3.4 Prompt-based Inference

After prompt learning, our model is able to generate diverse captions using different prompts. In the manual prompt framework, after encoding the special token [BOS], we sequentially embed the prompt tokens via $f(\mathbf{p})$ and feed them to the language model to generate the caption in an autoregressive manner. In the auto-prompt framework, we directly concatenate the token embedding of [BOS] and learned prompt embeddings \mathbf{P} as the input of the language model.

Machine Translated by Google

Mất mát tư ơ ng ph ần đ o l ư ờ ng s ự gi ờ ng nh au củ a các c ặp h ình ả nh-v ă n b ả n th ờ ng qu a m ột ph ư ơ ng ph ả p h ợ p nh ất nh ẹ nh ư t ích đ i ể m, tr ờ ng k hi m ấ t m ấ t k h ớ p đ o l ư ờ ng s ự gi ờ ng nh au củ a h ình ả nh-v ă n b ả n th ờ ng qu a m ột ph ư ơ ng ph ả p h ợ p nh ất n ặ ng nh ư chú ý ch éo. Ng ư ờ i ta đ ă c ờ ng nh ậ n r ộ ng r ằ ng cả hai đ ều có th ể t ạo đ i ề u k i ệ n cho vi ệ c c ă n ch ỉnh ch éo ph ư ơ ng th ứ c (Li et al. 2022). Đ o đ ố , m ặ c dù ch úng t ờ i t ậ p tr ư ng v ào chú th ích h ình ả nh, ch úng t ờ i c ũ ng b ao g ồ m LContrast và LMatch tr ờ ng g ă i đ o ă n ti ề n đ ă o t ă o. Đ ể bi ế t th êm ch i ti ế t, vui l ò ng th ă m k ă o BLIP (Li et al. 2022).

Kỹ thu ậ t nh ắc nh ờ. S ă u k hi đ ă o t ă o tr ư ớ c, m ô h ình đ ă c ố đ ư ợ c k ă n ă ng chú th ích zero-shot nh ờ m ô h ình m ấ t ng ờ n g ữ LLM. Đ o đ ố , các tr ình chú th ích h ình ả nh đ ư a tr ê n LM tr ư ớ c đ ă y nh ư SimVLM (Wang và c ộ ng s ự 2021b) và BLIP (Li và c ộ ng s ự 2022) t ậ n d ư ng l ờ i nh ắc đ ư c x ă c đ ị nh tr ư ớ c nh ư "m ột b ứ c ả nh củ a" ho ặ c "m ột b ứ c ả nh củ a" đ ể t ă o đ i ề u k i ệ n th u ậ n l ợ i cho vi ệ c chú th ích h ình ả nh. Tr ờ ng c ờ ng tr ình n ă y, ch úng t ờ i h ư ớ ng đ ế n vi ệ c ti ế t l ộ ti ề m n ă ng củ a m ô h ình tr ờ ng vi ệ c t ă o r ă các chú th ích đ ă d ạng th ờ ng qu a l ờ i nh ắc.

Ng ư ợ c l ă i v ớ i kỹ thu ậ t nh ắc nh ờ đ ơ n l ễ , tr ờ ng g ă i đ o ă n t ỉ nh ch ỉ nh, ch úng t ờ i th i ế t k ế nh i ề u nh ắc nh ờ l ă m đ i ể m neo đ ể ph ă n bi ế t d ữ li ệ u đ ă o t ă o v ớ i các mi ề n k ă c nh ă u. Th ề o c ă c h n ă y, các chú th ích c ă c h đ i ề u k ă c nh ă u k ờ ng l ă m h ư ớ ng đ ế n các ph ă n t ư ơ ng ứ ng củ a ch úng và c ũ ng nh ă u g ố p ph ă n t ă o n ề n m ô h ình m ặ n h ơ n. Các nh ắc nh ờ đ ư c th i ế t k ế th ử c ờ ng đ ư c m ỉ nh h ợ a tr ờ ng B ă ng 1. (i) Đ ố i v ớ i k ị ch b ả n li ề n mi ề n, v ớ i d ự , đ ă nh g ă m ột m ô h ình tr ê n cả COCO (Lin et al. 2014) và TextCaps (Sidorov et al. 2020), vi ệ c ch ỉ đ ị nh các l ờ i nh ắc k ă c nh ă u cho các t ậ p d ữ li ệ u n ă y đ ể h ợ c các m ô t ă c ự th ể cho t ũ ng mi ề n l ă đ i ề u đ ễ d ă ng. (ii) Đ ố i v ớ i vi ệ c k i ể m s ă o ă t đ ộ d ă i chú th ích, ch úng t ờ i ch i ă c các chú th ích t ích c ự c và ti ề u c ự c tr ờ ng t ậ p d ữ li ệ u COCO đ ể t ă o th ă nh các t ậ p h ợ p con k ờ ng có th ực, ch ữ a các t ữ t ích c ự c đ ư c x ă c đ ị nh tr ư ớ c nh ư "t uyệt , đ ẹp, đ ễ th ư ơ ng" và các t ữ ti ề u c ự c nh ư "x ă u x í, t ệ h ă i, k ỉ nh t ờ m". M ặ c dù các chú th ích k ờ ng có th ực n ă y r ấ t h i ể m, ph ư ơ ng ph ả p củ a ch úng t ờ i v ă n h ợ c đ ư c các ph ờ ng c ă c h th ố a m ă n b ằ ng c ă c s ử d ư ng các m ă u h ă n ch ế , bi ệ n m ỉ nh cho k ă n ă ng h ợ c t ậ p ít l ă n củ a kỹ thu ậ t nh ắc nh ờ. T ồ n b ộ các t ữ t ích c ự c và ti ề u c ự c, c ũ ng các l ờ i nh ắc th ử c ờ ng có k ă n ă ng h i ệ u qu ă k ă c đ ều đ ư c tr ình b ă y tr ờ ng t ă i li ệ u b ố s ư ng.

Cu ố i c ũ ng, chú th ích h ình ả nh h i ệ n t ă i th ư ờ ng l ă s ự th ậ t. T ư y nh i ề n, m ỗ i h ình ả nh tr ờ ng t ậ p d ữ li ệ u COCO đ ư c đ ă n nh ă n b ờ i n ă m chú th ích v i ề n, t ă t y ế u ch ữ a các m ô t ă c ă m x ứ c. Đ ể đ ă t đ ư c m ự c đ ị ch n ă y, ch úng t ờ i th ư th ă p các chú th ích t ích c ự c và ti ề u c ự c tr ờ ng t ậ p d ữ li ệ u COCO đ ể t ă o th ă nh các t ậ p h ợ p con k ờ ng có th ực, ch ữ a các t ữ t ích c ự c đ ư c x ă c đ ị nh tr ư ớ c nh ư "t uyệt , đ ẹp, đ ễ th ư ơ ng" và các t ữ ti ề u c ự c nh ư "x ă u x í, t ệ h ă i, k ỉ nh t ờ m". M ặ c dù các chú th ích k ờ ng có th ực n ă y r ấ t h i ể m, ph ư ơ ng ph ả p củ a ch úng t ờ i v ă n h ợ c đ ư c các ph ờ ng c ă c h th ố a m ă n b ằ ng c ă c s ử d ư ng các m ă u h ă n ch ế , bi ệ n m ỉ nh cho k ă n ă ng h ợ c t ậ p ít l ă n củ a kỹ thu ậ t nh ắc nh ờ. T ồ n b ộ các t ữ t ích c ự c và ti ề u c ự c, c ũ ng các l ờ i nh ắc th ử c ờ ng có k ă n ă ng h i ệ u qu ă k ă c đ ều đ ư c tr ình b ă y tr ờ ng t ă i li ệ u b ố s ư ng.

T ỉ nh ch ỉ nh m ô h ình. Tr ờ ng k h ướ n k ỏ củ a ch úng t ờ i, nh i ề u b ộ đ ă o t ă o đ ư c tr ộ n l ă n v ớ i nh ă u đ ể đ ă o t ă o m ô h ình th ờ ng nh ấ t. S ă v ớ i C ờ ng th ứ c (1), ch úng t ờ i d ự đ ă n m ă th ờ ng b ă o x t đ ư a tr ê n các t ỉ nh n ă ng tr ực qu ă n g(v), nh ứ ng m ă th ờ ng b ă o nh ắc nh ờ f(p) và nh ứ ng m ă th ờ ng b ă o tr ư ớ c đ ố f(x<t). D ữ li ệ u c ă c h đ i ề u k ă c nh ă u đ ư c g ă n v ớ i m ột l ờ i nh ắc c ự th ể nh ư m ỉ nh h ợ a tr ờ ng B ă ng 1. Tr ờ ng qu ă tr ình đ ă o t ă o, ch úng t ờ i th êm các l ờ i nh ắc th ử c ờ ng n ă y v ào các m ă th ờ ng b ă o chú th ích đ ư ớ i d ạng m ô t ă v ă n b ả n củ a h ình ả nh. Ch úng t ờ i l ă p r ă p các b ộ d ữ li ệ u c ă c h đ i ề u k ă c nh ă u đ ể c ũ ng đ ă o t ă o m ô h ình chú th ích b ằ ng c ă c s ử d ư ng LProLM m ấ t m ấ t LM đ ư a tr ê n l ờ i nh ắc nh ư s ă u:

$$\text{LProLM} = \mathbb{E}_{(\mathbf{v}, \mathbf{p}_1, \dots, \mathbf{x}) \in \mathcal{D}_1} \log P(x_t | g(\mathbf{v}), f(\mathbf{p}_1), f(\mathbf{x}_{<t})), \quad (3)$$

Kiểu chú thích	Hướng dẫn sử dụng nhắc nhở p
phong cách COCO	một hình ảnh bình thường hiển thị một hình
Kiểu TextCap	ảnh văn bản hiển thị
Ngắn dài	một bức ảnh có chú thích ngắn cho thấy
Độ dài trung bình:	một bức ảnh có chú thích trung bình cho thấy
Chiều dài cao	một bức ảnh có chú thích dài cho thấy
Tích cực	một bức tranh tích cực cho thấy
Tiêu cực	một bức tranh tiêu cực cho thấy

Bảng 1: Minh họa lời nhắc trong hướng dẫn sử dụng.

tr ờ ng đ ố p i bi ể u th ị l ờ i nh ắc th ử c ờ ng cho t ậ p d ữ li ệ u th ứ i D i. L ư u ý r ằ ng các m ă th ờ ng b ă o nh ắc p i và m ă th ờ ng b ă o chú th ích x<t ch i ă s ẽ c ũ ng m ột l ớ p ả nh x ă nh ứ ng f(·). Tr ờ ng k h ướ n k ỏ n ă y, ch úng t ờ i g i ữ ng ư ề n k i ế n tr úc m ô h ình c ơ s ớ k ờ ng th ă y đ ố i m ă k ờ ng có các k ỏ i có th ể h ợ c đ ư c b ố s ư ng, đ i ề u n ă y h i ệ u qu ă v ề m ặ t th ă m s ố .

3.3 Học tự động nhắc nhở Để tránh việc

th i ế t k ế nh ắc nh ờ th ử c ờ ng t ồ n c ờ ng s ứ c tr ờ ng Ph ă n 3.2, ch úng t ờ i k h ướ n k h ướ n t ự đ ộ ng h ợ c các nh ắc nh ờ th ề o c ă c h đ ă u cu ố i, nh ư th ể h i ệ n tr ờ ng H ị nh 2. V ớ i m ột ch ướ i các m ă th ờ ng b ă o nh ắc nh ờ th ử c ờ ng nh ư "m ột h ình ả nh v ă n b ả n cho th ấ y", tr ư ớ c ti ề n m ô h ình s ẽ ả nh x ă t ũ ng m ă th ờ ng b ă o th ă nh m ột ID s ố d ư y nh ấ t b ằ ng kỹ thu ậ t WordPiece. S ă u đ ố , đ ố i v ớ i m ô h ình đ ư a tr ê n BERT, các ID m ă th ờ ng b ă o đ ư c ch i ế u th ă nh các nh ứ ng t ữ 768-dim th ờ ng qu a l ớ p nh ứ ng m ă th ờ ng b ă o f(·) l ă m đ ă u v ào củ a m ô h ình h ợ p nh ấ t ng ờ n g ữ th ị g ă c, t ứ c l ă f(p) R N×768, tr ờ ng đ ố N bi ể u th ị đ ộ d ă i nh ắc nh ờ. Th ă y v ớ i th i ế t k ế nh ắc nh ờ th ử c ờ ng, ch úng t ờ i đ ề x ă u ă t h ợ c các nh ứ ng nh ắc nh ờ chú th ích P nh ư s ă u:

$$\mathbf{P} = [\mathbf{P}]_1 [\mathbf{P}]_2 \cdots [\mathbf{P}]_N, \quad (4)$$

tr ờ ng đ ố m ỗ i v ế c t ờ nh ứ ng [P]k (k = 1, ···, N) có c ũ ng ch i ế u v ớ i nh ứ ng t ữ. N ố i c ă c h k ă c, N×768 P R đ ố ng v ă i t ờ l ă m ột g ă i ả p th ă y th ể cho nh ứ ng nh ắc nh ờ th ử c ờ ng f(p). Tr ờ ng g ă i đ o ă n đ ă o t ă o, nh ứ ng nh ắc P đ ư c t ồ i ư u h ă a ch ư ng v ớ i m ă ng chú th ích nh ư s ă u:

$$\text{LAutoProLM} = \mathbb{E}_{(\mathbf{v}, \mathbf{x}) \in \mathcal{D}_i} \log P(x_t | g(\mathbf{v}), \mathbf{P}_i, f(\mathbf{x}_{<t})). \quad (5)$$

K h ướ ng đ ề x ă u ă t h ợ c các nh ứ ng nh ắc nh ờ c ự th ể P i cho m ỗ i t ậ p d ữ li ệ u D i c ự th ể th ề o mi ề n . Tr ờ ng qu ă tr ình đ ă o t ă o đ ă u cu ố i, các g r ă d i ệ n t có th ể đ ư c tr ư ề n ng ư ợ c h i ệ u qu ă đ ể t ồ i ư u h ă a các nh ứ ng nh ắc nh ờ. Đ ể đ ă t đ ư c m ự c đ ị ch n ă y, ng ư ờ i chú th ích có th ể k ă m ph ă đ ă y đ ủ các bi ể u di ể n nh ắc nh ờ ph ù h ợ p tr ờ ng k ờ ng g ă n nh ứ ng t ữ li ề n t ự c.

3.4 Suy luận dựa trên lời nhắc Sau khi

h ợ c l ờ i nh ắc, m ô h ình củ a ch úng t ờ i có th ể t ă o r ă nh i ề u chú th ích k ă c nh ă u b ằ ng c ă c s ử d ư ng các l ờ i nh ắc k ă c nh ă u. Tr ờ ng k h ướ n k ỏ l ờ i nh ắc th ử c ờ ng, s ă u k hi m ă h ă a m ă th ờ ng b ă o đ ặ c b i ệ t [BOS], ch úng t ờ i nh ứ ng t ă n t ự các m ă th ờ ng b ă o l ờ i nh ắc th ờ ng qu a f(p) và đ ư a ch úng v ào m ô h ình ng ờ n g ữ đ ể t ă o r ă chú th ích th ề o c ă c h t ự h ồ i qu y. Tr ờ ng k h ướ n k ỏ l ờ i nh ắc t ự đ ộ ng, ch úng t ờ i tr ực ti ế p n ố i nh ứ ng m ă th ờ ng b ă o củ a [BOS] và các nh ứ ng l ờ i nh ắc đ ă h ợ c P l ă m đ ă u v ào củ a m ô h ình ng ờ n g ữ.

By switching different prompts, the proposed captioner is able to generate a certain stylized caption.

4 Experiment

4.1 Datasets and Metrics

Pre-training Data. In the experiments, following our base-line approach (Li et al. 2022), we collect the image-text pairs from Visual Genome (Krishna et al. 2017), COCO (Lin et al. 2014), SBU Captions (Ordonez, Kulkarni, and Berg 2011), Conceptual Captions (Sharma et al. 2018), Conceptual 12M (Changpinyo et al. 2021), and a filtered version of LAION (115M images) (Schuhmann et al. 2021) to form the pre-training data. Following BLIP (Li et al. 2022), these data are filtered by a large model to form the high-quality bootstrapped dataset. In total, the pre-training corpus consists of about 129 million images.

Evaluation Datasets and Metrics. We evaluate the proposed method on the COCO caption dataset (Lin et al. 2014) of Karpathy split (Karpathy and Fei-Fei 2015), No-Caps (Agrawal et al. 2019), and TextCaps (Sidorov et al. 2020). To evaluate the quality of the generated captions, we use standard metrics in the image captioning task, including BLEU@4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016). In the inference stage, beam search (beam size = 3) is adopted in all experiments. More details and visualization results can be found in the supplementary material.

4.2 Implementation Details

Our model is implemented in Python with PyTorch. In the pre-training stage, the model is trained on 32 V100 GPUs. The image encoder is initialized from ViT-B/16 pre-trained on the ImageNet (Dosovitskiy et al. 2020), and the text encoder is initialized from BERT-base (Devlin et al. 2018). We pre-train the whole model for 32 epochs using a batch size of 2880. We use AdamW optimizer (Loshchilov and Hutter 2017) with a weight decay of 0.05. The learning rate is warmed-up to 3×10^{-4} and decayed linearly with a rate of 0.85. We take random image crops of resolution 224×224 during pre-training.

In the fine-tuning stage, we train the model using a small learning rate of 1×10^{-5} and linearly decay it. The model is fine-tuned for 5 epochs. Following previous works (Wang et al. 2021b), the image resolution is increased to 384×384 during fine-tuning. As for the prompt embedding $\mathbf{P} \in \mathbb{R}^{N \times 768}$, we randomly initialize it and set $N = 16$. We optimize our algorithm using standard cross-entropy loss *without* reinforcement learning. The proposed Controllable Captioner is denoted as ConCap.

4.3 Ablation Study

Manual Prompt v.s. w/o Prompt. Previous works such as BLIP utilize a pre-defined prompt “a picture of” to facilitate the caption generation. However, as shown in Table 2, in the zero-shot evaluation without model fine-tuning (① and ②), an empty prompt is even more effective. After downstream model fine-tuning (④ and ⑤), we observe that

Configuration	COCO Test		TextCaps Val	
	B@4	C	B@4	C
w/o Fine-tuning (Frozen Model)				
① w/o Prompt	33.9	106.6	18.6	48.6
② Manual Prompt (i.e., a picture of)	23.7	83.8	14.5	38.4
③ Learned Prompt ($N = 16$)	38.3	125.1	20.7	56.7
Multi-dataset Individual Training				
④ w/o Prompt	39.1	132.4	30.4	113.4
⑤ Manual Prompt (i.e., a picture of)	39.4	132.6	30.1	111.2
⑥ Learned Prompt ($N = 16$)	40.5	133.5	31.2	115.9
Multi-dataset Joint Training				
⑦ w/o Prompt	39.2	131.9	30.1	111.6
⑧ Shared Manual Prompt (i.e., a picture of)	39.3	132.2	30.0	110.4
⑨ Multi-prompt (Manual Prompts in Table 1)	39.6	132.8	30.7	113.5
⑩ Multi-prompt (Auto Learning, $N = 16$)	40.5	133.7	31.3	116.7

Table 2: Ablation comparisons on the COCO Karpathy test split (Lin et al. 2014) and TextCaps validation set (Sidorov et al. 2020), where B@4 and C denote BLEU@4 and CIDEr scores, respectively.

this hand-crafted prompt is beneficial to COCO dataset but harmful to TextCaps. These results show that the heuristic prompt is not always a good choice, which potentially requires the laborious manual design for different datasets.

Effectiveness of Learned Prompt. For a *frozen* image captioner, we only optimize the prompt embeddings in setting ③ in Table 2. The results show that learned prompt embeddings greatly unveil the potential of a pre-trained model with a good zero-shot performance of 125.1 CIDEr on COCO. After joint training of prompt embeddings and captioning model, the performance of the learned prompt is still superior to the manual prompt (⑥ v.s. ⑤).

Multi-dataset Individual Training v.s. Joint Training. Previous works typically train the model individually on different datasets. In setting ⑤ in Table 2, we separately fine-tune the image captioner on COCO and TextCaps. In setting ⑧ in Table 2, we merge the datasets of COCO and TextCaps, and use a *shared* prompt “a picture of” for both datasets. By analyzing the results of ⑤ and ⑧, we can observe that simply combining two diverse datasets with different styles will degrade the performance. This is consistent with common sense that data from diverse domains will challenge the model training.

Single Prompt v.s. Multi-prompt. In setting ⑨ of Table 2, we still combine the COCO and TextCaps to jointly train a unified captioner, but separate multi-domain data using different prompts. It is interesting that “multi-prompt for joint training” (⑨) not only outperforms the “single prompt for joint training” (⑧), but also surpasses “single prompt for individual training” (⑤), indicating that multiple (even manually designed) prompts can effectively separate the data from different domains. Furthermore, the most promising characteristic of “multi-prompt” is that we can control the caption style by feeding different prompts, which is infeasible for the “single prompt” setting. Finally, we encourage the model to jointly optimize multiple learnable prompt embeddings in an end-to-end manner (⑩), which achieves the best results.

Auto-prompt Length. We further validate the influence of

Machine Translated by Google

Bằng cách chuyển đổi các lời nhắc khác nhau, người viết chú thích đã được đề xuất là có khả năng tạo ra một chú thích cách điệu nhất định.

4 Thí nghiệm

4.1 Bộ dữ liệu và số liệu

Dữ liệu tiền đào tạo. Trong các thí nghiệm, theo phương pháp tiếp cận cơ bản của chúng tôi (Li et al. 2022), chúng tôi thu thập các cặp hình ảnh-văn bản từ Visual Genome (Krishna et al. 2017), COCO (Lin et al. 2014), Chú thích SBU (Ordonez, Kulkarni và Berg 2011), Chú thích khái niệm (Sharma et al. 2018), Khái niệm 12M (Changpinyo và cộng sự 2021) và phiên bản lọc của LAION (115M hình ảnh) (Schuhmann et al. 2021) để tạo thành dữ liệu tiền đào tạo.

Theo BLIP (Li et al. 2022), những dữ liệu này được lọc bởi một mô hình lớn để tạo thành tập dữ liệu khởi động chất lượng cao. Tổng cộng, tập dữ liệu tiền đào tạo bao gồm khoảng 129 triệu hình ảnh.

Bộ dữ liệu đánh giá và số liệu. Chúng tôi đánh giá phương pháp được đề xuất trên bộ dữ liệu chú thích COCO (Lin et al.

2014) của Karpathy chia tách (Karpathy và Fei-Fei 2015), No-Caps (Agrawal và cộng sự 2019) và TextCaps (Sidorov và cộng sự.

2020). Để đánh giá chất lượng của phụ đề được tạo, chúng tôi sử dụng các số liệu chuẩn trong nhiệm vụ chú thích hình ảnh, bao gồm BLEU@4 (Papineni và cộng sự 2002), METEOR (Banerjee và Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, và Parikh 2015) và SPICE (Anderson et al. 2016). Trong giai đoạn suy luận, tìm kiếm chùm tia (kích thước chùm tia = 3) được áp dụng trong tất cả thí nghiệm. Có thể có thêm thông tin chi tiết và kết quả trực quan có trong tài liệu bổ sung.

4.2 Chi tiết triển khai

Mô hình của chúng tôi được triển khai bằng Python với PyTorch. Trong Giai đoạn tiền đào tạo, mô hình được đào tạo trên 32 GPU V100. Bộ mã hóa hình ảnh được khởi tạo từ ViT-B/16 được đào tạo trước trên ImageNet (Dosovitskiy và cộng sự 2020) và bộ mã hóa văn bản được khởi tạo từ cơ sở BERT (Devlin và cộng sự 2018). Chúng tôi đào tạo trước toàn bộ mô hình trong 32 kỷ nguyên bằng cách sử dụng kích thước lô của 2880. Chúng tôi sử dụng trình tối ưu hóa AdamW (Loshchilov và Hutter 2017) với mức giảm trọng số là 0,05. Tỷ lệ học tập là được làm nóng lên đến 3×10^{-4} và phân rã tuyến tính với tốc độ 0.85. Chúng tôi cắt ảnh ngẫu nhiên có độ phân giải 224×224

trong quá trình đào tạo trước. Trong giai đoạn tinh chỉnh, chúng tôi đào tạo mô hình bằng cách sử dụng tốc độ học nhỏ 1×10^{-5} và phân rã tuyến tính. mô hình được tinh chỉnh cho 5 thời đại. Theo các công trình trước đó (Wang et al. 2021b), độ phân giải hình ảnh được tăng lên 384×384 trong quá trình tinh chỉnh. Đối với việc nhúng lời nhắc $\mathbf{P} \in \mathbb{R}^{N \times 768}$, chúng ta khởi tạo ngẫu nhiên và đặt $N = 16$. Chúng tôi tối ưu hóa thuật toán của mình bằng cách sử dụng mất mát entropy chéo tiêu chuẩn không có học tăng cường. Đề xuất Kiểm soát _____ Người chú thích được ký hiệu là ConCap.

4.3 Nghiên cứu cắt bỏ

Manual Prompt so với w/o Prompt. Các tác phẩm trước đây như BLIP sử dụng lời nhắc được xác định trước “một bức ảnh của” để tạo điều kiện thuận lợi cho việc tạo chú thích. Tuy nhiên, như được thể hiện trong Bảng 2, trong đánh giá zero-shot mà không cần tinh chỉnh mô hình (và), một lời nhắc trống thậm chí còn hiệu quả hơn. Sau điều chỉnh mô hình hạ lưu (và), chúng tôi quan sát thấy rằng

	COCO Kiểm tra B@4C@4C	TextCaps Val
Cấu hình		
không có tinh chỉnh (Mô hình đóng băng)		
không có	33,9 106,6 18	6 48,6
lời nhắc	Lời nhắc thủ công (tức là hình ảnh)	23,7 83,8 14,5 38,4
ảnh)	Lời nhắc đã học (N = 16)	38,3 125,1 20,7 56,7
Đào tạo cá nhân đa bộ dữ liệu		
không có	39,1 132,4 30,4	113,4
lời nhắc	Lời nhắc thủ công (tức là hình ảnh)	39,4 132,6 30,1 111,2
ảnh)	Lời nhắc đã học (N = 16)	40,5 133,5 31,2 115,9
Đào tạo chung nhiều tập dữ liệu		
không có	39.2 131.9 30.1	111,6
lời nhắc	Lời nhắc hướng dẫn chung (tức là hình ảnh)	39,3 132,2 30,0 110,4
Nhiều lời nhắc (Lời nhắc thủ công trong Bảng 1)	39,6 132,8 30,7 113,5	
Nhiều lời nhắc (Tự động học, N = 16)	40,5 133,7 31,3	116,7

Bảng 2: So sánh sự cắt bỏ trên thử nghiệm COCO Karpathy tách (Lin et al. 2014) và bộ xác thực TextCaps (Sidorov et al. 2020), trong đó B@4 và C biểu thị BLEU@4 và CIDEr điểm số tương ứng.

lời nhắc thủ công này có lợi cho tập dữ liệu COCO như ng có hại cho TextCaps. Những kết quả này cho thấy phương pháp tìm kiếm lời nhắc không phải lúc nào cũng là lựa chọn tốt, điều này có khả năng đòi hỏi phải thiết kế thủ công tốn nhiều công sức cho các tập dữ liệu khác nhau. Hiệu quả của lời nhắc đã học. Đối với người chú thích hình ảnh đóng băng, chúng tôi chỉ tối ưu hóa những lời nhắc trong cài đặt trong Bảng 2. Kết quả cho thấy các nhúng nhắc nhở đã học được tiết lộ rất nhiều tiềm năng của một mô hình được đào tạo trước với thành tích không bán phát nào tốt là 125,1 CIDEr trên COCO. Sau khi đào tạo chung về những nhanh và chú thích mô hình, hiệu suất của lời nhắc đã học vẫn vượt trội hơn lời nhắc thủ công (so với). Đào tạo cá nhân đa tập dữ liệu so với đào tạo chung. Các công trình trước đây thường đào tạo mô hình riêng lẻ trên các tập dữ liệu khác nhau. Trong thiết lập trong Bảng 2, chúng tôi tinh chỉnh riêng biệt chú thích hình ảnh trên COCO và TextCaps. Trong thiết lập trong Bảng 2, chúng tôi hợp nhất các tập dữ liệu của COCO và TextCaps và sử dụng lời nhắc chia sẻ “một bức ảnh của” cho cả hai tập dữ liệu. Bằng cách phân tích kết quả của và , chúng ta có thể lưu ý rằng việc chỉ kết hợp hai tập dữ liệu đa dạng với các phong cách khác nhau sẽ làm giảm hiệu suất. Điều này phù hợp với lẽ thường rằng dữ liệu từ các miền khác nhau sẽ thách thức mô hình đào tạo.

Lời nhắc đơn so với lời nhắc đa. Trong cài đặt của Bảng 2, chúng tôi vẫn kết hợp COCO và TextCaps để cùng nhau đào tạo một chú thích thống nhất, nhưng dữ liệu đa miền riêng biệt bằng cách sử dụng các lời nhắc khác nhau. Thật thú vị khi “nhiều lời nhắc cho đào tạo” () không chỉ vượt trội hơn “lời nhắc duy nhất cho “đào tạo chung” (), nhưng cũng vượt qua “lời nhắc duy nhất cho đào tạo cá nhân” (), cho thấy nhiều lời nhắc (thậm chí được thiết kế thủ công) có thể tách dữ liệu khỏi các miền khác nhau. Hơn nữa, đặc điểm hứa hẹn nhất của “multi-prompt” là chúng ta có thể kiểm soát chú thích phong cách bằng cách đưa ra các lời nhắc khác nhau, điều này là không khả thi đối với “thiết lập lời nhắc đơn”. Cuối cùng, chúng tôi khuyến khích mô hình cùng nhau tối ưu hóa nhiều nhúng nhắc nhở có thể học được trong một cách thức từ đầu đến cuối (), đạt được kết quả tốt nhất. Chiều dài tự động nhắc nhở. Chúng tôi xác nhận thêm ảnh hưởng của

	$N = 1$		$N = 4$		$N = 8$		$N = 16$		$N = 24$	
	B@4	C	B@4	C	B@4	C	B@4	C	B@4	C
TextCaps Val	30.7	114.5	30.8	115.4	31.1	115.4	31.3	116.7	31.5	115.9

Table 3: Ablation of the prompt embedding length N on the TextCaps validation set (Sidorov et al. 2020).

Prompt Style	B@4	M	C	S
Short-length Prompt	39.9	30.2	132.3	23.0
Medium-length Prompt	35.1	30.9	122.9	23.9
High-length Prompt	26.9	30.7	71.6	25.0
Positive Prompt	27.0	25.8	97.6	20.7
Negative Prompt	37.0	29.3	121.5	22.9
TextCap-style Prompt	22.1	25.9	66.0	20.5
COCO-style Prompt	40.5	30.9	133.7	23.8

Table 4: Performance comparisons of different prompts on the COCO Karpathy test split (Lin et al. 2014), where B@4, M, C, S denote BLEU@4, METEOR, CIDEr, and SPICE.

prompt embedding length N . In Table 3, the prompt embeddings of different lengths are randomly initialized. We test different lengths of $N = 1, 4, 8, 16, 24$ and observe that increasing the prompt embedding length N can consistently improve the performance. In our experiments, We choose $N = 16$ as it already yields saturated results.

Evaluation of Different Prompts on COCO. Finally, we evaluate the performance of different automatically learned prompts on the COCO Karpathy test split. The results are shown in Table 4. There is no doubt that ‘‘COCO-style Prompt’’ overall performs best, which leverages the entire training set of COCO for model fine-tuning. ‘‘TextCap-style Prompt’’ exhibits poor results on the COCO dataset, which justifies the domain gap between COCO and TextCaps datasets. Finally, it is interesting that CIDEr metric (Vedan-tam, Lawrence Zitnick, and Parikh 2015) prefers a short caption and ‘‘Short-length Prompt’’ is even comparable to the best ‘‘COCO-style Prompt’’ in CIDEr metric, while SPICE metric (Anderson et al. 2016) prefers a longer caption and ‘‘High-length Prompt’’ clearly outperforms the strong ‘‘COCO-style Prompt’’ in SPICE. Visualization results of different prompts are shown in the next Section 4.4.

4.4 Qualitative Evaluation

Results on COCO (Lin et al. 2014). In Figure 3, we exhibit the captioning results on the COCO dataset. By feeding different prompts, our ConCap method is able to generate diverse captions including COCO-style [■], Positive [■], Negative [■], Short-length [■], Medium-length [■], and High-length [■]. Besides, we observe that the percentage of emotional captions is only about 2% of the entire COCO dataset. The proposed ConCap merely utilizes limited positive or negative captions in COCO to learn such styles. This is consistent with the observation that prompt learning is suitable for few-shot domain transfer (Liu et al. 2021). As shown in Figure 3, our ConCap is able to briefly describe an image or in a more detailed manner. The high-length captions [■] produced by our ConCap are much longer than the ground-truth captions and yield additional

	
COCO-style Caption A yellow fire hydrant with a face drawn on it	COCO-style Caption An airplane flying over a field of trees
Positive Caption A yellow fire hydrant with a happy face drawn on it	Positive Caption A plane flying over a beautiful park with blooming trees
Negative Caption A yellow fire hydrant with a sad face drawn on it	Negative Caption A plane flying over a field of dead grass
Short-length Caption A fire hydrant with eyes drawn on it	Short-length Caption An airplane flying over a field of trees
Medium-length Caption A yellow fire hydrant with a face drawn on it	Medium-length Caption A plane flying over a field with trees in the foreground
High-length Caption A yellow fire hydrant with a face drawn on it in front of a large building	High-length Caption A plane flying over a field of trees with a building in the background and mountains in the distance
Ground-truth Caption A yellow fire hydrant with a couple of eyes drawing on it	Ground-truth Caption An airplane that is flying in the sky

Figure 3: Image captioning examples from COCO (Karpathy and Fei-Fei 2015) with different styles including COCO-style [■], Positive [■], Negative [■], Short-length [■], Medium-length [■] and High-length [■].

meaningful semantics, e.g., ‘‘a large building’’ in the first image and ‘‘mountains in the distance’’ in the second image. Furthermore, our approach generates the positive words such as ‘‘happy, beautiful’’ [■] or the negative words such as ‘‘sad, dead’’ [■] to describe the same image in opposite personality traits. Since the COCO-caption dataset rarely contains the image with OCR contents, we showcase the results of ‘‘TextCap-style Prompt’’ on the TextCaps dataset in Figure 4.



	
COCO-style Caption A group of trash cans sitting in front of a building	COCO-style Caption A pink bus driving down a street next to a tall building
TextCap-style Caption A sign on a building that says heart break	TextCap-style Caption A pink bus that says target travel on the front
Ground-truth Caption A sign is painted with a broken heart and a scroll that says Heartbreak	Ground-truth Caption A pink bus has Target Travel painted on it in several locations

Figure 4: Image captioning examples from TextCaps dataset (Sidorov et al. 2020) with different styles including COCO-style [■] and TexCap-style [■]. Best view in zoom in.

Results on TextCaps (Sidorov et al. 2020). Figure 4 exhibits the results on the TextCaps dataset, where we show the COCO-style [■] and TextCap-style [■] captions for style comparison. Different styles focus on different aspects of the image. For example, in the first image, the TextCap-style caption as well as the ground-truth annotation aim to de-

Translated by Google

	N = 1	N = 4	N = 8	N = 16	N = 24		
	B@4	CB@4	CB@4	CB@4	CB@4	C	
Giá trị văn bản	30,7	114,5	30,8	115,4	31,1	115,4	31,3 116,7 31,5 115,9

Bảng 3: Sự cất bỏ chiều dài nhúng nhanh N trên

Bộ xác thực TextCaps (Sidorov và cộng sự, 2020).



	B@4 MCS	Chủ thích theo phong cách COCO
Phong cách nhắc nhở		Một chiếc máy bay đang bay qua một cánh đồng có cỏ phủ đầy
Yêu cầu ngắn	39,9	30,2 132,3 23,0
Yêu cầu độ dài trung bình	35,1	30,9 122,9 23,0
Yêu cầu dài	26,9	30,7 122,9 23,0
Lời nhắc tích cực	27,0	25,8 97,6 20,7
Lời nhắc tiêu cực	37,0	29,3 121,5 22,9
Đầu nhắc theo kiểu TextCap	22,1	25,9 66,0 20,5
Gợi ý theo phong cách COCO	40,5	30,9 133,7 23,8

Bảng 4: So sánh hiệu suất của các lời nhắc khác nhau trên phép chia tách thử nghiệm COCO Karpathy (Lin et al. 2014), trong đó B@4, M, C, S biểu thị BLEU@4, METEOR, CIDEr và SPICE.

chiều dài nhúng nhắc nhở N. Trong Bảng 3, các đầu nhắc nhúng có chiều dài khác nhau đư ợc khởi tạo ngẫu nhiên. Chúng tôi kiểm tra các độ dài khác nhau của N = 1, 4, 8, 16, 24 và quan sát thấy rằng việc tăng độ dài nhúng nhắc nhở N có thể liên tục cải thiện hiệu suất. Trong các thí nghiệm của chúng tôi, chúng tôi chọn N = 16 vì nó đã mang lại kết quả bảo hòa. Đánh giá các lời nhắc khác nhau trên COCO. Cuối cùng, chúng tôi đánh giá hiệu suất của các phư ơng pháp học tự động khác nhau nhắc nhở về bài kiểm tra COCO Karpathy chia tách. Kết quả là đư ợc thể hiện trong Bảng 4. Không còn nghi ngờ gì nữa rằng ‘‘phong cách COCO Nhìn chung, ‘‘Prompt’’ hoạt động tốt nhất, tận dụng toàn bộ bộ ạo tạo của COCO để tinh chỉnh mô hình. ‘‘TextCap-style ‘‘Nhắc nhở’’ cho thấy kết quả kém trên tập dữ liệu COCO, biện minh cho khoảng cách miễn giữa COCO và TextCaps bộ dữ liệu. Cuối cùng, điều thú vị là số liệu CIDEr (Vedan-tam, Lawrence Zitnick và Parikh 2015) thích chú thích ngắn và ‘‘Lời nhắc ngắn’’ thậm chí còn tư ơng đư ơng đư ơng với ‘‘Lời nhắc theo phong cách COCO’’ tốt nhất theo số liệu của CIDEr, trong khi SPICE số liệu (Anderson et al. 2016) thích chú thích dài hơn và ‘‘Lời nhắc dài’’ rõ ràng vư ợt trội hơn ‘‘Lời nhắc theo phong cách COCO’’ trong SPICE. Kết quả trực quan hóa của các lời nhắc khác nhau đư ợc hiển thị trong Phần 4.4 tiếp theo.

4.4 Đánh giá định tính

Kết quả trên COCO (Lin et al. 2014). Trong Hình 3, chúng tôi trình bày kết quả chú thích trên tập dữ liệu COCO. Bằng cách đư a ra các lời nhắc khác nhau, phư ơng pháp ConCap của chúng tôi có thể tạo ra nhiều chú thích khác nhau bao gồm [■] theo phong cách COCO, Positive [■], Tiêu cực [■], Ngắn [■], Trung bình [■], và Độ dài cao [■]. Bên cạnh đó, chúng tôi quan sát thấy rằng tỷ lệ phần trăm độ tuổi của phư ơng pháp chỉ chiếm khoảng 2% tổng số Bộ dữ liệu COCO. ConCap đư ợc đề xuất chỉ sử dụng các chú thích tích cực hoặc tiêu cực hạn chế trong COCO để tìm hiểu như vậy phong cách. Điều này phù hợp với quan sát rằng nhắc nhở học tập phù hợp với việc chuyển giao miễn it lần (Liu et al. 2021). Như thể hiện trong Hình 3, ConCap của chúng tôi có thể tạm thời mô tả một hình ảnh hoặc theo cách chi tiết hơn. Các chú thích dài [■] do ConCap của chúng tôi tạo ra có nhiều dài hơn các chú thích thực tế và mang lại thêm

	
Chủ thích theo phong cách COCO Một vòi cứu hỏa màu vàng có hình mặt đư ợc vẽ trên đó	Chủ thích theo phong cách COCO Một chiếc máy bay đang bay qua một cánh đồng cây
Chủ thích tích cực Một vòi cứu hỏa màu vàng có hình mặt cu ời đư ợc vẽ trên đó	Chủ thích tích cực Một chiếc máy bay đang bay qua một cánh vườn xinh đẹp với những hàng cây ra hoa
Tiêu đề tiêu cực Một vòi cứu hỏa màu vàng với khuôn mặt buồn đư ợc vẽ trên đó	Tiêu đề tiêu cực Một chiếc máy bay đang bay qua một cánh đồng có chết
Chủ thích ngắn Một vòi cứu hỏa có vẽ mặt	Chủ thích ngắn Một chiếc máy bay đang bay qua một cánh đồng cây
Chủ thích có độ dài trung bình Một vòi cứu hỏa màu vàng có hình mặt đư ợc vẽ trên đó	Chủ thích có độ dài trung bình Một chiếc máy bay đang bay qua một cánh đồng có cây cối ở phía trư ớc
Chủ thích dài Một vòi cứu hỏa màu vàng có hình mặt ngư ời đư ợc vẽ ở phía trư ớc một tòa nhà lớn	Chủ thích dài Một chiếc máy bay bay qua một cánh đồng cây với một tòa nhà ở phía sau và những ngọn núi ở đằng xa
Chủ thích thực tế Một vòi cứu hỏa màu vàng với một vãi con mặt vẽ trên đó	Chủ thích thực tế Một chiếc máy bay đang bay trên bầu trời

Hình 3: Ví dụ chú thích hình ảnh từ COCO (Karpa-thy và Fei-Fei 2015) với các phong cách khác nhau bao gồm phong cách COCO [■], Tích cực [■], Tiêu cực [■], Độ dài ngắn [■], Độ dài trung bình [■] và Độ dài cao [■].

ngữ nghĩa có ý nghĩa, ví dụ, ‘‘một tòa nhà lớn’’ trong hình ảnh đầu tiên và ‘‘những ngọn núi ở đằng xa’’ trong hình ảnh thứ hai. Hơn nữa, cách tiếp cận của chúng tôi tạo ra những từ tích cực như ‘‘vui vẻ, xinh đẹp’’ [■] hoặc những từ ngữ tiêu cực như ‘‘buồn, chết’’ [■] để mô tả cùng một hình ảnh trong các đặc điểm tính cách đối lập. Vì tập dữ liệu chú thích COCO hiếm khi chứa hình ảnh có nội dung OCR, chúng tôi trình bày kết quả của ‘‘TextCap-style Prompt’’ trên tập dữ liệu TextCaps trong Hình 4.

		
Chủ thích theo phong cách COCO Một nhóm thùng rác đặt trư ớc một tòa nhà	Chủ thích theo phong cách COCO Một chiếc xe buýt màu hồng chạy xuống phố bên cạnh một tòa nhà cao tầng	Kiểu COCO C
Chủ thích theo kiểu TextCap Một biển báo trên tòa nhà ghi trái tim tan vỡ	Chủ thích theo kiểu TextCap Một chiếc xe buýt màu hồng có dòng chữ tiêu đi chuyển ở phía trước	Kiểu TextCap
Chủ thích thực tế Một tấm biển đư ợc vẽ hình trái tim tan vỡ và một cuộn giấy ghi chữ Heartbreak	Chủ thích thực tế Một chiếc xe buýt màu hồng có dòng chữ Target Travel đư ợc sơn ở nhiều vị trí	Kiểu TextCap ai
		Sự thật thực tế
		Máy bay màu trắng ở bên cạnh

Hình 4: Ví dụ chú thích hình ảnh từ tập dữ liệu TextCaps (Sidorov et al. 2020) với nhiều kiểu khác nhau bao gồm kiểu COCO [■] và kiểu TexCap [■]. Xem tốt nhất khi phóng to.

Kết quả trên TextCaps (Sidorov và cộng sự 2020). Hình 4 minh họa kết quả trên tập dữ liệu TextCaps, trong đó chúng tôi hiển thị Tiêu đề theo kiểu COCO [■] và kiểu TextCap [■] cho kiểu so sánh. Các phong cách khác nhau tập trung vào các khía cạnh khác nhau của hình ảnh. Ví dụ, trong hình ảnh đầu tiên, kiểu TextCap chú thích cũng như chú thích thực tế nhằm mục đích xác định

Method	Pre-training Data	COCO Caption				NoCaps Validation							
		Karpathy Test				In-domain		Near-domain		Out-domain		Overall	
		B@4	M	C	S	C	S	C	S	C	S	C	S
BUTD (Anderson et al. 2018)	N/A	36.2	27.0	113.5	20.3	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
AoANet (Huang et al. 2019)	N/A	37.2	28.4	119.8	21.3	-	-	-	-	-	-	-	-
X-LAN (Pan et al. 2020)	N/A	38.2	28.8	122.0	21.9	-	-	-	-	-	-	-	-
Oscar _{base} (Li et al. 2020)	7M	36.5	30.3	123.7	23.1	83.4	12.0	81.6	12.0	77.6	10.6	81.1	11.7
ViTCAP (Fang et al. 2021)	10M	36.3	29.3	125.2	22.6	98.7	13.3	92.3	13.3	95.4	12.7	93.8	13.0
VinVL _{base} (Zhang et al. 2021)	9M	38.2	30.3	129.3	23.6	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5
LEMON _{base} (Hu et al. 2021)	200M	40.3	30.2	133.3	23.3	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1
BLIP _{base} (Li et al. 2022)	129M	39.7	-	133.3	23.3	111.8	14.9	108.6	14.8	111.5	14.2	109.6	14.7
SimVLM _{base} (Wang et al. 2021b)	1.8B	39.0	32.9	134.8	24.0	-	-	-	-	-	-	94.8	13.1
ConCap (Ours)	129M	40.5	30.9	133.7	23.8	113.4	14.9	108.4	14.6	113.2	14.4	110.2	14.8

Table 5: Performance comparisons on the COCO Karpathy test split (Lin et al. 2014) and NoCaps validation split (Agrawal et al. 2019), where B@4, M, C, S denote BLEU@4, METEOR, CIDEr, and SPICE scores. For a fair comparison, all the methods only adopt the standard cross-entropy without CIDEr optimization.

scribe the words in the sign (e.g., “heart break”) while ignoring the objects such as “trash cans”. In contrast, the COCO-style pays more attention to the objects and environment, e.g., “tall building” in the second image.

4.5 Quantitative Evaluation

COCO (Lin et al. 2014). In Table 5, we present the performance of state-of-the-art captioning methods on the COCO-caption Karpathy test split (Karpathy and Fei-Fei 2015). Compared with the recent LEMON (Hu et al. 2021) that leverages more pre-training data, our method achieves superior performance. BLIP (Li et al. 2022) can be regarded as the baseline of our approach. Compared with BLIP, our ConCap outperforms it on all metrics. More importantly, the proposed ConCap is able to simultaneously handle other domains and generate captions with different lengths and styles for each image, which is infeasible for BLIP. The recent SimVLM approach (Wang et al. 2021b) leverages a large-scale pre-training corpus including 1.8 billion image-text pairs, which is 10× larger than ours. Besides, SimVLM combines the ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2020) models as the visual extractor, which is stronger than our pure ViT structure.

NoCaps (Agrawal et al. 2019). NoCaps dataset covers more than 600 object categories and nearly 2/3 of them are unseen from the training set in COCO. The images in NoCaps are categorized into in-domain, near-domain, and out-of-domain based on whether these images are seen in the COCO training set. On this benchmark, we evaluate our ConCap using the “COCO-style” prompt. As shown in Table 5, the proposed ConCap outperforms all existing methods in terms of the overall performance, which verifies the generalizability of our method.

TextCaps (Sidorov et al. 2020). TextCaps is a recently proposed dataset containing 28K images and 145K captions, which is more challenging than COCO due to the existence of complex textual words. We compare the proposed method with the classic captioner such as AoANet (Huang et al. 2019) and the recent state-of-the-art methods including MMA-SR (Wang, Tang, and Luo 2020), CNMT (Wang et al. 2021a), and TAP (Yang et al. 2021).

Method	OCR Input	Validation		Test	
		B@4	C	B@4	C
BUTD (Anderson et al. 2018)	✗	20.1	41.9	14.9	33.8
AoANet (Huang et al. 2019)	✗	20.4	42.7	15.9	34.6
M4C (Hu et al. 2020)	✓	23.3	89.6	18.9	81.0
MMA-SR (Wang, Tang, and Luo 2020)	✓	24.6	98.0	19.8	88.0
CNMT (Wang et al. 2021a)	✓	24.8	101.7	20.0	93.0
TAP (Yang et al. 2021)	✓	25.8	109.2	21.9	103.2
ConCap (Ours)	✗	31.3	116.7	27.4	105.6

Table 6: Comparision results on the TextCaps validation set and test set (Sidorov et al. 2020), where B@4 and C denote BLEU@4 and CIDEr scores, respectively.

The comparison results are shown in Table 6. Our approach significantly outperforms the classic methods without pre-training such as AoANet (Huang et al. 2019). To the best of our knowledge, TAP (Yang et al. 2021) represents the current performance leader on the TextCaps dataset. TAP approach collects high-quality OCR-based image-text pre-training data, and performs the text-aware pre-training. Besides, TAP feeds the OCR detection results to the model, while our approach is free of such necessity. Without knowing the OCR results, our approach still surpasses the current state-of-the-art TAP method by a large margin of 7.5 CIDEr on the validation set. It is worth noting that our ConCap is not specially designed for TextCaps and is able to perform well on multiple domains including COCO, NoCaps, and TextCaps using a single model.

5 Conclusion

In this paper, we propose a conceptually simple yet effective prompt-based image captioning framework, which has been rarely investigated in the captioning community. By prompt engineering, the proposed approach is able to generate captions with diverse styles. To further explore the potential of prompt learning, we encourage the network to automatically learn the suitable prompt vectors in the continuous word embedding space. Extensive qualitative and quantitative experiments verify the effectiveness of the proposed framework.

Phư ơ ng pháp	Dữ liệu tiên đào tạo	Chú thích COCO		Xác thực NoCaps							
		Kiểm tra Karpathy		Trong miền		Gần miền		Ngoài miền		Tổng thể	
		B@4	MC	S	C	S	C	S	C	S	C
BUTD (Anderson và cộng sự 2018)	không có	36,2	27,0	113,5	20,3	80,0		12,0	73,6	11,3	66,4
AoANet (Huang và cộng sự 2019)	không có	37,2	28,4	119,8	21,3			-	-	-	-
X-LAN (Pan và cộng sự 2020)	không có	38,2	28,8	122,0	21,9			-	-	-	-
Oscarbase (Li và cộng sự, 2020)	7 phút	36,5	30,3	123,7	23,1	83,4	12,0	81,6	12,0	77,6	10,6
ViTCAP (Fang và cộng sự 2021)	10 triệu	36,3	29,3	125,2	22,6	98,7	13,3	92,3	13,3	95,4	12,7
VinVLbase (Zhang và cộng sự, 2021)	9M	38,2	30,3	129,3	23,6	103,1	14,2	96,1	13,8	88,3	12,1
LEMONbase (Hu và cộng sự 2021)	200 triệu	40,3	30,2	133,3	23,3	107,7	14,7	106,2	14,3	107,9	13,1
BLIPbase (Li và cộng sự, 2022)	129 triệu	39,7	-	133,3	23,3	111,8	14,9	108,6	14,8	111,5	14,2
SimVLMbase (Wang và cộng sự 2021b)	1,8 tỷ	39,0	32,9	134,8	24,0	-	-	-	-	-	94,8
ConCap (của chúng tôi)	129 triệu	40,5	30,9	133,7	23,8	113,4	14,9	108,4	14,6	113,2	14,4

Bảng 5: So sánh hiệu suất giữa phép chia kiểm tra COCO Karpathy (Lin et al. 2014) và phép chia xác thực NoCaps (Agrawal et al. 2019), trong đó B@4, M, C, S biểu thị điểm BLEU@4, METEOR, CIDEr và SPICE. Để so sánh công bằng, tất cả các phư ơ ng pháp chỉ áp dụng entropy chéo tiêu chuẩn mà không có tối ưu hóa CIDEr.

viết những từ trong biển báo (ví dụ, “trái tim tan vỡ”) trong khi bỏ qua các đối tượng như “thùng rác”. Ngư ợc lại, phong cách COCO chú ý nhiều hơn đến các đối tượng và môi trường, ví dụ, “tòa nhà cao tầng” trong hình ảnh thứ hai.

4.5 Đánh giá định lượng

COCO (Lin et al. 2014). Trong Bảng 5, chúng tôi trình bày hiệu suất của các phư ơ ng pháp chú thích hiện đại trên phép chia tách thử nghiệm COCO-caption Karpathy (Karpathy và Fei-Fei 2015).

So với LEMON gần đây (Hu et al. 2021) rằng tận dụng nhiều dữ liệu tiên đào tạo hơn n, phư ơ ng pháp của chúng tôi đạt đư ợc hiệu suất vư ợt trội. BLIP (Li et al. 2022) có thể đư ợc coi là như là cơ sở của cách tiếp cận của chúng tôi. So với BLIP, ConCap vư ợt trội hơn n nó về mọi mặt. Quan trọng hơn n, ConCap đư ợc đề xuất có thể xử lý đồng thời các miền và tạo chú thích với độ dài khác nhau và kiểu cho mỗi hình ảnh, điều này là không khả thi đối với BLIP. cách tiếp cận SimVLM gần đây (Wang et al. 2021b) tận dụng ngữ liệu tiên đào tạo quy mô lớn bao gồm 1,8 tỷ cặp hình ảnh-văn bản, lớn hơn 10 lần so với của chúng tôi. Bên cạnh đó, SimVLM kết hợp ResNet (He et al. 2016) và ViT (Dosovitskiy et al. 2020) mô hình như trích xuất trực quan, mạnh hơn hơn cấu trúc ViT tinh khiết của chúng tôi.

NoCaps (Agrawal và cộng sự 2019). Bộ dữ liệu NoCaps bao gồm nhiều hơn hơn n 600 danh mục đối tượng và gần 2/3 trong số đó không đư ợc nhìn thấy từ tập huấn luyện trong COCO. Các hình ảnh trong No-Caps đư ợc phân loại thành trong miền, gần miền và ngoài miền dựa trên việc những hình ảnh này có đư ợc nhìn thấy trong Bộ đào tạo COCO. Trên chuẩn mực này, chúng tôi đánh giá ConCap sử dụng lời nhắc “kiểu COCO”. Như đư ợc thể hiện trong Bảng 5, ConCap đư ợc đề xuất vư ợt trội hơn n tất cả các phư ơ ng pháp hiện có về hiệu suất tổng thể, điều này xác minh

khả năng khái quát hóa phư ơ ng pháp của chúng tôi. TextCaps (Sidorov và cộng sự 2020). TextCaps là một tập dữ liệu đư ợc đề xuất gần đây chứa 28K hình ảnh và 145K chú thích, cái này khó hơn n COCO do có sự tồn tại của các từ ngữ phức tạp trong văn bản. Chúng tôi so sánh đề xuất phư ơ ng pháp với chú thích cổ điển như AoANet (Huang et al. 2019) và các phư ơ ng pháp tiên tiến gần đây bao gồm MMA-SR (Wang, Tang và Luo 2020), CNMT (Wang et al. 2021a) và TAP (Yang và cộng sự, 2021).

Phư ơ ng pháp	Xác thực OCR		Đầu vào
	Đầu vào	B@4	CB@4
BUTD (Anderson và cộng sự 2018)	20,1	41,9	14,9
AoANet (Huang và cộng sự 2019)	20,4	42,7	15,9
M4C (Hu và cộng sự 2020)	23,3	89,6	18,9
MMA-SR (Wang, Tang và Luo 2020)	24,6	98,0	19,8
CNMT (Wang và cộng sự 2021a)	24,8	101,7	20,0
TAP (Yang và cộng sự, 2021)	25,8	109,2	21,9
ConCap (của chúng tôi)	31,3	116,7	27,4

Bảng 6: Kết quả so sánh trên bộ xác thực TextCaps và bộ kiểm tra (Sidorov et al. 2020), trong đó B@4 và C biểu thị Điểm BLEU@4 và CIDEr tương ứng.

Kết quả so sánh đư ợc thể hiện trong Bảng 6. Phư ơ ng pháp tiếp cận của chúng tôi vư ợt trội hơn đáng kể so với các phư ơ ng pháp cổ điển không có đào tạo trư ớc như AoANet (Huang et al. 2019). Đối với theo hiểu biết của chúng tôi, TAP (Yang et al. 2021) đại diện cho ngư ời dẫn đầu về hiệu suất hiện tại trên tập dữ liệu TextCaps. TAP phư ơ ng pháp tiếp cận thu thập dữ liệu đào tạo trư ớc hình ảnh-văn bản dựa trên OCR chất lượng cao và thực hiện đào tạo trư ớc nhận biết văn bản. Bên cạnh đó, TAP cung cấp kết quả phát hiện OCR cho mô hình, trong khi cách tiếp cận của chúng tôi không cần thiết như vậy. Nếu không biết kết quả OCR, cách tiếp cận của chúng tôi vẫn vư ợt trội hơn cách tiếp cận hiện tại phư ơ ng pháp TAP tiên tiến nhất với biên độ lớn là 7,5 CIDEr trên bộ xác thực. Điều đáng chú ý là ConCap của chúng tôi là không đư ợc thiết kế đặc biệt cho TextCaps và có thể thực hiện tốt trên nhiều miền bao gồm COCO, NoCaps và TextCaps sử dụng một mô hình duy nhất.

5 Kết luận

Trong bài báo này, chúng tôi đề xuất một khái niệm đơn giản như ng hiệu quả khung chú thích hình ảnh dựa trên lời nhắc, đã đư ợc hiếm khi đư ợc điều tra trong cộng đồng chú thích. Theo lời nhắc kỹ thuật, phư ơ ng pháp đề xuất có thể tạo ra các chú thích với nhiều phong cách khác nhau. Để khám phá thêm tiềm năng của học nhanh, chúng tôi khuyến khích mạng tự động tìm hiểu các vectơ nhắc nhở phù hợp trong không gian nhúng từ liên tục. Các thí nghiệm định tính và định lượng mở rộng xác minh tính hiệu quả của khuôn khổ đư ợc đề xuất.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *ICCV*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020a. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*.

Chen, T.; Zhang, Z.; You, Q.; Fang, C.; Wang, Z.; Jin, H.; and Luo, J. 2018. “Factual”or“Emotional”: Stylized Image Captioning with Adaptive Learning and Attention. In *ECCV*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Universal image-text representation learning. In *ECCV*.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-memory transformer for image captioning. In *CVPR*.

Deng, C.; Ding, N.; Tan, M.; and Wu, Q. 2020. Length-controllable image captioning. In *ECCV*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Liu, Z.; Zeng, M.; et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387*.

Fang, Z.; Wang, J.; Hu, X.; Liang, L.; Gan, Z.; Wang, L.; Yang, Y.; and Liu, Z. 2021. Injecting semantic concepts into end-to-end image captioning. *arXiv preprint arXiv:2112.05230*.

Fei, Z. 2022. Attention-Aligned Transformer for Image Captioning. In *AAAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2021. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *ICCV*.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *AAAI*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *TACL*, 8: 423–438.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kobus, C.; Crego, J.; and Senellart, J. 2016. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1): 32–73.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; and Ji, R. 2021. Dual-level collaborative transformer for image captioning. In *AAAI*.

Mathews, A.; Xie, L.; and He, X. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-linear attention networks for image captioning. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Machine Translated by Google

Tài liệu tham khảo

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; và Anderson, P. 2019. Nocaps: Chú thích đối tượng mới lạ ở quy mô lớn. Trong ICCV.

Anderson, P.; Fernando, B.; Johnson, M.; và Gould, S. 2016. Spice: Đánh giá chú thích hình ảnh mệnh đề ngữ nghĩa. Trong ECCV.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; và Zhang, L. 2018. Sự chú ý từ dưới lên và từ trên xuống để chú thích hình ảnh và trả lời câu hỏi trực quan. Trong CVPR.

Banerjee, S.; và Lavie, A. 2005. METEOR: Một phép đo tự động để đánh giá MT với mỗi tư ơng quan đư ợc cải thiện với các phán đoán của con ngư ời. Trong Hội thảo ACL.

Changpinyo, S.; Sharma, P.; Ding, N.; và Soricut, R. 2021. Khái niệm 12m: Đẩy mạnh quá trình đào tạo trư ớc hình ảnh-văn bản ở quy mô web để nhận dạng các khái niệm trực quan đư ời dài. Trong CVPR.

Chen, S.; Jin, Q.; Wang, P.; và Wu, Q. 2020a. Nói theo ý bạn muốn: Kiểm soát chi tiết việc tạo chú thích hình ảnh bằng đồ thị cảnh trư ờng. Trong CVPR.

Chen, T.; Zhang, Z.; You, Q.; Fang, C.; Wang, Z.; Jin, H.; và Luo, J. 2018. “Thực tế” hoặc “Cảm xúc”: Chú thích hình ảnh cách điệu với Học tập thích ứng và Chú ý. Trong ECCV.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; và Liu, J. 2020b. Uniter: Học tập biểu diễn-phân đối hình ảnh-văn bản phổ quát. Trong ECCV.

Cornia, M.; Baraldi, L.; và Cucchiara, R. 2019. Hiện thị, kiểm soát và kẻ: Một khuôn khổ để tạo chú thích có thể kiểm soát và có căn cứ. Trong CVPR.

Cornia, M.; Stefanini, M.; Baraldi, L.; và Cucchiara, R. 2020. Bộ chuyển đổi bộ nhớ dạng lưu ời để chú thích hình ảnh. Trong CVPR.

Đặng, C.; Đình, N.; Tân, M.; và Wu, Q. 2020. Chú thích hình ảnh có thể kiểm soát độ dài. Trong ECCV.

Devlin, J.; Chang, M.-W.; Lee, K.; và Toutanova, K. 2018. Bert: Đào tạo trư ớc các bộ biến đổi song hư ớng sâu để hiểu ngôn ngữ. Bản in trư ớc arXiv arXiv:1810.04805.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. Một hình ảnh có giá trị bằng 16x16 từ: Bộ biến đổi để nhận dạng hình ảnh theo tỷ lệ. Bản in trư ớc arXiv arXiv:2010.11929.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Liu, Z.; Zeng, M.; et al. 2021. Một nghiên cứu thực nghiệm về đào tạo các bộ chuyển đổi ngôn ngữ và thị giác đầu cuối. Bản in trư ớc arXiv arXiv:2111.02387.

Fang, Z.; Vũ ơng, J.; Hu, X.; Lưu ơng, L.; Cẩm, Z.; Vũ ơng, L.; Dư ơng, Y.; và Lưu, Z. 2021. Đưa a các khái niệm ngữ nghĩa vào chú thích hình ảnh từ đầu đến cuối. bản in trư ớc arXiv arXiv:2112.05230.

Fei, Z. 2022. Bộ chuyển đổi căn chỉnh sự chú ý để chú thích hình ảnh. Trong AAAI.

He, K.; Zhang, X.; Ren, S.; và Sun, J. 2016. Học sâu dư thừa để nhận dạng hình ảnh. Trong CVPR.

Hu, R.; Singh, A.; Darrell, T.; và Rohrbach, M. 2020. Dự đoán câu trả lời lặp đi lặp lại với bộ chuyển đổi đa phư ơng thức tăng cường con trỏ cho textvqa. Trong CVPR.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; và Wang, L. 2021. Mở rộng quy mô đào tạo trư ớc ngôn ngữ thị giác để chú thích hình ảnh. Bản in trư ớc arXiv arXiv:2111.12233.

Huang, L.; Wang, W.; Chen, J.; và Wei, X.-Y. 2019. Chú ý đến sự chú ý cho chú thích hình ảnh. Trong ICCV.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; và Fu, J. 2021. Nhìn ra ngoài khuôn khổ: Đào tạo trư ớc toàn diện cho việc học biểu diễn ngôn ngữ thị giác. Trong CVPR.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; và Ji, R. 2021. Cải thiện chú thích hình ảnh bằng cách tận dụng biểu diễn toàn cầu trong và giữa các lớp trong mạng biến áp. Trong AAAI.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, QV; Sung, Y.; Li, Z.; và Duerig, T. 2021. Mở rộng quy mô học biểu diễn ngôn ngữ thị giác và thị giác với giám sát văn bản nhiều. Trong ICML.

Jiang, Z.; Xu, FF; Araki, J.; và Neubig, G. 2020. Làm sao chúng ta có thể biết đư ợc các mô hình ngôn ngữ biết những gì? TACL, 8: 423-438.

Karpathy, A.; và Fei-Fei, L. 2015. Sự liên kết ngữ nghĩa-hình ảnh sâu sắc để tạo ra mô tả hình ảnh. Trong CVPR.

Kobus, C.; Crego, J.; và Senellart, J. 2016. Kiểm soát miền cho dịch máy thần kinh. Bản in trư ớc arXiv arXiv:1612.06140.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, DA; et al. 2017. Bộ gen thị giác: Kết nối ngôn ngữ và thị giác bằng cách sử dụng chú thích hình ảnh dày đặc do cộng đồng đóng góp. IJCV, 123(1): 32-73.

Lester, B.; Al-Rfou, R.; và Constant, N. 2021. Sức mạnh của quy mô để điều chỉnh nhanh chóng hiệu quả theo tham số. Bản in trư ớc arXiv arXiv:2104.08691.

Li, J.; Li, D.; Xiong, C.; và Hoi, S. 2022. BLIP: Khởi động quá trình đào tạo trư ớc ngôn ngữ-hình ảnh để hiểu và tạo ngôn ngữ-thị giác thống nhất. Bản in trư ớc arXiv arXiv:2201.12086.

Lý, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; và Hội, S. CH 2021. Căn chỉnh trư ớc khi kết hợp: Học biểu diễn ngôn ngữ và thị giác với chú ng cắt động lưu ợng. Trong NeurIPS.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Đào tạo trư ớc theo ngữ nghĩa đối tượng cho các nhiệm vụ ngôn ngữ thị giác. Trong ECCV.

Li, XL; và Liang, P. 2021. Điều chỉnh tiền tố: Tối ưu hóa lời nhắc liên tục để tạo. Bản in trư ớc arXiv arXiv:2101.00190.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; và Zitnick, CL 2014. Microsoft coco: Các đối tượng chung trong ngữ cảnh. Trong ECCV.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; và Neubig, G. 2021. Đào tạo trư ớc, nhắc nhở và dự đoán: Khảo sát có hệ thống về các phư ơng pháp nhắc nhở trong xử lý ngôn ngữ tự nhiên. Bản in trư ớc của arXiv arXiv:2107.13586.

Loshchilov, I.; và Hutter, F. 2017. Chuẩn hóa suy giảm trọng lưu ợng tách biệt. Bản in trư ớc arXiv arXiv:1711.05101.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; và Ji, R. 2021. Bộ chuyển đổi cộng tác hai cấp để chú thích hình ảnh. Trong AAAI.

Mathews, A.; Xie, L.; và He, X. 2016. Senticap: Tạo mô tả hình ảnh với cảm xúc. Trong AAAI.

Ordonez, V.; Kulkarni, G.; và Berg, T. 2011. Im2text: Mô tả hình ảnh bằng 1 triệu bức ảnh có chú thích. Trong NeurIPS.

Pan, Y.; Yao, T.; Li, Y.; và Mei, T. 2020. Mạng chú ý tuyến tính X cho chú thích hình ảnh. Trong CVPR.

Papineni, K.; Roukos, S.; Ward, T.; và Zhu, W.-J. 2002. Bleu: một phư ơng pháp đánh giá tự động bản dịch máy. Trong ACL.

Radford, A.; Kim, JW; Hallacy, C.; Ramesh, A.; Goh, G.; Agar-wal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; và cộng sự 2021. Học các mô hình trực quan có thể chuyển giao từ tầm nhìn siêu ngôn ngữ tự nhiên. Bản in trư ớc arXiv arXiv:2103.00020.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6): 1137–1149.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; and Weston, J. 2019. Engaging image captioning via personality. In *CVPR*.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*.

Song, Z.; Zhou, X.; Mao, Z.; and Tan, J. 2021. Image captioning with context-aware auxiliary guidance. In *AAAI*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, J.; Tang, J.; and Luo, J. 2020. Multimodal attention with image text spatial relationship for ocr-based image captioning. In *ACM MM*.

Wang, Y.; Xu, J.; and Sun, Y. 2022. End-to-End Transformer Based Model for Image Captioning. In *AAAI*.

Wang, Z.; Bao, R.; Wu, Q.; and Liu, S. 2021a. Confidence-aware non-repetitive multimodal transformers for textcaps. In *AAAI*.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; and Huang, F. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804*.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florencio, D.; Wang, L.; Zhang, C.; Zhang, L.; and Luo, J. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *CVPR*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Zhang, W.; Shi, H.; Guo, J.; Zhang, S.; Cai, Q.; Li, J.; Luo, S.; and Zhuang, Y. 2022. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *AAAI*.

Zhong, Z.; Friedman, D.; and Chen, D. 2021. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*.

Ren, S.; He, K.; Girshick, R.; và Sun, J. 2016. Faster R-CNN: hướng tới phát hiện đối tượng theo thời gian thực với mạng đề xuất vùng. Tạp chí IEEE TPAMI, 39(6): 1137-1149.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; và Komatsuzaki, A. 2021. Laion-400m: Bộ dữ liệu mở gồm 400 triệu cặp hình ảnh-văn bản được lọc theo clip. Bản in trước arXiv arXiv:2111.02114.

Sharma, P.; Ding, N.; Goodman, S.; và Soricut, R. 2018. Chủ thích khái niệm: Một tập dữ liệu văn bản thay thế hình ảnh đã được làm sạch, có siêu ẩn danh để tạo chủ thích hình ảnh tự động. Trong ACL.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; và Singh, S. 2020. Tự động nhắc: Thu thập kiến thức từ các mô hình ngôn ngữ bằng lời nhắc được tạo tự động. Bản in trước arXiv arXiv:2010.15980.

Shuster, K.; Humeau, S.; Hu, H.; Biên giới, A.; và Weston, J. 2019.

Chủ thích hình ảnh hấp dẫn thông qua tính cách. Trong CVPR.

Sidorov, O.; Hu, R.; Rohrbach, M.; và Singh, A. 2020. Textcaps: một tập dữ liệu để chủ thích hình ảnh với khả năng đọc hiểu. Trong ECCV.

Song, Z.; Zhou, X.; Mao, Z.; và Tan, J. 2021. Chủ thích hình ảnh với hướng dẫn phụ trợ nhận biết ngữ cảnh. Trong AAAI.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, Ł.; và Polosukhin, I. 2017. Tất cả những gì bạn cần là chú ý. Trong NeurIPS.

Vedantam, R.; Lawrence Zitnick, C.; và Parikh, D. 2015. Cider: Đánh giá mô tả hình ảnh dựa trên sự đồng thuận. Trong CVPR.

Vinyals, O.; Toshev, A.; Bengio, S.; và Erhan, D. 2015. Hiện thị và kể lại: Một trình tạo chủ thích hình ảnh thần kinh. Trong CVPR.

Wang, J.; Tang, J.; và Luo, J. 2020. Sự chú ý đa phương thức với mối quan hệ không gian văn bản hình ảnh để chủ thích hình ảnh dựa trên ocr. Trong ACM MM.

Wang, Y.; Xu, J.; và Sun, Y. 2022. Mô hình dựa trên biến áp đầu cuối để chủ thích hình ảnh. Trong AAAI.

Wang, Z.; Bao, R.; Wu, Q.; và Liu, S. 2021a. Bộ chuyển đổi đa phương thức không lặp lại có nhận thức về độ tin cậy cho textcaps. Trong AAAI.

Vương, Z.; Yu, J.; Yu, A. W.; Đại, Z.; Tsvetkov, Y.; và Cao, Y. 2021b. Simvlm: Huấn luyện trước mô hình ngôn ngữ thị giác hình ảnh đơn giản với khả năng giám sát yếu. bản in trước arXiv arXiv:2108.10904.

Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; và Huang, F. 2021. E2E-VLP: Đào tạo trước ngôn ngữ thị giác đầu cuối được tăng cường bằng học trực quan. Bản in trước arXiv arXiv:2106.01804.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florencio, D.; Wang, L.; Zhang, C.; Zhang, L.; và Luo, J. 2021. Tap: Đào tạo trước nhận biết văn bản cho text-vqa và text-caption. Trong CVPR.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; và Gao, J. 2021. Vinvl: Xem xét lại các biểu diễn trực quan trong các mô hình ngôn ngữ thị giác. Trong CVPR.

Zhang, W.; Shi, H.; Guo, J.; Zhang, S.; Cai, Q.; Li, J.; Luo, S.; và Zhuang, Y. 2022. Magic: Suy luận đối nghịch đồ thị quan hệ đa phương thức cho chủ thích hình ảnh dựa trên văn bản đa dạng và không ghép nối. Trong AAAI.

Zhong, Z.; Friedman, D.; và Chen, D. 2021. Khảo sát thực tế là [mặt nạ]: Học so với học cách nhớ lại. Bản in trước arXiv arXiv:2104.05240.

Zhou, K.; Yang, J.; Loy, C. C.; và Liu, Z. 2021. Học cách nhắc nhở cho các mô hình ngôn ngữ thị giác. Bản in trước arXiv arXiv:2109.01134.

Chu, L.; Palangi, H.; Truong, L.; Hu, H.; Corso, J.; và Gao, J. 2020. Đào tạo trước ngôn ngữ thị giác thống nhất cho chủ thích hình ảnh và vqa. Trong AAAI.

A Social Impact and Ethics Statement

The proposed framework has the following potential positive impacts: (1) this work focuses on a meaningful direction of image captioning, i.e., maintaining both style controllability and state-of-the-art performance on prevalent benchmarks; (2) without bells and whistles, this simple pipeline is general and can be easily combined with existing captioning methods; (3) this is the first attempt to absorb the continuous prompt learning idea in the image captioning community, which may inspire future works to explore better formulations in this direction for superior results.

Nevertheless, since our method aims to describe the image from different views, it also potentially raises the risk of privacy invasion (e.g., describing a person in a negative view), which is a common concern in the captioning area.

B Method Limitation

The proposed framework optimizes *learnable* prompts for image captioners. The merits are three-fold: (1) it avoids laborious manual design; (2) it enhances the model controllable capability and allows the captioner to generate stylized captions using one model; (3) it achieves better results compared with the manual prompt.

Nevertheless, compared with the manual prompt such as “a picture of”, the main limitation is that the learned prompt representations are difficult to visualize. Since learned prompts are optimized in the *continuous* word embedding space, they fail to map back to the words in the dictionary. To this end, we search within the vocabulary for words that are closest to the learned prompt embeddings based on the dot-product similarity. Table 7 shows some examples of the approximate learned prompts (token IDs are mapped to tokens via BERT tokenizer). Although these approximate auto-prompts are difficult to be understood by humans, they are suitable for the image captioner.

Besides, our generated captions also contain some unsatisfactory descriptions. The failure cases include (1) some high-length captions include redundant expressions; (2) some images fail to generate positive or negative captions. However, we think some failures are also reasonable. For example, forcing a positive image (e.g., a smiling person) to generate negative captions is somewhat difficult.

C More Implementation Details

Pre-training Framework. Following previous works (Zhang et al. 2021; Hu et al. 2021; Li et al. 2022; Wang et al. 2021b), we also adopt the large-scale pre-training on the noisy image-text corpus to improve the downstream captioning task. We combine three widely used pre-training losses including the language modeling (LM) loss, image-text contrastive loss (Radford et al. 2021; Jia et al. 2021), and image-text matching loss (Chen et al. 2020b; Li et al. 2021, 2022) to jointly optimize the visual encoder and cross-modal fusion model, as follows:

$$\mathcal{L}_{\text{Pre-train}} = \mathcal{L}_{\text{Contrast}} + \mathcal{L}_{\text{Match}} + \mathcal{L}_{\text{LM}}. \quad (6)$$

Similar to the CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), the contrastive loss $\mathcal{L}_{\text{Contrast}}$ measures the similarity of the image-text pairs via a late fusion manner such

as dot-product. This loss aims to align the feature representations of the visual input and text input by encouraging positive image-text pairs to have similar representations in contrast to the negative pairs. The matching loss $\mathcal{L}_{\text{Match}}$ measures the image-text similarity via a deep cross-modal fusion manner such as cross-attention. Then, the output embedding of the special token [CLS] is used to conduct the binary classification to judge whether an image-text pair is positive (matched) or negative (unmatched). The language modeling loss \mathcal{L}_{LM} optimizes a cross-entropy loss which trains the model to maximize the likelihood of the text in an autoregressive manner. It has been widely recognized that the aforementioned losses can facilitate cross-modal alignment (Li et al. 2022). Therefore, although we focus on the image captioning, we additionally include the $\mathcal{L}_{\text{Contrast}}$ and $\mathcal{L}_{\text{Match}}$ in the pre-training stage. For more details, please refer to BLIP (Li et al. 2022).

Setting	Approximate Prompt Tokens
COCO Prompt	toned extend blank turnissbyzone shells lady elderea tagwing painted phone electronics
TextCap Prompt	edit extend pierce turnissbyzone shellsali elderea tagzo painted phone fuel

Table 7: Examples of the approximate auto-prompts via searching for the words that are closest to the learned prompt embeddings according to the dot-product similarity.

	Emotional Words
Pre-defined Positive Words	happy, nice, awesome, tasty, great pretty, beautiful, cute, good, delilicious
Pre-defined Negative Words	stupid, bad, lonely, disgusting, silly dead, ugly, crazy, terrible, dirty

Table 8: Details of the pre-defined emotional words.

	COCO	VG	SBU	CC3M	CC12M	LAION
Image	113K	100K	860K	3M	10M	115M
Text	567K	769K	860K	3M	10M	115M

Table 9: Details of the pre-training datasets.

Inference Stage. In the inference stage, we use beam search with a beam size of 3 to generate captions. For COCO-style, TextCap-style, Positive, Negative, and Short-length captions, the maximum generation length is set to 40 since the learned prompts have already occupied 16 tokens. For Medium-length and High-length captions, the maximum generation length is set to 60.

Training Data. To divide the training data by different emotions, we select the following 10 positive words and 10 negative words, as shown in Table 8. Note that the total positive and negative captions in the COCO dataset (Lin et al. 2014) are rare (only 1.5% of the total COCO dataset). Prompt learning facilitates the *few-shot* domain transfer using limited training samples.

Table 9 shows the details of the pre-training datasets.

Tuyên bố về tác động xã hội và đạo đức

Khung đề xuất có những tác động tích cực tiềm tàng sau: (1) công trình này tập trung vào một hướng có ý nghĩa của chủ thích hình ảnh, tức là, duy trì cả khả năng kiểm soát phong cách và hiệu suất tiên tiến trên các chuẩn mực phổ biến; (2) không có chuông và còi, dự định ống đơ n giản này là chung và có thể dễ dàng kết hợp với phụ đề hiện có phương pháp; (3) đây là nỗ lực đầu tiên để hấp thụ liên tục ý tưởng học tập nhanh chóng trong cộng đồng chủ thích hình ảnh, điều này có thể truyền cảm hứng cho các công trình trong tương lai nhằm khám phá những công thức tốt hơn theo hướng này để đạt được kết quả vượt trội. Tuy nhiên, vì phương pháp của chúng tôi nhằm mục đích mô tả hình ảnh từ các góc nhìn khác nhau nên nó cũng có khả năng làm tăng rủi ro xâm phạm quyền riêng tư (ví dụ, mô tả một người theo cách tiêu cực xem), đây là mối quan tâm phổ biến trong lĩnh vực chủ thích.

B Phương pháp hạn chế

Khung đề xuất tối ưu hóa các lời nhắc có thể học được cho người chủ thích hình ảnh. Ưu điểm là ba mặt: (1) nó tránh được thiết kế thủ công tốn công sức; (2) nó tăng cường khả năng kiểm soát nhãn mô hình và cho phép người chủ thích tạo ra các chủ thích sử dụng một mô hình; (3) nó đạt được kết quả tốt hơn so với lời nhắc thủ công.

Tuy nhiên, so với lời nhắc thủ công như vậy như “một bức tranh của”, hạn chế chính là các biểu diễn nhắc nhở đã học rất khó hình dung. Vì các lời nhắc đã học được tối ưu hóa trong từ liên tục nhưng không gian, chúng không thể ánh xạ trở lại các từ trong từ điển. Để đạt được mục đích này, chúng tôi tìm kiếm trong vốn từ vựng cho những từ gần nhất với các nhúng nhắc nhở đã học

dựa trên sự tương đồng của tích vô hướng. Bảng 7 cho thấy một số ví dụ về các lời nhắc học được gần đúng (ID mã thông báo là được ánh xạ tới các mã thông báo thông qua bộ mã thông báo BERT). Mặc dù con người khó có thể hiểu được các lời nhắc tự động gần đúng này, nhưng chúng lại phù hợp với người chủ thích hình ảnh.

Bên cạnh đó, các chủ thích do chúng tôi tạo ra cũng chứa một số mô tả không thỏa đáng. Các từ đồng hợp lỗi bao gồm (1) một số chủ thích dài bao gồm các biểu thức thừa; (2) một số hình ảnh không tạo ra được chủ thích tích cực hoặc tiêu cực. Tuy nhiên, chúng tôi nghĩ rằng một số thất bại cũng là hợp lý. Đối với ví dụ, ép buộc một hình ảnh tích cực (ví dụ, một người đang cư ời) việc tạo ra các chủ thích tiêu cực có phần khó khăn.

C Chi tiết triển khai khác

Khung đào tạo trước. Theo dõi các công trình trước đó (Zhang và cộng sự 2021; Hu và cộng sự 2021; Li và cộng sự 2022; Wang và cộng sự. 2021b), chúng tôi cũng áp dụng chương trình đào tạo trước quy mô lớn về dữ liệu hình ảnh-văn bản nhiều để cải thiện nhiệm vụ chủ thích hạ lưu. Chúng tôi kết hợp ba tổn thất trước khi đào tạo được sử dụng rộng rãi bao gồm mất mát mô hình ngôn ngữ (LM), mất mát tương phản hình ảnh-văn bản (Radford et al. 2021; Jia et al. 2021) và mất mát khớp hình ảnh-văn bản (Chen et al. 2020b; Li et al. 2021, 2022) để tối ưu hóa đồng thời bộ mã hóa hình ảnh và mô hình hợp nhất đa phương thức như sau:

$$\mathcal{L}_{\text{Pre-train}} = \mathcal{L}_{\text{Contrast}} + \mathcal{L}_{\text{Match}} + \mathcal{L}_{\text{LM}}. \quad (6)$$

Tương tự như CLIP (Radford et al. 2021) và ALIGN (Jia et al. 2021), mất độ tương phản LContrast đo lường sự giống nhau của các cặp hình ảnh-văn bản thông qua một cách hợp nhất muộn như vậy

như tích vô hướng. Sự mất mát này nhằm mục đích sắp xếp các biểu diễn đặc trưng của đầu vào trực quan và đầu vào văn bản bằng cách khuyến khích các cặp hình ảnh-văn bản tích cực có các biểu diễn tương tự trái ngược với các cặp tiêu cực. Sự mất mát phù hợp LMatch đo lường sự tương đồng giữa hình ảnh-văn bản thông qua sự kết hợp đa phương thức sâu sắc cách như sự chú ý chéo. Sau đó, nhúng đầu ra của mã thông báo đặc biệt [CLS] được sử dụng để thực hiện nhị phân phân loại để đánh giá xem cặp hình ảnh-văn bản là dự định (phù hợp) hay âm (không phù hợp). Mất mát mô hình hóa ngôn ngữ LLM tối ưu hóa mất mát entropy chéo đào tạo mô hình để tối đa hóa khả năng của văn bản theo cách tự hồi quy. Người ta đã công nhận rộng rãi rằng những tổn thất đã đề cập ở trên có thể tạo điều kiện cho sự liên kết đa phương thức (Li et al. 2022). Do đó, mặc dù chúng tôi tập trung vào hình ảnh chủ thích, chúng tôi cũng bao gồm LContrast và LMatch trong giai đoạn tiền đào tạo. Để biết thêm chi tiết, vui lòng tham khảo BLIP (Li và cộng sự, 2022).

Cài đặt	Mã thông báo nhắc nhở gần đúng
COCO Nhắc Nhở	toned mở rộng trống turnissbyzone vô quý cô elderea tagwing sơ n điện thoại điện tử
Đầu nhắc TextCap	chỉnh sửa mở rộng xuyên qua turnissbyzone shellsali elderea tagzo sơ n điện thoại nhiên liệu

Bảng 7: Ví dụ về các lời nhắc tự động gần đúng thông qua tìm kiếm những từ gần nhất với lời nhắc đã học những theo độ tương đồng của tích vô hướng.

	Từ ngữ cảm xúc
Từ ngữ tích cực dự định nghĩa trước	vui vẻ, tốt đẹp, tuyệt vời, ngon, tuyệt vời xinh xắn, đẹp, dễ thương, tốt, ngon
Từ phủ định dự định nghĩa trước	ngu ngốc, xấu, cô đơn, kinh tởm, ngột ngạt chết, xấu xí, điên rồ, khủng khiếp, bẩn thỉu

Bảng 8: Chi tiết về các từ ngữ biểu thị cảm xúc được định nghĩa trước.

	COCO	VG	SBU	CC3M	CC12M	LAION
Hình ảnh	113K	100K	860K	3M	10 triệu	115 triệu
Chữ	567K	769K	860K	3M	10 triệu	115 triệu

Bảng 9: Chi tiết về các tập dữ liệu tiền đào tạo.

Giai đoạn suy luận. Trong giai đoạn suy luận, chúng tôi sử dụng tìm kiếm chùm tia với kích thước chùm tia là 3 để tạo chủ thích. Dành cho kiểu COCO, kiểu TextCap, kiểu Positive, Negative và Short-length chủ thích, độ dài thể hệ tối đa được đặt thành 40 kể từ các lời nhắc đã học đã chiếm 16 mã thông báo. Đối với Phụ đề có độ dài trung bình và độ dài cao, tối đa Độ dài thể hệ được đặt thành 60. Dữ liệu đào tạo. Để chia dữ liệu đào tạo theo các cảm xúc khác nhau, chúng tôi chọn 10 từ tích cực và 10 từ tiêu cực sau đây, như thể hiện trong Bảng 8. Lưu ý rằng tổng số tích cực và chủ thích tiêu cực trong tập dữ liệu COCO (Lin et al. 2014) rất hiếm (chỉ chiếm 1,5% tổng số tập dữ liệu COCO). Nhắc nhở việc học tạo điều kiện thuận lợi cho việc chuyển giao miền ít lần bằng cách sử dụng các mẫu đào tạo giới hạn.

Bảng 9 hiển thị thông tin chi tiết về các tập dữ liệu tiền đào tạo.

Method	Validation Set				Test Set			
	B@4	M	C	S	B@4	M	C	S
BUTD (Anderson et al. 2018)	20.1	17.8	41.9	11.7	14.9	15.2	33.8	8.8
AoANet (Huang et al. 2019)	20.4	18.9	42.7	13.2	15.9	16.6	34.6	10.5
M4C (Hu et al. 2020)	23.3	22.0	89.6	15.6	18.9	19.8	81.0	12.8
M4C (GT OCR) (Hu et al. 2020)	26.0	23.2	104.3	16.2	21.3	21.1	97.2	13.5
MMA-SR (Wang, Tang, and Luo 2020)	24.6	23.0	98.0	16.2	19.8	20.6	88.0	13.2
CNMT (Wang et al. 2021a)	24.8	23.0	101.7	16.3	20.0	20.8	93.0	13.4
TAP (Yang et al. 2021)	25.8	23.8	109.2	17.1	21.9	21.8	103.2	14.6
ConCap (Ours)	31.3	26.0	116.7	19.6	27.4	23.9	105.6	17.3

Table 10: Comparison results on the TextCaps (Sidorov et al. 2020) validation set and test set.

Manual Prompt	B@4	M	C	S
a picture of	23.7	21.2	83.8	15.8
a photo of	25.3	22.0	88.3	16.4
a picture contains	27.3	22.8	92.5	18.0
a picture with	22.3	20.2	79.2	15.0
a picture that shows	24.5	22.3	88.2	16.4
a beautiful picture that shows	11.7	15.9	58.5	11.4
a terrible picture that shows	17.8	19.0	71.9	14.1
a normal picture that shows	19.7	19.7	76.1	14.7

Table 11: Performance of different manual prompts on the COCO Karpathy test split (Lin et al. 2014).

D Detailed Results on TextCaps

Due to the limited space, in the main paper, we only exhibit the BLEU@4 and CIDEr results on the TextCaps dataset (Sidorov et al. 2020). The compact comparison results are shown in Table 10.

E Experiments on Manual Prompt

In this section, we exhibit the captioning performance of different manual prompts. To avoid the time-consuming model training, we evaluate the pre-trained model on downstream dataset COCO (Lin et al. 2014) without fine-tuning (i.e., zero-shot setting). As shown in Table 11, we validate the zero-shot performance with different manual prompts such as “a picture of”, “a picture contains”, “a photo of”, “a picture that shows”, etc. From the results, we can observe that a slight word change will cause a clear positive or negative impact on the captioning performance. To ensure each dataset has a specific manual prompt, we cannot utilize the common prompts such as “a picture of”. To this end, we heuristically design “a normal picture that shows” and “a textual picture that shows” for COCO and TextCaps datasets, respectively.

F More Visualization Results

In Figure 5 and Figure 6, we exhibit more visualization results on the COCO (Lin et al. 2014) and TextCaps (Sidorov et al. 2020) datasets, respectively.

To split the training data by different emotions, we pre-define some positive and negative words, as illustrated in Table 8. Interestingly, our approach can generalize to more emotional words such as “funny” and “messy” that are

not contained in Table 8. The emotional captioning results are shown in Figure 5. These visualization results justify that prompt learning facilitates the few-shot learning with limited training samples.

Hướng dẫn sử dụng	B@4M		C	S
một bức ảnh của	23,7	21.2	83,8	15.8
một bức ảnh của	25.3	22.0	88,3	16.4
một bức ảnh chứa một bức ảnh	27,3	22,8	92,5	18.0
với một bức ảnh cho	22.3	20.2	79,2	15.0
thấy một bức ảnh đẹp cho thấy	24,5	22.3	88,2	16.4
một bức ảnh khiến khiếp cho thấy một bức ảnh	11.7	15,9	58,5	11.4
bình thư ờng cho thấy	17.8	19.0	71,9	14.1
	19,7	19,7	76,1	14,7

Bảng 11: Hiệu suất của các lời nhắc thủ công khác nhau trên

Phân chia thử nghiệm COCO Karpathy (Lin và cộng sự 2014).

D Kết quả chi tiết trên TextCaps

Do không gian có hạn, trong bài báo chính, chúng tôi chỉ trình bày

kết quả BLEU@4 và CIDEr trên tập dữ liệu TextCaps

(Sidorov và cộng sự 2020). Kết quả so sánh nhỏ gọn là

được thể hiện trong Bảng 10.

E Thí nghiệm trên Hướng dẫn nhắc nhở

Trong phần này, chúng tôi trình bày hiệu suất chú thích của các lời

nhắc thủ công khác nhau. Để tránh mô hình tốn thời gian

đào tạo, chúng tôi đánh giá mô hình được đào tạo trước ở hạ lưu

tập dữ liệu COCO (Lin et al. 2014) mà không cần tinh chỉnh (tức là,

thiết lập zero-shot). Như thể hiện trong Bảng 11, chúng tôi xác nhận

hiệu suất không biến với các lời nhắc thủ công khác nhau như

như “một bức tranh của”, “một bức tranh chứa đựng”, “một

ảnh của”, “một bức ảnh cho thấy”, v.v. Từ

kết quả, chúng ta có thể quan sát thấy một sự thay đổi nhỏ về từ ngữ sẽ

gây ra tác động tích cực hoặc tiêu cực rõ ràng đến hiệu suất chú

thích. Để đảm bảo mỗi tập dữ liệu có một

nhắc nhở thủ công, chúng tôi không thể sử dụng các nhắc nhở chung

chẳng hạn như “một bức tranh của”. Để đạt được mục đích này, chúng tôi

thiết kế theo phương pháp thử nghiệm “một bức tranh bình thường cho thấy” và

“một hình ảnh văn bản cho thấy” đối với COCO và

Bộ dữ liệu TextCaps tương ứng.

F Kết quả trực quan hơn

Trong Hình 5 và Hình 6, chúng tôi trình bày nhiều kết quả trực quan

hơn trên COCO (Lin et al. 2014) và TextCaps (Sidorov

et al. 2020) tập dữ liệu tương ứng.

Để phân chia dữ liệu đào tạo theo các cảm xúc khác nhau, chúng tôi

xác định trước một số từ tích cực và tiêu cực, như minh họa trong

Bảng 8. Điều thú vị là cách tiếp cận của chúng tôi có thể khái quát hóa thành nhiều

những từ ngữ mang tính cảm xúc như “buồn cười” và “lộn xộn”

không có trong Bảng 8. Kết quả chú thích cảm xúc

được hiển thị trong Hình 5. Những kết quả trực quan này chứng minh

rằng việc học nhanh chóng tạo điều kiện cho việc học ít lần với

mẫu đào tạo hạn chế.



COCO-style Caption
A dog wearing a green and red hat

Positive Caption
A very **cute** small dog wearing a **funny** hat

Negative Caption
A dog wearing a **silly** hat in the snow

Short-length Caption
A dog wearing a green and red hat

Medium-length Caption
A brown and white dog wearing a green and red hat

High-length Caption
A brown and white dog wearing a green and red hat sitting in the snow

Ground-truth Caption
A close up of a dog sitting wearing a hat



COCO-style Caption
A sandwich on a plate on a wooden table

Positive Caption
A very **tasty looking** sandwich on a plate

Negative Caption
A plate that has a very **bad looking** sandwich on it

Short-length Caption
A sandwich on a plate on a wooden table

Medium-length Caption
A plate with a sandwich on it sitting on a wooden table

High-length Caption
A white plate with a blue rim holds a sandwich that has a bite taken out of it

Ground-truth Caption
A sandwich on a white plate on a wooden table



COCO-style Caption
A black couch sitting in a living room next to a wooden wall

Positive Caption
A very **nice looking** living room with a black couch

Negative Caption
A **dirty** black couch in a **dirty** room

Short-length Caption
A living room with a black couch in it

Medium-length Caption
A black couch sitting in a living room next to a wooden wall

High-length Caption
A living room with a black couch and a wooden wall with a lamp on it

Ground-truth Caption
A low black couch with lots of pillows



Chủ thích theo phong cách COCO
Một con chó đội mũ xanh và đỏ

Chủ thích tích cực
Một chú chó nhỏ rất dễ thương đội một chiếc mũ ngộ nghĩnh

Tiêu đề tiêu cực
Một chú chó đội chiếc mũ ngớ ngẩn trên tuyết

Chủ thích ngắn
Một con chó đội mũ xanh và đỏ

Chủ thích có độ dài trung bình
Một con chó màu nâu và trắng đội mũ xanh và đỏ

Chủ thích dài
Một chú chó nâu và trắng đội mũ xanh và đỏ đang ngồi trên tuyết

Chủ thích thực tế
Cận cảnh một chú chó đang ngồi đội mũ



Chủ thích theo phong cách COCO
Một chiếc bánh sandwich trên đĩa trên bàn gỗ

Chủ thích tích cực
Một chiếc bánh sandwich trông rất ngon trên đĩa

Tiêu đề tiêu cực
Một chiếc đĩa có một chiếc bánh sandwich trông rất xấu xí

Chủ thích ngắn
Một chiếc bánh sandwich trên đĩa trên bàn gỗ

Chủ thích có độ dài trung bình
Một chiếc đĩa đựng một chiếc bánh sandwich đặt trên một chiếc bàn gỗ

Chủ thích dài
Một chiếc đĩa trắng có viền xanh đựng một chiếc bánh sandwich đã bị cắn mất một miếng

Chủ thích thực tế
Một chiếc bánh sandwich trên một chiếc đĩa trắng trên một chiếc bàn gỗ



Chủ thích theo phong cách COCO
Một chiếc ghế dài màu đen đặt trong phòng khách cạnh bức tư ông gỗ

Chủ thích tích cực
Một phòng khách **trông rất đẹp** với chiếc ghế dài màu đen

Tiêu đề tiêu cực
Một chiếc ghế dài màu đen bẩn thỉu trong một căn phòng bẩn thỉu

Chủ thích ngắn
Một phòng khách có một chiếc ghế dài màu đen

Chủ thích có độ dài trung bình
Một chiếc ghế dài màu đen đặt trong phòng khách cạnh bức tư ông gỗ

Chủ thích dài
Một phòng khách có ghế sofa màu đen và một bức tư ông gỗ có đèn trên đó

Chủ thích thực tế
Một chiếc ghế dài màu đen thấp với nhiều gối



COCO-style Caption
A herd of horses walking across a river

Positive Caption
A herd of **beautiful** horses running through a river

Negative Caption
A herd of wild horses walking across a **dirty** river

Short-length Caption
A herd of horses walking across a river

Medium-length Caption
A herd of horses walking across a river next to a forest

High-length Caption
A herd of horses running through a body of water with trees in the background

Ground-truth Caption
A bunch of horses that are standing in the water



COCO-style Caption
A man wearing a white shirt and a tie

Positive Caption
A man wearing a **very nice looking** tie

Negative Caption
A man wearing a tie with a **silly** design on it

Short-length Caption
A man wearing a white shirt and a tie

Medium-length Caption
A man wearing a white shirt and a tie with a cartoon design

High-length Caption
A man wearing a white shirt and a tie with a colorful design on it

Ground-truth Caption
A person wearing a white shirt and colorful tie



COCO-style Caption
An unmade bed with a blanket on top of it

Positive Caption
A **very nice looking** bed in a room

Negative Caption
A **messy** bed with a **messy** blanket on top of it

Short-length Caption
A bed with a blanket on top of it

Medium-length Caption
A bed with a blanket on top of it in a bedroom

High-length Caption
An unmade bed with a blanket on top of it in a room with a clock on the wall

Ground-truth Caption
A bedroom with a giant clock hanging in the wall



Chủ thích theo phong cách COCO
Một đàn ngựa đang đi qua sông

Chủ thích tích cực
Một đàn ngựa **đẹp** đang chạy qua một con sông

Tiêu đề tiêu cực
Một đàn ngựa hoang đang đi qua một con sông bẩn

Chủ thích ngắn
Một đàn ngựa đang đi qua sông

Chủ thích có độ dài trung bình
Một đàn ngựa đang đi qua một con sông bên cạnh một khu rừng

Chủ thích dài
Một đàn ngựa chạy qua một vùng nư ớc có cây cối ở phía sau

Chủ thích thực tế
Một đàn ngựa đang đứng dư ới nư ớc



Chủ thích theo phong cách COCO
Một ngư ời đàn ông mặc áo sơ mi trắng và cà vạt

Chủ thích tích cực
Một ngư ời đàn ông đeo một chiếc cà vạt **trông rất đẹp**

Tiêu đề tiêu cực
Một ngư ời đàn ông đeo cà vạt có thiết kế ngớ ngẩn

Chủ thích ngắn
Một ngư ời đàn ông mặc áo sơ mi trắng và cà vạt

Chủ thích có độ dài trung bình
Một ngư ời đàn ông mặc áo sơ mi trắng và cà vạt có thiết kế hoạt hình

Chủ thích dài
Một ngư ời đàn ông mặc áo sơ mi trắng và cà vạt có họa tiết nhiều màu sắc

Chủ thích thực tế
Một ngư ời mặc áo sơ mi trắng và cà vạt nhiều màu



Chủ thích theo phong cách COCO
Một chiếc gi ường ch ứa a dọn với một chiếc ch ần ở trên

Chủ thích tích cực
Một chiếc gi ường **trông rất đẹp** trong phòng

Tiêu đề tiêu cực
Một chiếc gi ường lộn xộn với một chiếc ch ần lộn xộn ở trên

Chủ thích ngắn
Một chiếc gi ường có ch ần ở trên

Chủ thích có độ dài trung bình
Một chiếc gi ường có ch ần phủ lên trên trong phòng ngủ

Chủ thích dài
Một chiếc gi ường ch ứa a dọn với một chiếc ch ần phủ lên trên trong một căn phòng có đồng hồ trên tư ờng

Chủ thích thực tế
Một phòng ngủ có chiếc đồng hồ khổng lồ treo trên tư ờng

Figure 5: Image captioning examples from COCO (Karpathy and Fei-Fei 2015) with different styles including COCO-style [COCO], Positive [Positive], Negative [Negative], Short-length [Short-length], Medium-length [Medium-length], and High-length [High-length].

Hình 5: Các ví dụ chú thích hình ảnh từ COCO (Karpathy và Fei-Fei 2015) với các phong cách khác nhau bao gồm phong cách COCO [COCO], Tích cực [Positive], Tiêu cực [Negative], Độ dài ngắn [Short-length], Độ dài trung bình [Medium-length] và Độ dài cao [High-length].



COCO-style Caption
A baseball player standing on a baseball field

TextCap-style Caption
A baseball player with the number **12** on his jersey

Ground-truth Caption
A baseball player for the Whitecaps team wearing number 12 goes to throw a baseball.



COCO-style Caption
A white ipod sitting on top of a table

TextCap-style Caption
An ipod with a screen that says **long kong phoeey sublime** on it

Ground-truth Caption
An iPod with a Sublime song currently playing



COCO-style Caption
A large jetliner taking off from an airport runway

TextCap-style Caption
A group of people holding a sign that says **tax dodgers**

Ground-truth Caption
The people protesting are holding a sign saying Tax Dodgers



Chú thích theo phong cách COCO
Một cầu thủ bóng chày đang đứng trên sân bóng chày

Chú thích theo kiểu TextCap
Một cầu thủ bóng chày có số **12** trên áo đấu

Chú thích thực tế
Một cầu thủ bóng chày của đội Whitecaps mặc áo số 12 đang ném bóng chày.



Chú thích theo phong cách COCO
Một chiếc ipod màu trắng đặt trên bàn

Chú thích theo kiểu TextCap
Một chiếc ipod có màn hình ghi chữ **long kong phoeey tuyệt vời** trên đó

Chú thích thực tế
Một chiếc iPod đang phát bài hát Sublime



Chú thích theo phong cách COCO
Một máy bay phản lực lớn cất cánh từ đư ờng băng sân bay

Chú thích theo kiểu TextCap
Một nhóm người cầm tấm biển ghi dòng chữ **trốn thuế**

Chú thích thực tế
Những ngư ời biểu tình đang cầm một tấm biển ghi rằng Những kẻ trốn thuế



COCO-style Caption
A clock on a wall with a picture of a child on it

TextCap-style Caption
A clock on a wall with a sign that says **unattended children will be given espresso and free kitten**

Ground-truth Caption
A wall with a clock and a sign that says unattended children will be given espresso and a free kitten.



COCO-style Caption
A bottle of liquid sitting on top of a tree stump

TextCap-style Caption
A bottle of liquid with a yellow label that says **syrup of the night** on it

Ground-truth Caption
A laxative with pleasant flavor packed in a bottle with a yellow flavor



COCO-style Caption
A bunch of coins are on a table

TextCap-style Caption
A pile of coins with one of them saying **united states of america** on it

Ground-truth Caption
A pile of shiny American coins with the words United States of America.



Chú thích theo phong cách COCO
Một chiếc đồng hồ trên tư ờng có hình ảnh một đứa trẻ

Chú thích theo kiểu TextCap
Một chiếc đồng hồ trên tư ờng có biển báo ghi rằng **trẻ em không đư ợc trông nom sẽ đư ợc tặng cà phê espresso và mèo con miễn phí**

Chú thích thực tế
Một bức tư ờng có đồng hồ và biển báo ghi rằng trẻ em không có ngư ời trông coi sẽ đư ợc tặng cà phê espresso và một chú mèo con miễn phí.



Chú thích theo phong cách COCO
Một chai chất lỏng đặt trên gốc cây

Chú thích theo kiểu TextCap
Một chai chất lỏng có nhãn màu vàng ghi là **xi- rô của đêm**

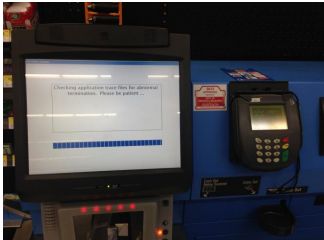
Chú thích thực tế
Thuốc nhuận tràng có hư ơng vị dễ chịu đư ợc đóng gói trong chai có hư ơng vị màu vàng



Chú thích theo phong cách COCO
Một đồng tiền xu ở trên bàn

Chú thích theo kiểu TextCap
Một đồng tiền xu có một trong số chúng ghi dòng chữ **Hoa Kỳ** trên đó

Chú thích thực tế
Một đồng tiền xu Mỹ sáng bóng có dòng chữ Hoa Kỳ.



COCO-style Caption
A machine that has a screen on it

TextCap-style Caption
An atm machine with a screen that says **checking application for additional permission** on it

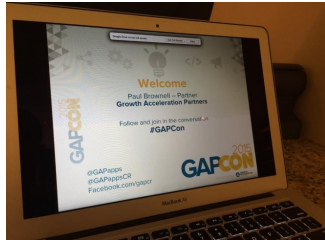
Ground-truth Caption
A screen next to a phone is displaying a message about abnormal termination



COCO-style Caption
A television screen showing a large crowd at a concert

TextCap-style Caption
A tv screen shows a large crowd at a concert with the words the **senate** at the bottom

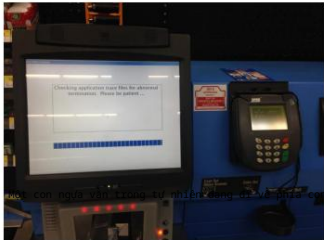
Ground-truth Caption
A Samsung tv display shows a live recording of The Senate against Obama and McCain



COCO-style Caption
A laptop computer sitting on top of a table

TextCap-style Caption
A **macbook air** is open to a **welcome** screen

Ground-truth Caption
A MacBook Air screen showing website Gap Con



Chú thích theo phong cách COCO
Một máy có màn hình trên đó

Chú thích theo kiểu TextCap
Một máy ATM có màn hình ghi là **đang kiểm tra đơ n xin cấp phép bỏ sung**

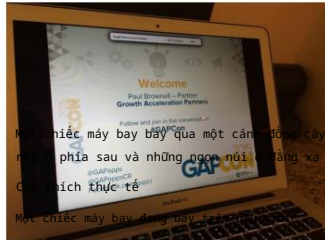
Chú thích thực tế
Màn hình bên cạnh điện thoại đang hiển thị thông báo về việc chấm dứt bất thư ờng



Chú thích theo phong cách COCO
Một màn hình tivi đang chiếu cảnh đám đông lớn tại một buổi hòa nhạc

Chú thích theo kiểu TextCap
Một màn hình tivi cho thấy một đám đông lớn tại một buổi hòa nhạc với dòng chữ **thư ờng viện** ở phía dư ới

Chú thích thực tế
Màn hình tivi Samsung chiếu trực tiếp cảnh Thư ờng viện đấu với Obama và McCain



Chú thích theo phong cách COCO
Một máy tính xách tay đặt trên bàn

Chú thích theo kiểu TextCap
Macbook **air** mở màn hình **chào mừng**

Chú thích thực tế
Màn hình MacBook Air hiển thị trang web Gap Con

Figure 6: Image captioning examples from TextCaps validation split (Sidorov et al. 2020) with different styles including COCO-style [] and TextCap-style []. Best view in color and zoom in.

Hình 6: Các ví dụ chú thích hình ảnh từ phân tách xác thực TextCaps (Sidorov và cộng sự 2020) với các kiểu khác nhau bao gồm kiểu COCO [] và kiểu TextCap []. Xem tốt nhất bằng màu sắc và phóng to.

Một con ngựa vẫn trong tự nhiên đang đi về phía con đư ờng đất

Một vôi cứu hỏa màu vàng với một vài con mắt vẽ trên đó

Một chiếc máy bay bay qua một cánh đồng cây với một tòa nhà ở phía sau và những ngọn núi ở đằng xa

Chú thích thực tế

Một chiếc máy bay đang bay trên bầu trời