

Caption Anything: Interactive Image Description with Diverse Multimodal Controls

Teng Wang^{*†}, Jinrui Zhang*, Junjie Fei*, Hao Zheng, Yunlong Tang, Zhe Li,
Mingqi Gao, Shanshan Zhao
SUSTech VIP Lab

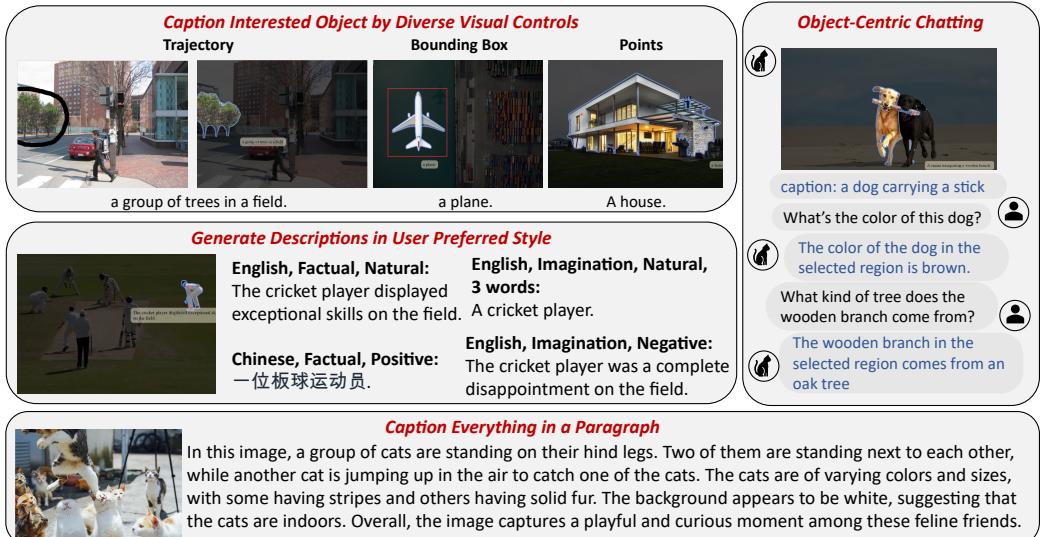


Figure 1: Caption Anything supports a diverse range of visual and language controls, making it effortlessly adaptable for object-centric chatting and image paragraph captioning.

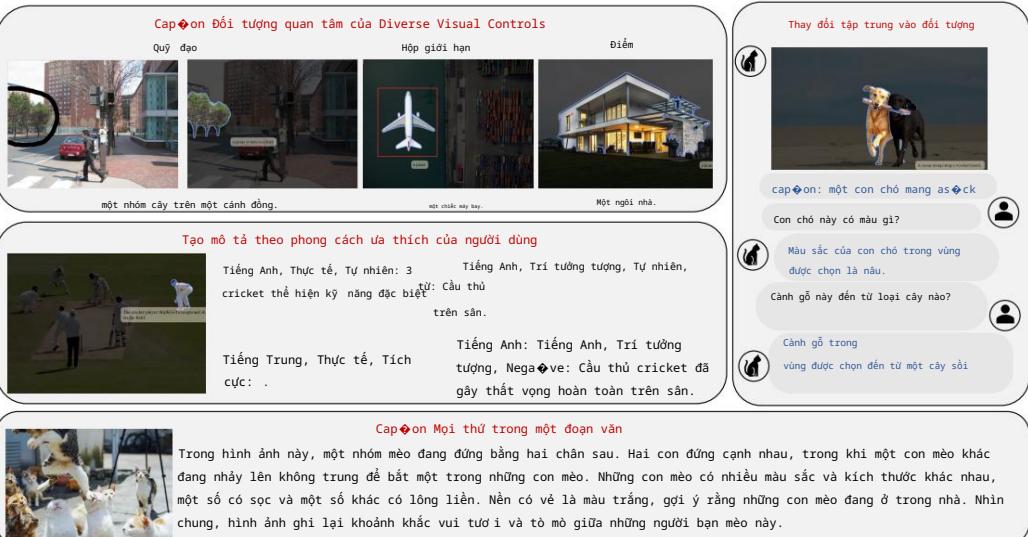
Abstract

Controllable image captioning is an emerging multimodal topic that aims to describe the image with natural language following human purpose, *e.g.*, looking at the specified regions or telling in a particular text style. State-of-the-art methods are trained on annotated pairs of input controls and output captions. However, the scarcity of such well-annotated multimodal data largely limits their usability and scalability for interactive AI systems. Leveraging unimodal instruction-following foundation models is a promising alternative that benefits from broader sources of data. In this paper, we present Caption AnyThing (CAT), a foundation model augmented image captioning framework supporting a wide range of multimodal controls: 1) visual controls, including points, boxes, and trajectories; 2) language controls, such as sentiment, length, language, and factuality. Powered by Segment Anything Model (SAM) and ChatGPT, we unify the visual and language prompts into a modularized framework, enabling the flexible combination between different controls. Extensive case studies demonstrate the user intention alignment capabilities of our framework, shedding light on effective user interaction modeling in vision-language applications. Our code is publicly available at <https://github.com/ttengwang/Caption-Anything>.

* Equal contribution. † Work done during internship in ARC Lab, Tencent PCG. We thank Yixiao Ge and Ying Shan for their support and constructive discussions.

Chú thích bất kỳ: Mô tả hình ảnh tương tác với nhiều điều khiển đa phương thức

Đặng Võng †, Jinrui Zhang , Junjie Fei , Hách Chính, Đường Văn Long, Triết Lý,
Cao Minh Kỳ, Triệu San San
Phòng thí nghiệm VIP của SUSTech



Hình 1: Caption Anything hỗ trợ nhiều loại điều khiển ngôn ngữ và hình ảnh, giúp dễ dàng thích ứng với việc trò chuyện theo đối tượng và chú thích đoạn văn hình ảnh.

Tóm tắt

Chú thích hình ảnh có thể kiểm soát là một chủ đề đa phương thức mới nhằm mục đích mô tả hình ảnh bằng ngôn ngữ tự nhiên theo mục đích của con người, ví dụ, xem các vùng đã chỉ định hoặc kể theo một kiểu văn bản cụ thể. Các phương pháp tiên tiến được đào tạo trên các cặp điều khiển đầu vào và chú thích đầu ra được chú thích. Tuy nhiên, sự khan hiếm của dữ liệu đa phương thức được chú thích tốt như vậy phần lớn hạn chế khả năng sử dụng và khả năng mở rộng của chúng đối với các hệ thống AI tương tác. Tận dụng các mô hình nền tảng theo hướng dẫn nêu phong cách là một giải pháp thay thế đầy hứa hẹn, được hưởng lợi từ các nguồn dữ liệu rộng hơn. Trong bài báo này, chúng tôi trình bày Caption AnyThing (CAT), một khuôn khổ chú thích hình ảnh tăng cường mô hình nền tảng hỗ trợ nhiều loại điều khiển đa mô hình: 1) điều khiển trực quan, bao gồm các điểm, hộp và quỹ đạo; 2) điều khiển ngôn ngữ, chẳng hạn như tình cảm, độ dài, ngôn ngữ và tính thực tế. Được hỗ trợ bởi Segment Anything Model (SAM) và ChatGPT, chúng tôi hợp nhất các lời nhắc trực quan và ngôn ngữ thành một khuôn khổ mô-đun hóa, cho phép kết hợp linh hoạt giữa các điều khiển khác nhau. Các nghiên cứu trường hợp mở rộng chứng minh khả năng cân chỉnh ý định của người dùng trong khuôn khổ của chúng tôi, làm sáng tỏ mô hình tương tác người dùng hiệu quả trong các ứng dụng ngôn ngữ thị giác. Mã của chúng tôi có sẵn công khai tại <https://github.com/ttengwang/Caption-Anything>.

Đóng góp ngang nhau. † Công việc thực hiện trong thời gian thực tập tại ARC Lab, Tencent PCG. Chúng tôi cảm ơn Yixiao Ge và Ying Shan vì sự hỗ trợ và thảo luận mang tính xây dựng của họ.

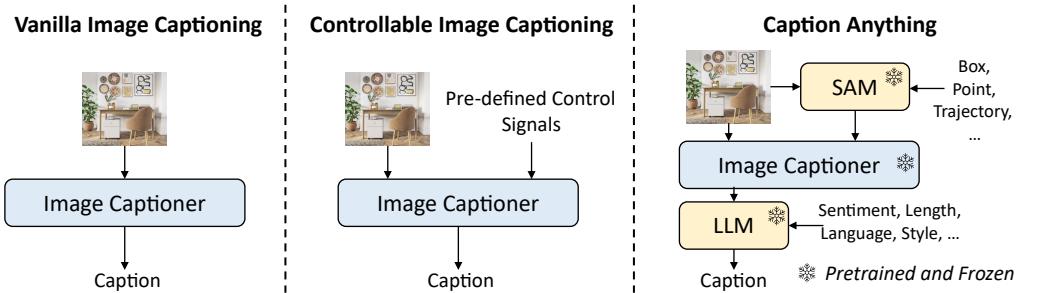


Figure 2: Comparison between image captioning pipelines. Vanilla image captioning methods lack explicit control, making them unsuitable to interact with users. Previous controllable image captioning methods mainly rely on limited-scale human-annotated data with specific control signals, and they only support pre-defined control signals. The proposed CAT is training-free and supports diverse visual controls and language controls.

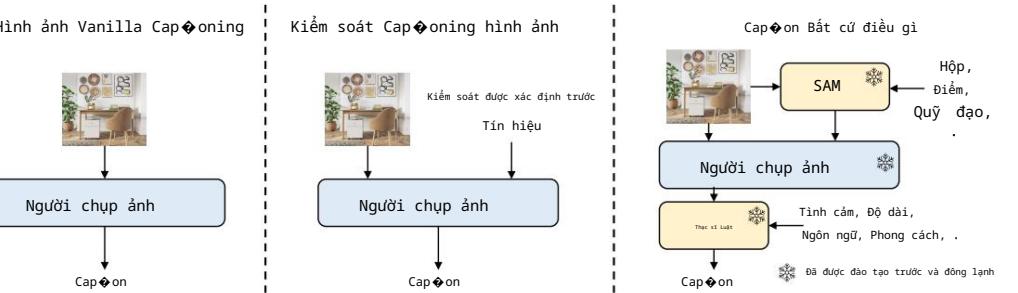
1 Introduction

Describing images in natural language is a critical problem of vision-language learning. Fig. 2 demonstrates the difference between image captioning systems. Vanilla image captioning [11, 8, 20, 9, 40] and dense captioning [13, 47, 43] typically present an image’s salient features using objective and pragmatic language, either in a single sentence or a set of sentences. However, such approaches may produce output that is excessively concise or overly complex, rendering them unsuitable for user interaction due to the lack of explicit control that aligns with user intention. Controllable Image Captioning (CIC) is a promising research direction that aligns language output with user intent. Subsequent methods [6, 3, 5, 29] have been proposed to incorporate diverse control signals into image captioning models. However, their applicability is limited by two primary factors: (1) Existing CIC models typically rely on training with human-annotated (image, text, control signal) tuples [10, 25]. The limited scale of the datasets constrains these models’ capacity to comprehend control signals; (2) These models only support pre-defined single or several control signals, which limits their flexibility of combining different controls and introducing new dimensions of controllability.

To tackle these issues, we propose Caption AnyThing (CAT), a zero-shot controllable image captioning framework augmented by pre-trained foundation models. Specifically, CAT integrates pre-trained image captioners [18, 17, 38] with SAM [14] and an instruction-tuned LLM. The image and the visual controls are first processed by SAM, which generates a pixel-level mask that corresponds to the selected region, thereby facilitating the perception centered on user-interested objects. Benefiting from the various visual prompts (*e.g.*, points, bounding boxes) used during the training of SAM, CAT supports flexible visual controls when interacting with users. The output sentences are further refined by an instruction-tuned LLM. Prominent LLMs, such as GPT-4 [27] and FLAN-T5 [4], are tuned with human feedback, thereby enabling CAT to accommodate a variety of language controls and align more effectively with user intent.

In contrast to existing controllable captioning approaches, CAT leverages foundation models to establish its controllability rather than solely relying on learning pre-defined control signals from training data. This approach not only reduces the reliance on human-annotated data, leading to a training-free model but also enhances the model’s transferability by harnessing the knowledge embedded in vast pre-training data. Furthermore, CAT supports a diverse range of control signals, making it highly adaptable and extendable for interactive use. Currently, it supports 3 vision (click, boxes, trajectory) and 4 language (sentiment, length, language, factuality) control signals, which can be flexibly combined to generate diverse and personalized captions. In addition, our model provides unified representations for both two types of controls. To be specific, the visual controls and language controls are unified to pixel-level masks and textual prompts, respectively. With this design, CAT could be readily expanded with any control signals that can be transposed into these unified representations, thereby augmenting its flexibility and scalability.

We present the strong user-interactive capabilities of CAT through a comprehensive array of qualitative examples. As shown in Fig. 1, users could select the interested objects via various visual controls to generate captions in their preferred styles. Moreover, by incorporating additional OCR and VQA



Hình 2: So sánh giữa các đường ống chú thích hình ảnh. Các phuong pháp chú thích hình ảnh vani thiieu khac nang kiem soat rõ ràng, khien chung khong phu hop de tuong tac voi nguoi dung. Các phuong pháp chut thich hinh anh co the kiem soat truoc day chut yeu duoc vao du lieu co chut thich cua con nguoi o quy mo hanh che voi cac tin hiệu dien khiien cu the va chung chi hoi truc cac tin hiệu dien khiien duoc xác định trước. CAT duoc de xuất khong can đào tạo và hoi truc nheu dien khiien truc quan va dien khiien ngon ngug.

1 Giới thiệu

Mô tả hình ảnh bằng ngôn ngữ tự nhiên là một vấn đề quan trọng của việc học ngôn ngữ thị giác. Hình 2 minh họa sự khác biệt giữa các hệ thống chú thích hình ảnh. Chú thích hình ảnh vani [11, 8, 20, 9, 40] và chú thích dày đặc [13, 47, 43] thường trình bày các đặc điểm nổi bật của hình ảnh bằng ngôn ngữ khách quan và thực dụng, trong một câu duy nhất hoặc một tập hợp các câu. Tuy nhiên, các cách tiếp cận như vậy có thể tạo ra đầu ra quá súc tích hoặc quá phức tạp, khiến chúng không phù hợp với tương tác của người dùng do thiếu kiểm soát rõ ràng phù hợp với ý định của người dùng. Chú thích hình ảnh có thể kiểm soát (CIC) là một hướng nghiên cứu đầy hứa hẹn, phù hợp với đầu ra ngôn ngữ với ý định của người dùng. Các phuong pháp tiếp theo [6, 3, 5, 29] đã được đề xuất để kết hợp các tín hiệu điều khiển đa dạng vào các mô hình chú thích hình ảnh. Tuy nhiên, khả năng áp dụng của chúng bị hạn chế bởi hai yếu tố chính: (1) Các mô hình CIC hiện có thường dựa vào việc đào tạo với các bộ chú thích của con người (hình ảnh, văn bản, tín hiệu điều khiển) [10, 25]. Quy mô hạn chế của các tập dữ liệu hạn chế khả năng hiểu các tín hiệu điều khiển của các mô hình này; (2) Các mô hình này chỉ hỗ trợ một hoặc nhiều tín hiệu điều khiển được xác định trước, điều này hạn chế tính linh hoạt trong việc kết hợp các điều khiển khác nhau và đưa ra các chiều hướng điều khiển mới.

Để giải quyết những vấn đề này, chúng tôi đề xuất Caption AnyThing (CAT), một khuôn khổ chú thích hình ảnh có thể điều khiển bằng zero-shot được tăng cường bởi các mô hình nền tảng được đào tạo trước. Cụ thể, CAT tích hợp các trình chú thích hình ảnh được đào tạo trước [18, 17, 38] với SAM [14] và LLM được điều chỉnh theo hướng dẫn. Hình ảnh và các điều khiển trực quan trước tiên được xử lý bởi SAM, tạo ra mặt nạ pixel tương ứng với vùng đã chọn, do đó tạo điều kiện cho nhận thức tập trung vào các đối tượng mà người dùng quan tâm. Tận dụng lợi thế từ nhiều lời nhắc trực quan khác nhau (ví dụ: điểm, hộp giới hạn) được sử dụng trong quá trình đào tạo SAM, CAT hỗ trợ các điều khiển trực quan linh hoạt khi tương tác với người dùng. Các câu đầu ra được tinh chỉnh thêm bằng LLM được điều chỉnh theo hướng dẫn. Các LLM nổi bật, chẳng hạn như GPT-4 [27] và FLAN-T5 [4], được điều chỉnh theo phản hồi của con người, do đó cho phép CAT chứa nhiều điều khiển ngôn ngữ khác nhau và phù hợp hiệu quả hơn với ý định của người dùng.

Ngược lại với các phuong pháp chú thích có thể kiểm soát hiện có, CAT tận dụng các mô hình nền tảng để thiết lập khả năng kiểm soát của nó thay vì chỉ dựa vào việc học các tín hiệu điều khiển được xác định trước từ dữ liệu đào tạo. Phương pháp này không chỉ làm giảm sự phụ thuộc vào dữ liệu do con người chú thích, dẫn đến một mô hình không cần đào tạo mà còn tăng cường khả năng chuyển giao của mô hình bằng cách khai thác kiến thức được nhúng trong dữ liệu đào tạo trước không lõi. Hơn nữa, CAT hỗ trợ nhiều loại tín hiệu điều khiển khác nhau, giúp nó có khả năng thích ứng và mở rộng cao để sử dụng tương tác. Hiện tại, nó hỗ trợ 3 tín hiệu điều khiển thị giác (nhấp, hộp, quỹ đạo) và 4 ngôn ngữ (tính cảm, độ dài, ngôn ngữ, tính thực tế), có thể được kết hợp linh hoạt để tạo ra các chủ đề đa dạng và được cá nhân hóa. Ngoài ra, mô hình của chúng tôi cung cấp các biểu diễn thống nhất cho cả hai loại điều khiển. Cụ thể hơn, các điều khiển trực quan và điều khiển ngôn ngữ được thông nhất thành mặt nạ pixel và lời nhắc văn bản. Với thiết kế này, CAT có thể dễ dàng mở rộng với bất kỳ tín hiệu điều khiển nào có thể được chuyển thành các biểu diễn thống nhất này, do đó tăng cường tính linh hoạt và khả năng mở rộng của nó.

Chúng tôi trình bày khả năng tương tác mạnh mẽ của CAT thông qua một loạt các ví dụ định tính toàn diện. Như thể hiện trong Hình 1, người dùng có thể chọn các đối tượng quan tâm thông qua nhiều điều khiển trực quan khác nhau để tạo chủ thích theo phong cách ưa thích của họ. Hơn nữa, bằng cách kết hợp OCR và VQA bổ sung

tools, CAT could be easily extended to two multimodal applications, namely, object-centric chatting and image paragraph captioning. The former enables users to chat around specific objects, facilitating a more in-depth understanding of interested objects. The latter effectively integrates knowledge from various domains originating from distinct foundation models, enabling the generation of detailed and logically coherent descriptions. Overall, diverse cases show that CAT is a highly interactive multimodal system, exhibiting considerable potential for real-world applications.

In summary, the contributions of this paper are three-fold: 1) We propose a training-free CIC framework that is built upon foundation models, leading to reduced reliance on human-annotated data. 2) Our approach supports a more diverse range of controls and offers unified representations for both visual and language controls, facilitating extensibility to incorporate new aspects of controllability 3) Experiments demonstrate strong user-interactive capabilities of CAT.

2 Related Work

Image Captioning and Dense Captioning. Image captioning is a multimodal task that aims to generate descriptions for a given image. The prevailing approaches mainly employ the encoder-decoder paradigm [36, 44, 1, 12, 18, 17, 38] to solve this task. To enrich the detailed understanding for complex scenes, dense captioning [13, 47, 19, 48, 31, 43] is proposed to generate localized captions for all salient objects in an image. Yin et al. [48] propose a multi-scale contextual information-sharing technique to capture fine-grained global context information. Wu et al. [43] propose a unified framework for dense captioning and object detection, achieving better object understanding capability. Nonetheless, these methods generate descriptions based on the image’s salient features only, rendering them unsuitable for user interaction as they lack an explicit control mechanism that aligns with the user’s intention. Compare with them, our method focuses on interactive image description and supports diverse multimodal controls to provide better user alignment.

Controllable Image Captioning. The controllable capability to generate desired descriptions related to user-specified objects in the image is useful in practical applications. On the one hand, the visual controls in controllable image captioning usually involve the bounding box [5], mouse behavior [29, 46]. Cornia et al. [5] propose to generate the corresponding caption based on a sequence or a set of image regions. Pont-Tuset et al. [29] release an interactive image captioning dataset where the annotators are required to describe the image region the mouse trajectory covered. LoopCAG [46] further improves the generation quality of captions and interactive controllability. On the other hand, the text style of the generated caption can be flexibly changed according to different application scenarios. Typically, the controllable text style can be summarized as: length controllability [6], sentiment controllability [25], imaginary controllability (*e.g.*, romantic or humorous descriptions) [51, 10]. Some works [49, 39] unify these controllable styles into a single architecture. BLIP2 [17], a language-image pre-training model, performs zero-shot image-to-text generation following natural language prompts. In this paper, we further extend the flexibility of controllable image captioning, where the interaction can be points, boxes, or trajectory specified by users, and the language style of generated descriptions is able to meet users’ requirements to the greatest extent.

Interactive Image Segmentation. Interactive image segmentation [22, 37, 45, 32] is an important research problem where the model is required to segment the image according to the user’s intention with interaction (*e.g.*, point, trajectory). Xu et al. [45] propose deep-learning-based interactive image segmentation firstly. Li et al present an end-to-end interactive image segmentation [21]. The whole architecture is divided into two convolutional networks, where the first is used to synthesize various masks according to the user’s input, and the second is trained to select a single solution among these masks. FCA-Net [23] makes better use of the first click to improve the interactive segmentation result. SAM [14] builds a promptable foundation model for segmentation, where the prompt can be any information indicating what to segment in an image, *e.g.*, a set of points, a rough box or mask, or free-form text.

Large Language Models. LLMs have arisen significant interest due to their strong transfer capabilities across a diverse range of language processing tasks, as well as their emerging capabilities to interact with humans. Recently, the most important breakthrough was made by GPT-3 [2], a model with 175B parameters, which unveiled the emerging potential of the few-shot learning techniques. This remarkable performance motivated lots of subsequent LLMs [50, 4, 35, 30]. In order to enhance

công cụ, CAT có thể dễ dàng mở rộng thành hai ứng dụng đa phương thức, cụ thể là trò chuyện lấy đối tượng làm trung tâm và chú thích đoạn văn hình ảnh. Ứng dụng trước cho phép người dùng trò chuyện xung quanh các đối tượng cụ thể, tạo điều kiện hiểu sâu hơn về các đối tượng quan tâm. Ứng dụng sau tích hợp hiệu quả kiến thức từ nhiều lĩnh vực khác nhau có nguồn gốc từ các mô hình nền tảng riêng biệt, cho phép tạo ra các mô tả chi tiết và mạch lạc về mặt logic. Nhìn chung, nhiều trường hợp khác nhau cho thấy CAT là một hệ thống đa phương thức có tính tương tác cao, thể hiện tiềm năng đáng kể cho các ứng dụng trong thế giới thực.

Tóm lại, bài báo này có ba đóng góp: 1) Chúng tôi đề xuất một khuôn khổ CIC không cần đào tạo được xây dựng dựa trên các mô hình nền tảng, giúp giảm sự phụ thuộc vào dữ liệu do con người chủ thịch. 2) Phương pháp tiếp cận của chúng tôi hỗ trợ nhiều loại điều khiển đa dạng hơn và cung cấp các biểu diễn thống nhất cho cả điều khiển trực quan và ngôn ngữ, tạo điều kiện mở rộng để kết hợp các khía cạnh mới của khả năng điều khiển 3) Các thí nghiệm chứng minh khả năng tương tác mạnh mẽ với người dùng của CAT.

2 Công trình liên quan

Chú thích hình ảnh và chú thích dày đặc. Chú thích hình ảnh là một nhiệm vụ đa phương thức nhằm mục đích tạo ra các mô tả cho một hình ảnh nhất định. Các phương pháp tiếp cận phổ biến chủ yếu sử dụng mô hình mã hóa-giải mã [36, 44, 1, 12, 18, 17, 38] để giải quyết nhiệm vụ này. Để làm phong phú thêm sự hiểu biết chi tiết cho các cảnh phức tạp, chú thích dày đặc [13, 47, 19, 48, 31, 43] được đề xuất để tạo ra các chú thích cục bộ cho tất cả các đối tượng nổi bật trong một hình ảnh. Yin et al. [48] đề xuất một kỹ thuật chia sẻ thông tin theo ngữ cảnh đa thang để nắm bắt thông tin ngữ cảnh toàn cầu chi tiết. Wu et al. [43] đề xuất một khuôn khổ thống nhất cho chú thích dày đặc và phát hiện đối tượng, đạt được khả năng hiểu đối tượng tốt hơn. Tuy nhiên, các phương pháp này chỉ tạo ra các mô tả dựa trên các đặc điểm nổi bật của hình ảnh, khiến chúng không phù hợp với tương tác của người dùng vì chúng thiếu cơ chế kiểm soát rõ ràng phù hợp với ý định của người dùng. So với chúng, phương pháp của chúng tôi tập trung vào mô tả hình ảnh tương tác và hỗ trợ nhiều điều khiển đa phương thức khác nhau để cung cấp sự linh hoạt tốt hơn với người dùng.

Chú thích hình ảnh có thể kiểm soát. Khả năng có thể kiểm soát để tạo ra các mô tả mong muốn liên quan đến các đối tượng do người dùng chỉ định trong hình ảnh rất hữu ích trong các ứng dụng thực tế. Một mặt, các điều khiển trực quan trong chú thích hình ảnh có thể kiểm soát thường liên quan đến hộp giới hạn [5], hành vi của chuột [29, 46]. Cornia và cộng sự [5] đề xuất tạo chú thích tương ứng dựa trên một chuỗi hoặc một tập hợp các vùng hình ảnh. Pont-Tuset và cộng sự [29] phát hành một tập dữ liệu chú thích hình ảnh tương tác, trong đó người chú thích được yêu cầu mô tả vùng hình ảnh mà quỹ đạo chuột đã đi qua. LoopCAG [46] cải thiện thêm chất lượng tạo chú thích và khả năng kiểm soát tương tác. Mặt khác, kiểu văn bản của chú thích được tạo có thể được thay đổi linh hoạt theo các tình huống ứng dụng khác nhau. Thông thường, kiểu văn bản có thể kiểm soát có thể được tóm tắt như sau: khả năng kiểm soát độ dài [6], khả năng kiểm soát tinh cảm [25], khả năng kiểm soát tương tương (ví dụ: mô tả lãng mạn hoặc hài hước) [51, 10]. Một số tác phẩm [49, 39] thống nhất các phong cách có thể kiểm soát này thành một kiến trúc duy nhất. BLIP2 [17], một mô hình tiền đào tạo ngôn ngữ-hình ảnh, thực hiện tạo hình ảnh thành văn bản không cần chụp theo các lời nhắc ngôn ngữ tự nhiên. Trong bài báo này, chúng tôi mở rộng thêm tính linh hoạt của chú thích hình ảnh có thể kiểm soát, trong đó tương tác có thể là các điểm, hộp hoặc quỹ đạo do người dùng chỉ định và phong cách ngôn ngữ của các mô tả được tạo ra có thể đáp ứng các yêu cầu của người dùng ở mức độ lớn nhất.

Phân đoạn hình ảnh tương tác. Phân đoạn hình ảnh tương tác [22, 37, 45, 32] là một vấn đề nghiên cứu quan trọng trong đó mô hình được yêu cầu phân đoạn hình ảnh theo ý định của người dùng với tương tác (ví dụ: điểm, quỹ đạo). Xu et al. [45] đề xuất phân đoạn hình ảnh tương tác dựa trên học sâu trước tiên. Li et al trình bày một phân đoạn hình ảnh tương tác đầu cuối [21]. Toàn bộ kiến trúc được chia thành hai mảng tích chập, trong đó mảng đầu tiên được sử dụng để tổng hợp nhiều mặt nạ khác nhau theo đầu vào của người dùng và mảng thứ hai được đào tạo để chọn một giải pháp duy nhất trong số các mặt nạ này. FCA-Net [23] sử dụng cú nhấp chuột đầu tiên tốt hơn để cải thiện kết quả phân đoạn tương tác. SAM [14] xây dựng một mô hình nền tảng có thể nhắc nhớ để phân đoạn, trong đó lời nhắc có thể là bất kỳ thông tin nào cho biết nội dung cần phân đoạn trong hình ảnh, ví dụ: một tập hợp các điểm, hộp hoặc mặt nạ hoặc văn bản dạng tự do.

Các mô hình ngôn ngữ lớn. LLM đã tạo nên sự quan tâm đáng kể do khả năng chuyển giao mạnh mẽ của chúng trên nhiều loại nhiệm vụ xử lý ngôn ngữ khác nhau, cũng như khả năng tương tác với con người mới nổi của chúng. Gần đây, bước đột phá quan trọng nhất đã được thực hiện bởi GPT-3 [2], một mô hình có 175B tham số, đã tiết lộ tiềm năng mới nổi của các kỹ thuật học tập ít lần. Hiệu suất đáng chú ý này đã thúc đẩy rất nhiều LLM tiếp theo [50, 4, 35, 30]. Để nâng cao

the interactive capacity of LLMs with human users, state-of-the-art models are commonly fine-tuned with human feedback [28, 33, 26, 27, 34]. This approach enables LLMs to effectively adapt to diverse language instructions, thereby aligning more closely with user intent. Inspired by the user alignment capabilities of LLMs, we unify language controls to textual prompts and apply LLMs to generate image descriptions that align with user preferences.

3 Caption Anything

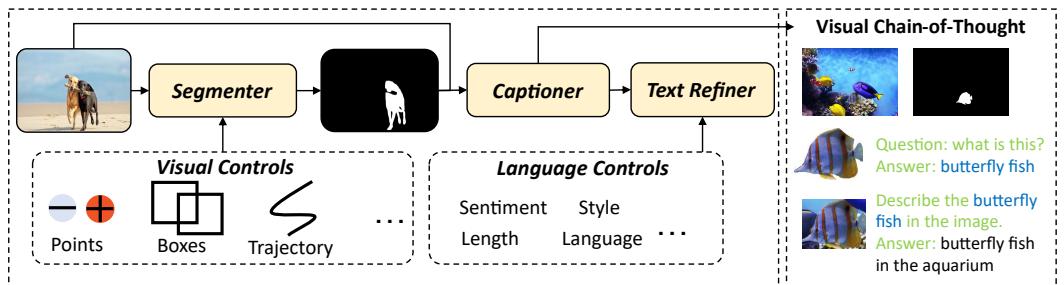


Figure 3: The overall framework of Caption Anything. It introduces multimodal controls to image captioning, rendering a variety of visual focuses and language styles aligned with human intention. The visual prompt is firstly converted into the mask prompt by the *segmenter*. Subsequently, the *captioner* predicts a raw caption for the region delineated by the mask. To make the *captioner* focus on the user-interested object, we use a simple visual chain-of-thought technique to conduct step-by-step inference. Finally, both the text prompt and the raw caption are fed into the *text refiner*, which generates a user-preferred caption in accordance with the desired genre.

To enhance the user-centric interactivity of current image captioning systems, we propose a foundational model augmentation strategy to accommodate image captioners with a variety of multimodal controls. Specifically, our approach could be formulated as a triplet solver {*segmenter*, *captioner*, *text refiner*}. As illustrated in Fig. 3, *segmenter* first takes the interactive visual controls (e.g., points, boxes, trajectory) and represents the user-interested regions via pixel-level masks. Subsequently, the *captioner* generates raw descriptions in relation to the specified region based on the original image and the provided mask. In order to facilitate the *captioner* focus on the user-interested object, we design a visual chain-of-thought technique with step-by-step inference. Lastly, *text refiner* refines the raw descriptions by incorporating user-defined language controls, thereby tailoring the language style according to user preferences.

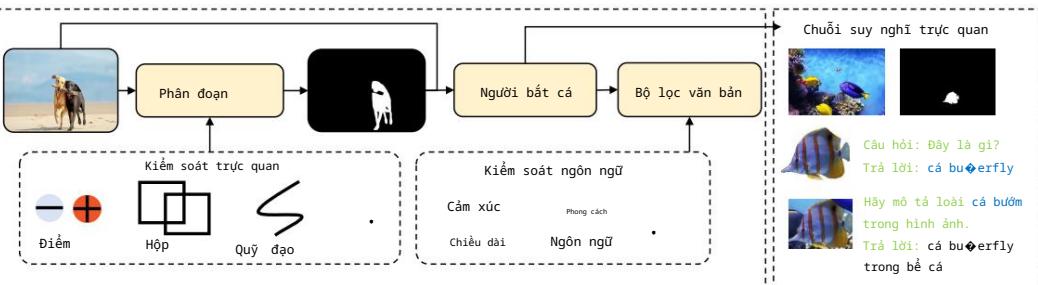
Segmenter. The ideal *segmenter* is capable of segmenting any part of an image according to the visual controls. SAM [14] meets the requirements well and has impressive zero-shot transferability to new image domain, benefiting from promptable pre-training and the SA-1B dataset¹ (the largest segmentation dataset with 1 billion masks on 11M images). SAM adapts interactive segmentation [24] to achieve promptable ability, where a prompt, any interaction (e.g., points, boxes) indicating what to segment in an image, is used to prompt SAM to return a valid segmentation mask. Once we obtain the user-specified segmentation mask, it is easy to generate the desired caption according to the original image and mask prompt.

Captioner. To describe any user-specific object in the image, the *captioner* is expected to perform strong zero-shot captioning performance. In other words, the ideal *captioner* should generate reasonable descriptions among various novel objects and different image distributions. We use BLIP2 as the captioner. It leverages frozen pre-trained image encoders and frozen LLMs together with a querying transformer to bridge the modality gap, achieving excellent zero-shot performance.

Text Refiner. In most cases, image-related descriptions should follow users' preferences. However, refining the raw caption generated by the *captioner* based on user instructions is a nontrivial task. To achieve this goal, we introduce ChatGPT as an API to generate more expressive and controllable

khả năng tương tác của LLM với người dùng, các mô hình tiên tiến thường được tinh chỉnh với phản hồi của con người [28, 33, 26, 27, 34]. Cách tiếp cận này cho phép LLM thích ứng hiệu quả với các hướng dẫn ngôn ngữ đa dạng, do đó phù hợp hơn với ý định của người dùng. Lấy cảm hứng từ khả năng cẩn chỉnh người dùng của LLM, chúng tôi hợp nhất các điều khiển ngôn ngữ thành lời nhắc văn bản và áp dụng LLM để tạo mô tả hình ảnh phù hợp với sở thích của người dùng.

3 Chú thích bắt cứ điều gì



Hình 3: Khung tổng thể của Caption Anything. Nó giới thiệu các điều khiển đa phương thức cho chú thích hình ảnh, tạo ra nhiều tiêu điểm trực quan và phong cách ngôn ngữ phù hợp với ý định của con người. Đầu tiên, lời nhắc trực quan được chuyển đổi thành lời nhắc mặt nạ bởi bộ phân đoạn. Sau đó, bộ chú thích dự đoán một chú thích thô cho vùng được phân định bởi mặt nạ. Để làm cho bộ chú thích tập trung vào đối tượng mà người dùng quan tâm, chúng tôi sử dụng một kỹ thuật chuỗi suy nghĩ trực quan đơn giản để thực hiện suy luận từng bước. Cuối cùng, cả lời nhắc văn bản và chú thích thô đều được đưa vào bộ tinh chỉnh văn bản, bộ này sẽ tạo ra một chú thích mà người dùng ưa thích theo thể loại mong muốn.

Để tăng cường tính tương tác lấy người dùng làm trung tâm của các hệ thống chú thích hình ảnh hiện tại, chúng tôi đã xuất một chiến lược tăng cường mô hình cơ bản để hỗ trợ người chú thích hình ảnh với nhiều điều khiển đa phương thức. Cụ thể, cách tiếp cận của chúng tôi có thể được xây dựng như một bộ giải bài ba {*segmenter*, *captioner*, *text refiner*}. Như minh họa trong Hình 3, trước tiên, *segmenter* sẽ lấy các điều khiển trực quan tương tác (ví dụ: điểm, hộp, quỹ đạo) và biểu diễn các vùng mà người dùng quan tâm thông qua mặt nạ cấp pixel. Sau đó, *captioner* tạo ra các mô tả thô liên quan đến vùng đã chỉ định dựa trên hình ảnh gốc và mặt nạ được cung cấp. Để tạo điều kiện cho người chú thích tập trung vào đối tượng mà người dùng quan tâm, chúng tôi thiết kế một kỹ thuật chuỗi suy nghĩ trực quan với suy luận từng bước. Cuối cùng, *text refiner* tinh chỉnh các mô tả thô bằng cách kết hợp các điều khiển ngôn ngữ do người dùng xác định, do đó điều chỉnh phong cách ngôn ngữ theo sở thích của người dùng.

Bộ phân đoạn. Bộ phân đoạn lý tưởng có khả năng phân đoạn bất kỳ phần nào của hình ảnh theo các điều khiển trực quan. SAM [14] đáp ứng tốt các yêu cầu và có khả năng chuyển đổi không cần chỉnh sửa ẩn tượng sang miền hình ảnh mới, được hưởng lợi từ quá trình đào tạo trước có thể nhắc nhở và tập dữ liệu SA-1B (tập dữ liệu phân đoạn lớn nhất với 1 tỷ mặt nạ trên 11 triệu hình ảnh). SAM điều chỉnh phân đoạn tương tác [24] để đạt được khả năng có thể nhắc nhở, trong đó lời nhắc, bất kỳ tương tác nào (ví dụ: điểm, hộp) chỉ ra phần nào cần phân đoạn trong hình ảnh, được sử dụng để nhắc SAM trả về mặt nạ phân đoạn hợp lệ. Khi chúng tôi có được mặt nạ phân đoạn do người dùng chỉ định, thật dễ dàng để tạo chú thích mong muốn theo hình ảnh gốc và lời nhắc mặt nạ.

Captioner. Để mô tả bất kỳ đối tượng cụ thể nào của người dùng trong hình ảnh, *captioner* được kỳ vọng sẽ thực hiện hiệu suất captioning zero-shot mạnh mẽ. Nói cách khác, *captioner* lý tưởng phải tạo ra các mô tả hợp lý giữa nhiều đối tượng mới lạ và các phân phối hình ảnh khác nhau. Chúng tôi sử dụng BLIP2 làm *captioner*. Nó tận dụng các bộ mã hóa hình ảnh được đào tạo trước đồng lạnh và LLM đồng lạnh cùng với một bộ chuyển đổi truy vấn để thu hẹp khoảng cách phương thức, đạt được hiệu suất zero-shot tuyệt vời.

Trình tinh chỉnh văn bản. Trong hầu hết các trường hợp, mô tả liên quan đến hình ảnh phải tuân theo sở thích của người dùng. Tuy nhiên, việc tinh chỉnh chú thích thô do trình tạo chú thích tạo ra dựa trên hướng dẫn của người dùng là một nhiệm vụ không hề đơn giản. Để đạt được mục tiêu này, chúng tôi giới thiệu ChatGPT như một API để tạo ra các cuộc trò chuyện biểu cảm và có thể kiểm soát được hơn

¹The project link: <https://segment-anything.com>



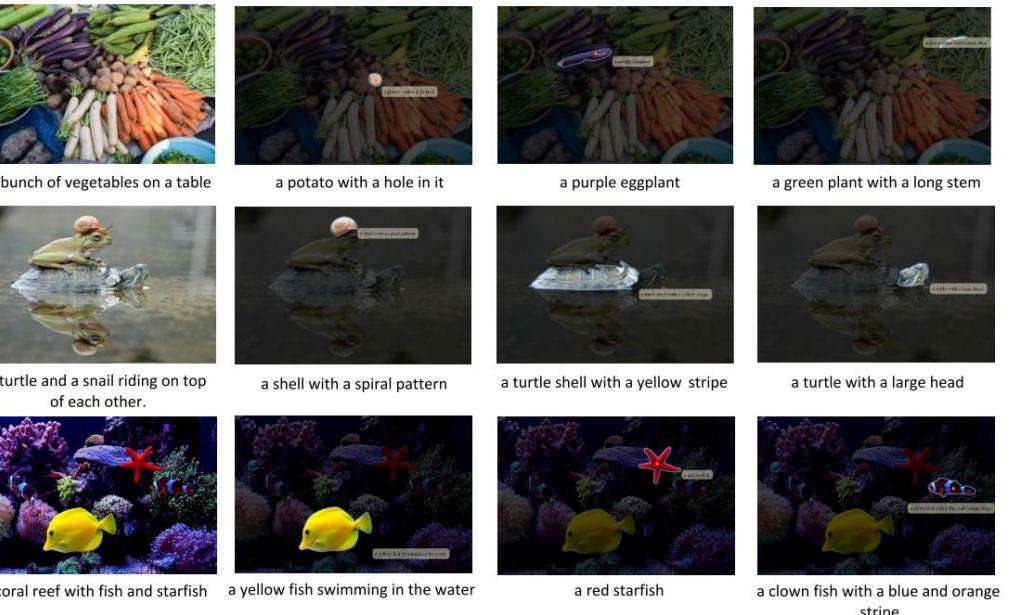
Figure 4: Visualization of describing the image with point-based visual controls.

descriptions from the raw caption. The *text refiner* can be replaced with open-source alternative LLMs easily, such as LLaMA [35], OPT-IML [50], BLOOM [30].

Visual Chain-of-Thought. We empirically found that the generated object captions are easily affected by the background information. Inspired by the chain-of-thought (CoT) prompting originated from NLP [41, 15], we bootstrap the *captioner* with step-by-step text generation to ensure that the generated description focuses on the user-selected region. As Fig. 3 shows, we first retain the user-selected object and replace the background with white and ask the *captioner* to identify the category of the interested object. Subsequently, the *captioner* takes the generated text and the cropped image with the background as a prompt to generate the final caption. In this way, the *captioner* is capable of focusing on the selected object during the caption generation process.

Extension to Object-Centric Chatting. Drawing inspiration from recent advancements in multimodal dialog systems, we investigate the potential of visual dialog specifically targeting objects identified by visual prompts. Unlike dominant chat systems that are designed for a global comprehension of the entire image, local region chat has broader applications for complex, informative, and high-resolution images. This includes visual navigation, embodied intelligence, and early-child education. Given the segmentation mask of an object and user query, we use an off-the-shelf visual question answering model [17] as a visual API that empowers the ChatGPT to understand detailed visual cues by asking questions. Specifically, we include the generated caption to the initial prompt and follow [42] to use LangChain [16] as the control hub to predict the chain of API calls.

Extension to Paragraph Captioning. Our framework is adaptable to the image paragraph captioning task by leveraging ChatGPT to summarize the dense captions and scene texts of an image into a paragraph. Specifically, we generate dense captions by initially using the SAM to segment everything within the image, followed by captioning each object with the CAT pipeline. To incorporate scene text information into the paragraph, we utilize additional OCR tools (*e.g.*, EasyOCR [7]) to identify the text present in the image. The dense captions and scene texts are subsequently merged into a predefined prompt template, which is then employed to instruct ChatGPT in summarizing the scene information into a cohesive paragraph.



Hình 4: Hình ảnh mô tả hình ảnh bằng các điều khiển trực quan dựa trên điểm.

mô tả từ chủ thích thô. Bộ lọc văn bản có thể được thay thế dễ dàng bằng các LLM thay thế nguồn mở, chẳng hạn như LLaMA [35], OPT-IML [50], BLOOM [30].

mô tả từ chủ thích thô. Bộ lọc văn bản có thể được thay thế dễ dàng bằng các LLM thay thế nguồn mở, chẳng hạn như LLaMA [35], OPT-IML [50], BLOOM [30].

Chuỗi suy nghĩ trực quan. Chúng tôi thấy rằng chủ thích đối tượng được tạo ra dễ bị ảnh hưởng bởi thông tin nền. Lấy cảm hứng từ lời nhắc chuỗi suy nghĩ (CoT) có nguồn gốc từ NLP [41, 15], chúng tôi khởi động trình chủ thích với việc tạo văn bản từng bước để đảm bảo rằng mô tả được tạo ra tập trung vào vùng do người dùng chọn. Như Hình 3 cho thấy, trước tiên chúng tôi giữ nguyên đối tượng do người dùng chọn và thay thế nền bằng màu trắng và yêu cầu người chủ thích xác định danh mục của đối tượng quan tâm. Sau đó, người chủ thích lấy văn bản được tạo ra và hình ảnh được cắt có nền làm lời nhắc để tạo chủ thích cuối cùng. Theo cách này, người chủ thích có khả năng tập trung vào đối tượng được chọn trong quá trình tạo chủ thích.

Mở rộng cho Trò chuyện lấy Đối tượng làm trung tâm. Lấy cảm hứng từ những tiến bộ gần đây trong các hệ thống đối thoại đa phương thức, chúng tôi nghiên cứu tiềm năng của hộp thoại trực quan nhằm mục tiêu cụ thể vào các đối tượng được xác định bằng lời nhắc trực quan. Không giống như các hệ thống trò chuyện thông thường được thiết kế để hiểu toàn bộ hình ảnh, trò chuyện theo vùng cục bộ có ứng dụng rộng hơn cho hình ảnh phức tạp, nhiều thông tin và có độ phân giải cao. Điều này bao gồm điều hướng trực quan, trí thông minh nhân tạo và giáo dục trẻ em. Với mặt nạ phân đoạn của đối tượng và truy vấn của người dùng, chúng tôi sử dụng mô hình trả lời câu hỏi trực quan có sẵn [17] làm API trực quan giúp ChatGPT hiểu các tín hiệu trực quan chi tiết bằng cách đặt câu hỏi. Cụ thể, chúng tôi bao gồm chủ thích được tạo vào lời nhắc ban đầu và theo [42] để sử dụng LangChain [16] làm trung tâm điều khiển để dự đoán chuỗi lệnh gọi API.

Mở rộng cho chủ thích đoạn văn. Khung của chúng tôi có thể thích ứng với nhiệm vụ chủ thích đoạn văn hình ảnh bằng cách tận dụng ChatGPT để tóm tắt các chủ thích dày đặc và văn bản cảnh của một hình ảnh thành một đoạn văn. Cụ thể, chúng tôi tạo các chủ thích dày đặc bằng cách ban đầu sử dụng SAM để phân đoạn mọi thứ trong hình ảnh, sau đó chủ thích từng đối tượng bằng đường ống CAT. Để kết hợp thông tin văn bản cảnh vào đoạn văn, chúng tôi sử dụng các công cụ OCR bổ sung (*ví dụ*: EasyOCR [7]) để xác định văn bản có trong hình ảnh. Sau đó, các chủ thích dày đặc và văn bản cảnh được hợp nhất thành một mẫu nhắc được xác định trước, sau đó được sử dụng để hướng dẫn ChatGPT tóm tắt thông tin cảnh thành một đoạn văn gắn kết.

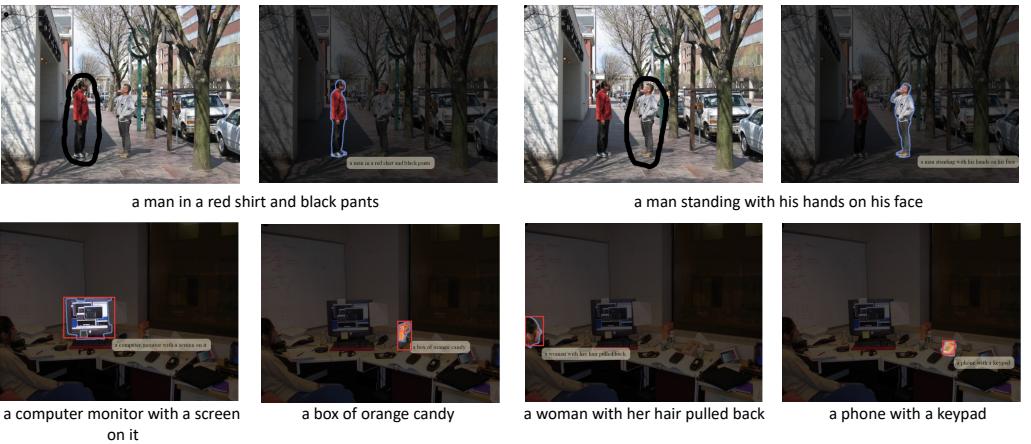


Figure 5: Visualization of describing the image with visual controls (trajectory and box).

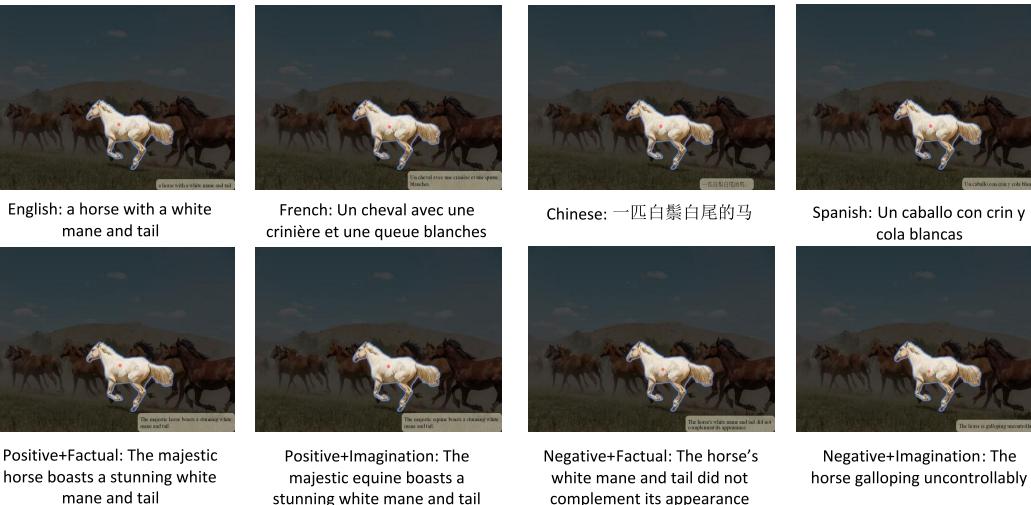


Figure 6: Visualization of generated captions with language controls (first row: multiple languages; second row: sentiment and factuality).

4 Experiments

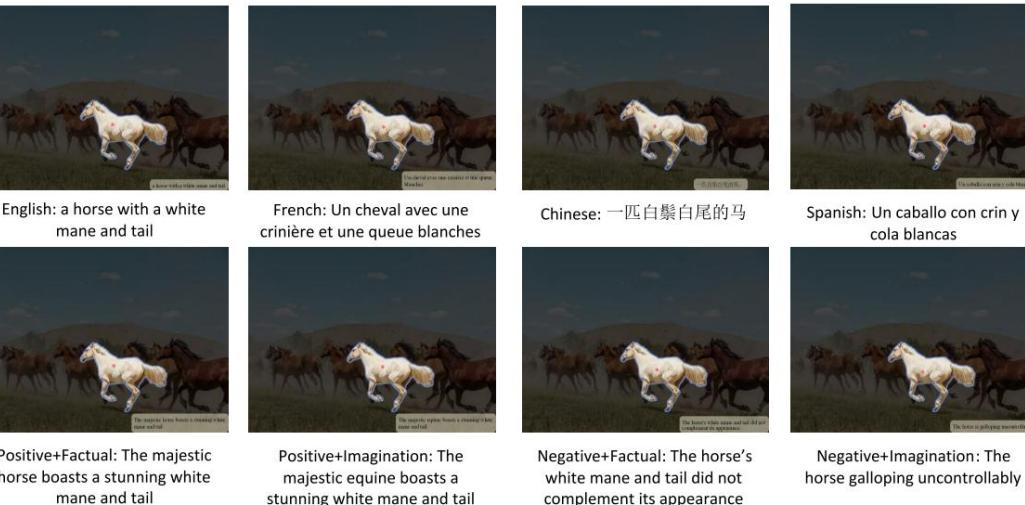
Visual Controls. As shown in Fig. 4, we present some qualitative results that showcase the remarkable visual control capabilities of CAT. By placing click-point prompts in various locations within the images, CAT is able to accurately identify and describe corresponding objects, demonstrating its exceptional ability to caption a diverse range of objects in any given image. Furthermore, as shown in Fig. 5, CAT’s visual controls can be trajectory-based or bounding box-based, further highlighting its versatility and adaptability in generating accurate descriptions for a wide range of image content.

Language Controls. We provide additional qualitative results that showcase the impressive language control capabilities of our model, as shown in Fig. 6. Leveraging its advanced language controls (*i.e.*, sentiment and factuality), CAT can generate captions with a diverse range of language styles, ranging from casual and conversational to formal and informative. These results highlight the model’s remarkable ability to adapt its language output based on the specific needs and preferences of the user.

Object-Centric Chatting. In Fig. 7, we present further evidence of CAT’s exceptional object-centric chatting ability. Specifically, our model is capable of performing visual question answering



Hình 5: Hình ảnh mô tả hình ảnh bằng các điều khiển trực quan (quỹ đạo và hộp).



Hình 6: Hình ảnh hóa phụ đề được tạo với các điều khiển ngôn ngữ (hàng đầu tiên: nhiều ngôn ngữ; hàng thứ hai: tình cảm và thực tế).

4 Thí nghiệm

Kiểm soát trực quan. Như được thể hiện trong Hình 4, chúng tôi trình bày một số kết quả định tính cho thấy khả năng kiểm soát trực quan đáng chú ý của CAT. Bằng cách đặt lời nhắc diếm nhập ở nhiều vị trí khác nhau trong hình ảnh, CAT có khả năng xác định và mô tả chính xác các đối tượng tương ứng, chứng minh khả năng đặc biệt để chú thích một loạt các đối tượng khác nhau trong bất kỳ hình ảnh nào. Hơn nữa, như đã hiển thị trong Hình 5, các điều khiển trực quan của CAT có thể dựa trên quỹ đạo hoặc dựa trên hộp giới hạn, làm nổi bật tính linh hoạt và khả năng thích ứng trong việc tạo ra các mô tả chính xác cho nhiều nội dung hình ảnh.

Kiểm soát ngôn ngữ. Chúng tôi cung cấp các kết quả định tính bổ sung cho thấy khả năng kiểm soát ngôn ngữ ám tương của mô hình của chúng tôi, như thể hiện trong Hình 6. Tận dụng ngôn ngữ tiên tiến của nó kiểm soát (tức là tình cảm và tính thực tế), CAT có thể tạo chú thích với nhiều ngôn ngữ khác nhau phong cách, từ bình thường và đàm thoại đến trang trọng và thông tin. Những kết quả này làm nổi bật khả năng đáng chú ý của mô hình trong việc điều chỉnh đầu ra ngôn ngữ của nó dựa trên nhu cầu và sở thích cụ thể của người dùng.

Trò chuyện lấy đối tượng làm trung tâm. Trong Hình 7, chúng tôi trình bày thêm chứng về khả năng trò chuyện lấy đối tượng làm trung tâm đặc biệt của CAT. Cụ thể, mô hình của chúng tôi có khả năng thực hiện trả lời câu hỏi trực quan



Figure 7: Examples of object-centric chatting.



Hình 7: Ví dụ về trò chuyện lấy đối tượng làm trung tâm.

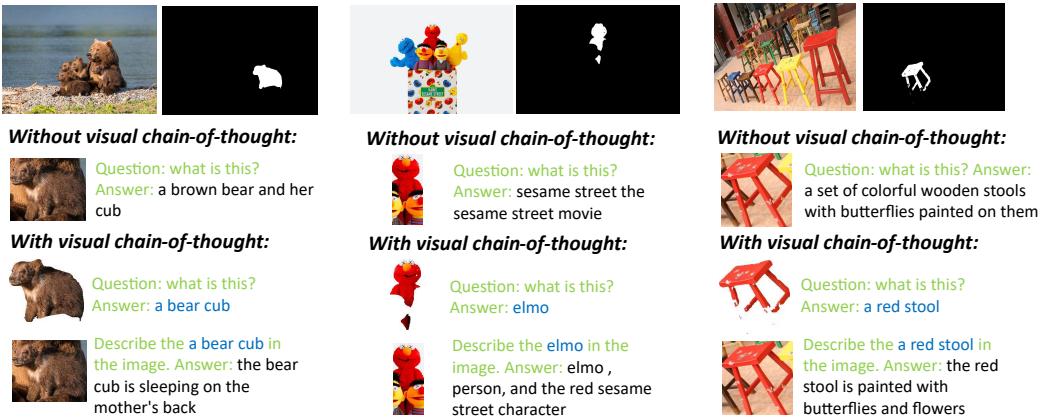
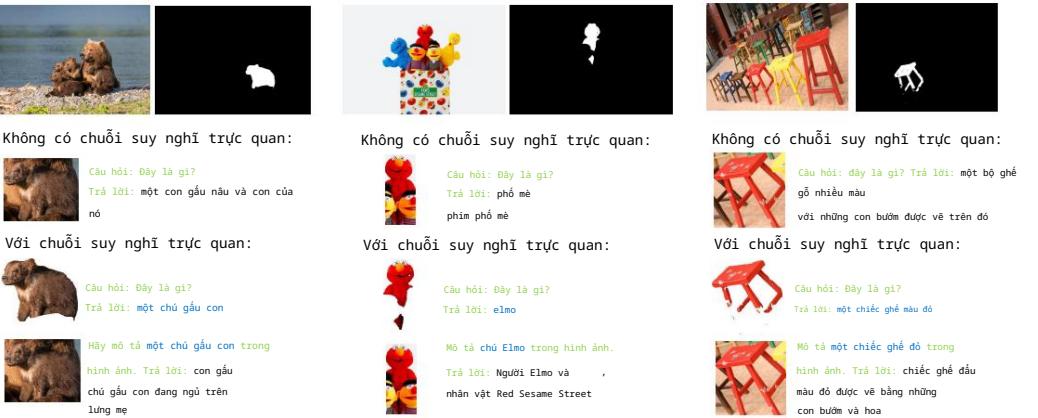


Figure 8: Examples of visual chain-of-thought.



Hình 8: Ví dụ về chuỗi suy nghĩ trực quan.

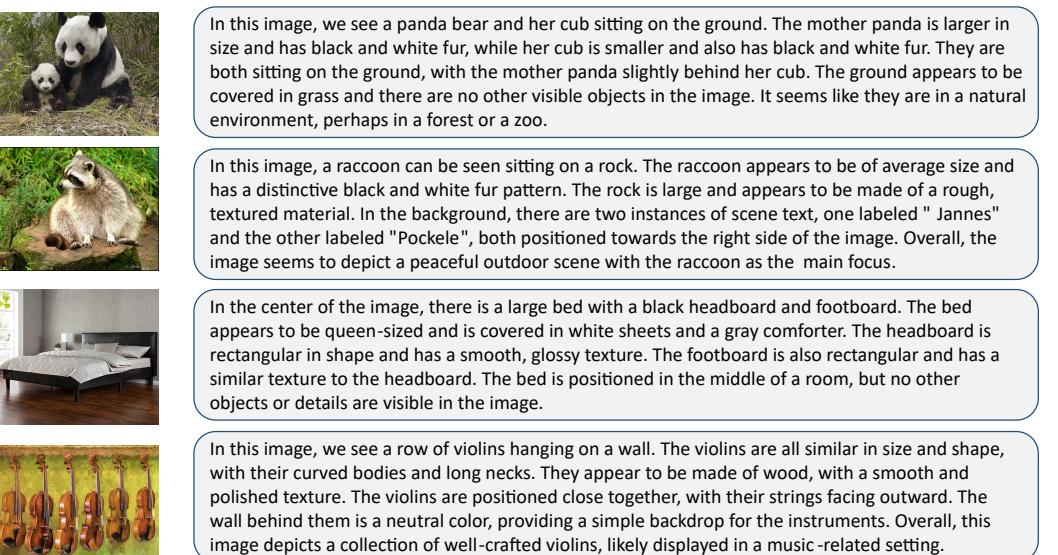
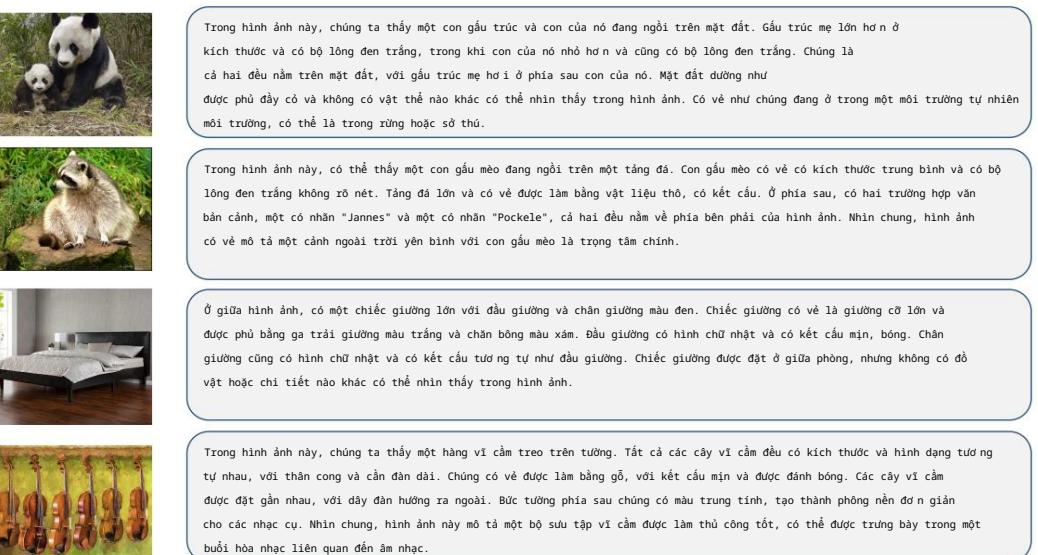


Figure 9: Examples of image paragraph captioning.



Hình 9: Ví dụ về chủ thích đoạn văn trong hình ảnh.

focused around selected objects within an image, showcasing its remarkable ability to engage in rich and meaningful conversations centered around specific visual content.

Visual Chain-of-Thought. Fig. 8 presents a variety of examples illustrating the efficacy of the visual CoT. As shown in this figure, performing direct inference on images containing backgrounds tends to be adversely influenced by the background content, thus hindering the captioner’s ability to concentrate on the interested object. By incorporating a step-by-step thought process, the generated captions exhibit enhanced focus. Concurrently, the visual chain-of-thought facilitates the disclosure of more intricate details pertaining to the object of interest, thereby enabling the entire system to more effectively align with the user’s intent.

Caption Everything in a Paragraph. Fig. 9 presents several examples of the resultant paragraphs. These generated paragraphs encompass the majority of the objects and textual information in the scene and even include some reasoning information. The overall descriptions provided by the paragraphs are accurate in general and logically coherent.

5 Conclusion

In this paper, we propose a foundation model augmented controllable image captioning framework, Caption AnyThing (CAT), that addresses the limitations of existing CIC approaches. Our proposed framework leverages pre-trained image captioners and integrate them with the SAM and an instruction-tuned LLM, thereby mitigating the reliance on human-annotated data and expanding the range of supported control signals. The unified representation of control signals in CAT enhances the model’s flexibility and scalability, making it easily adaptable for interactive use and extension with new aspects of controllability. The experiments demonstrate that CAT provides a training-free and adaptable solution for controllable image captioning tasks, offering strong user-interactive capabilities.

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3
- [3] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16856, 2021. 2
- [4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2, 3
- [5] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019. 2, 3
- [6] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. In *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 712–729. Springer, 2020. 2, 3
- [7] EasyOCR. <https://www.jaisted.ai/easyocr>, 2021. 5
- [8] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18009–18019, 2022. 2

tập trung vào các đối tượng được chọn trong hình ảnh, thể hiện khả năng đáng chú ý của nó trong việc tham gia vào các cuộc trò chuyện phong phú và có ý nghĩa xoay quanh nội dung hình ảnh cụ thể.

Chuỗi suy nghĩ trực quan. Hình 8 trình bày nhiều ví dụ minh họa hiệu quả của CoT trực quan. Như thể hiện trong hình này, việc thực hiện suy luận trực tiếp trên hình ảnh có chứa nền có xu hướng bị ảnh hưởng xấu bởi nội dung nền, do đó cần trở khả năng tập trung vào đối tượng quan tâm của người chủ thích. Bằng cách kết hợp quy trình suy nghĩ từng bước, các chủ thích được tạo ra thể hiện sự tập trung được tăng cường. Đồng thời, chuỗi suy nghĩ trực quan tạo điều kiện cho việc tiết lộ các chi tiết phức tạp hơn liên quan đến đối tượng quan tâm, do đó cho phép toàn bộ hệ thống phù hợp hơn với ý định của người dùng.

Chú thích Mọi thứ trong một đoạn văn. Hình 9 trình bày một số ví dụ về các đoạn văn kết quả. Những đoạn văn được tạo ra này bao gồm phần lớn các đối tượng và thông tin văn bản trong cảnh và thậm chí bao gồm một số thông tin lý luận. Các mô tả tổng thể do các đoạn văn cung cấp nhìn chung là chính xác và mạch lạc về mặt logic.

5 Kết luận

Trong bài báo này, chúng tôi đề xuất một khuôn khổ chú thích hình ảnh có thể điều khiển được tăng cường mô hình nền tảng, Caption Anything (CAT), giải quyết các hạn chế của các phương pháp tiếp cận CIC hiện có. Khuôn khổ đề xuất của chúng tôi tận dụng các chú thích hình ảnh được đào tạo trước và tích hợp chúng với SAM và LLM được điều chỉnh theo hướng dẫn, do đó giảm thiểu sự phụ thuộc vào dữ liệu do con người chủ thích và mở rộng phạm vi các tín hiệu điều khiển được hỗ trợ. Biểu diễn thống nhất các tín hiệu điều khiển trong CAT tăng cường tính linh hoạt và khả năng mở rộng của mô hình, giúp dễ dàng thích ứng để sử dụng tương tác và mở rộng với các khía cạnh mới về khả năng điều khiển. Các thí nghiệm chứng minh rằng CAT cung cấp một giải pháp không cần đào tạo và có thể thích ứng cho các tác vụ chú thích hình ảnh có thể điều khiển được, mang lại khả năng tương tác mạnh mẽ với người dùng.

Tài liệu tham khảo

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould và Lei Zhang. Sự chú ý từ dưới lên và từ trên xuống để chú thích hình ảnh và trả lời câu hỏi trực quan. Trong *Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu*, trang 6077–6086, 2018. 3
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Các mô hình ngôn ngữ là những người học ít lần. *Những tiến bộ trong hệ thống xử lý thông tin thần kinh*, 33:1877–1901, 2020. 3
- [3] Long Chen, Zhihong Jiang, Jun Xiao và Wei Liu. Chú thích hình ảnh có thể điều khiển giống con người với vai trò ngữ nghĩa cụ thể của động từ. Trong *Biên bản báo cáo Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu*, trang 16846–16856, 2021. 2
- [4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, và những người khác. Mở rộng các mô hình ngôn ngữ được tinh chỉnh theo hướng dẫn. *bản in trước arXiv arXiv:2210.11416*, 2022. 2, 3
- [5] Marcella Cornia, Lorenzo Baraldi và Rita Cucchiara. Hiển thị, kiểm soát và kể: Một khuôn khổ để tạo ra các chú thích có thể kiểm soát và có cơ sở. Trong *Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu*, trang 8307–8316, 2019. 2, 3
- [6] Đặng Chaorui, Ning Ding, Mingkui Tan và Qi Wu. Chú thích hình ảnh có thể kiểm soát độ dài. Trong *Computer Vision-ECCV 2020: Hội nghị Châu Âu lần thứ 16, Glasgow, Vương quốc Anh, ngày 23–28 tháng 8 năm 2020, Kỷ yếu, Phần XIII 16*, trang 712–729. Mùa xuân, 2020. 2, 3
- [7] EasyOCR. <https://www.jaisted.ai/easyocr>, 2021. 5
- [8] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang và Zicheng Liu. Đưa các khái niệm ngữ nghĩa vào chú thích hình ảnh đầu cuối. Trong *Biên bản báo cáo của hội nghị IEEE/CVF về thị giác máy tính và nhận dạng mẫu*, trang 18009–18019, 2022. 2

- [9] Zhengcong Fei, Xu Yan, Shuhui Wang, and Qi Tian. Deecap: dynamic early exiting for efficient image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216–12226, 2022. 2
- [10] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3137–3146, 2017. 2, 3
- [11] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 2
- [12] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019. 3
- [13] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 2, 3
- [14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3, 4
- [15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022. 5
- [16] LangChain. <https://github.com/hwchase17/langchain>, 2022. 5
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 5
- [18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning*, pages 12888–12900, 2022. 2, 3
- [19] Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8650–8657, 2019. 3
- [20] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17990–17999, 2022. 2
- [21] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–585, 2018. 3
- [22] JunHao Liew, Yunchao Wei, Wei Xiong, Sim-Heng Ong, and Jiashi Feng. Regional interactive image segmentation networks. *2017 IEEE international conference on computer vision (ICCV)*, pages 2746–2754, 2017. 3
- [23] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13339–13348, 2020. 3
- [24] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. *arXiv preprint arXiv:1805.04398*, 2018. 4
- [25] Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 2, 3
- [9] Zhengcong Fei, Xu Yan, Shuhui Wang và Qi Tian. Deecap: thoát sớm động để chú thích hình ảnh hiệu quả. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Tâm nhìn máy tính và Nhận dạng mẫu, trang 12216-12226, 2022. 2
- [10] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao và Li Deng. Stylenet: Tạo chú thích trực quan hấp dẫn với các kiểu. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 3137-3146, 2017. 2, 3
- [11] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zichen Liu, Yumao Lu và Lijuan Wang. Mở rộng quy mô đào tạo trước ngôn ngữ thị giác cho chú thích hình ảnh. Trong Kỷ yếu của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 17980-17989, 2022. 2
- [12] Lun Huang, Wenmin Wang, Jie Chen và Xiao-Yong Wei. Chú ý đến chú ý cho chú thích hình ảnh. Trong Biên bản báo cáo hội nghị quốc tế IEEE/CVF về tâm nhìn máy tính, trang 4634-4643, 2019. 3
- [13] Justin Johnson, Andrej Karpathy và Li Fei-Fei. Densecap: Mạng định vị tích chập hoàn toàn cho chú thích dày đặc. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 4565-4574, 2016. 2, 3
- [14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár và Ross Girshick. Phân đoạn bắt cứ thứ gì. bản in trước arXiv arXiv:2304.02643, 2023. 2, 3, 4
- [15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo và Yusuke Iwasawa. Các mô hình ngôn ngữ lớn là những nhà lý luận không cần phải bắn. bản in trước arXiv arXiv:2205.11916, 2022. 5
- [16] LangChain. <https://github.com/hwchase17/langchain>, 2022. 5
- [17] Junnan Li, Dongxu Li, Silvio Savarese và Steven Hoi. Blip-2: Khởi động quá trình đào tạo trước ngôn ngữ-hình ảnh với bộ mã hóa hình ảnh đông lạnh và các mô hình ngôn ngữ lớn. Bản in trước arXiv arXiv:2301.12597, 2023. 2, 3, 5
- [18] Junnan Li, Dongxu Li, Caiming Xiong và Steven Hoi. Blip: Khởi động quá trình đào tạo trước ngôn ngữ-hình ảnh để hiểu và tạo ra ngôn ngữ-thị giác thống nhất. Hội nghị quốc tế về học máy, trang 12888-12900, 2022. 2, 3
- [19] Xiangyang Li, Shuqiang Jiang, và Jungong Han. Ngữ cảnh đối tượng học tập cho chú thích dày đặc. Trong Biên bản báo cáo hội nghị AAAI về trí tuệ nhân tạo, tập 33, trang 8650-8657, 2019. 3
- [20] Yehao Li, Yingwei Pan, Ting Yao và Tao Mei. Hiểu và sắp xếp ngữ nghĩa cho chú thích hình ảnh. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Tâm nhìn máy tính và Nhận dạng mẫu, trang 17990-17999, 2022. 2
- [21] Zhuwen Li, Qifeng Chen và Vladlen Koltun. Phân đoạn hình ảnh tương tác với sự đa dạng tiềm ẩn. Biên bản báo cáo Hội nghị IEEE về Tâm nhìn máy tính và Nhận dạng mẫu, trang 577-585, 2018. 3
- [22] JunHao Liew, Yunchao Wei, Wei Xiong, Sim-Heng Ong và Jiashi Feng. Mạng phân đoạn hình ảnh tương tác khu vực. Hội nghị quốc tế IEEE 2017 về thị giác máy tính (ICCV), trang 2746-2754, 2017. 3
- [23] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng và Shao-Ping Lu. Phân đoạn hình ảnh tương tác với sự chú ý nhập chuột đầu tiên. Biên bản hội nghị IEEE/CVF về thị giác máy tính và nhận dạng mẫu, trang 13339-13348, 2020. 3
- [24] Sabarinath Mahadevan, Paul Voigtlaender và Bastian Leibe. Phân đoạn tương tác được đào tạo lặp đi lặp lại. Bản in trước arXiv arXiv:1805.04398, 2018. 4
- [25] Alexander Mathews, Lexing Xie và Xuming He. Senticap: Tạo mô tả hình ảnh bằng cảm xúc. Trong Biên bản báo cáo hội nghị AAAI về trí tuệ nhân tạo, tập 30, 2016. 2, 3

- [26] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022. 4
- [27] OpenAI. Gpt-4 technical report, 2023. 2, 4
- [28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 4
- [29] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664, 2020. 2, 3
- [30] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 3, 5
- [31] Zhuang Shao, Jungong Han, Demetris Marnerides, and Kurt Debattista. Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3
- [32] Konstantin Sofiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking back-propagating refinement for interactive segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 3
- [33] Rohan Taori, Ishaaq Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 4
- [34] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poultton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 4
- [35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 5
- [36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 3
- [37] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018. 3
- [38] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2, 3
- [39] Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. Controllable image captioning via prompting. *arXiv preprint arXiv:2212.01803*, 2022. 3
- [40] Teng Wang, Yixiao Ge, Feng Zheng, Ran Cheng, Ying Shan, Xiaohu Qie, and Ping Luo. Accelerating vision-language pretraining with free language modeling. *arXiv preprint arXiv:2303.14038*, 2023. 2
- [26] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Tổng quát hóa xuyên ngôn ngữ thông qua tinh chỉnh đa nhiệm vụ. Bản in trước arXiv arXiv:2211.01786, 2022. 4
- [27] OpenAI. Báo cáo kỹ thuật Gpt-4, 2023. 2, 4
- [28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Đào tạo các mô hình ngôn ngữ để tuân theo hướng dẫn với phản hồi của con người. Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, 35:27730–27744, 2022. 4
- [29] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut và Vittorio Ferrari. Kết nối tầm nhìn và ngôn ngữ với các câu chuyện địa phương. Trong Computer Vision-ECCV 2020: Hội nghị châu Âu lần thứ 16, Glasgow, Vương quốc Anh, 23-28 tháng 8 năm 2020, Biên bản báo cáo, Phần V 16, trang 647-664, 2020. 2, 3
- [30] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, và những người khác. Bloom: Mô hình ngôn ngữ đa ngôn ngữ truy cập mở tham số 176b. bản in trước arXiv arXiv:2211.05100, 2022. 3, 5
- [31] Zhuang Shao, Jungong Han, Demetris Marnerides và Kurt Debattista. Chú thích dày đặc nhận biết mối quan hệ đối tượng-khu vực thông qua bộ chuyển đổi. Giao dịch IEEE về mạng nơ-ron và hệ thống học tập, 2022. 3
- [32] Konstantin Sofiuk, Ilia Petrov, Olga Barinova và Anton Konushin. f-brs: Xem xét lại quá trình tinh chỉnh lan truyền ngược cho phân đoạn tương tác. Biên bản báo cáo Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu, trang 8623-8632, 2020. 3
- [33] Rohan Taori, Ishaaq Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang và Tatsunori B. Hashimoto. Stanford alpaca: Một mô hình llama làm theo hướng dẫn. https://github.com/tatsu-lab/stanford_alpaca, 2023. 4
- [34] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poultton, Viktor Kerkez và Robert Stojnic. Galactica: Một mô hình ngôn ngữ lớn cho khoa học. bản in trước arXiv arXiv:2211.09085, 2022. 4
- [35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timo-thée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Các mô hình ngôn ngữ nền tảng mở và hiệu quả. bản in trước arXiv arXiv:2302.13971, 2023. 3, 5
- [36] Oriol Vinyals, Alexander Toshev, Samy Bengio và Dumitru Erhan. Hiển thị và kể: Một trình tạo chú thích hình ảnh thần kinh. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 3156-3164, 2015. 3
- [37] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: một khuôn khổ trắc địa tương tác sâu để phân đoạn hình ảnh y tế. Giao dịch IEEE về phân tích mẫu và trí tuệ máy móc, 41(7):1559–1572, 2018. 3
- [38] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zichen Liu, Ce Liu và Lijuan Wang. Git: Một công cụ chuyển đổi hình ảnh thành văn bản mang tính tổng quát cho tầm nhìn và ngôn ngữ. bản in trước arXiv arXiv:2205.14100, 2022. 2, 3
- [39] Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia và Linlin Li. Chú thích hình ảnh có thể kiểm soát thông qua lời nhắc. bản in trước arXiv arXiv:2212.01803, 2022. 3
- [40] Teng Wang, Yixiao Ge, Feng Zheng, Ran Cheng, Ying Shan, Xiaohu Qie và Ping Luo. Tăng tốc quá trình đào tạo trước ngôn ngữ thị giác với mô hình ngôn ngữ miễn phí. Bản in trước arXiv arXiv:2303.14038, 2023. 2

- [41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 5
- [42] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 5
- [43] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 2, 3
- [44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 3
- [45] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381, 2016. 3
- [46] Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, and Shuai Ma. Control image captioning spatially and temporally. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2014–2025, 2021. 3
- [47] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202, 2017. 2, 3
- [48] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6241–6250, 2019. 3
- [49] Zequn Zeng, Hao Zhang, Zhengjue Wang, Ruiying Lu, Dongsheng Wang, and Bo Chen. Conzic: Controllable zero-shot image captioning by sampling-based polishing. *arXiv preprint arXiv:2303.02437*, 2023. 3
- [50] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3, 5
- [51] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12984–12992, 2020. 3
- [41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le và Denny Zhou. Chuỗi suy nghĩ thúc đẩy lý luận trong các mô hình ngôn ngữ lớn. Bản in trước arXiv arXiv :2201.11903, 2022. 5
- [42] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zechen Tang, và Nan Duan. Visual chatgpt: Nói chuyện, vẽ và chỉnh sửa bằng các mô hình nền tảng trực quan. Bản in trước arXiv arXiv:2303.04671, 2023. 5
- [43] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zichen Liu, Junsong Yuan và Lijuan Wang. Grit: Một công cụ biến đổi vùng thành văn bản tổng quát để hiểu đối tượng. bản in trước arXiv arXiv:2212.00280, 2022. 2, 3
- [44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel và Yoshua Bengio. Hiển thị, tham dự và kể: Tạo chú thích hình ảnh thần kinh với sự chú ý trực quan. Trong Hội nghị quốc tế về máy học, trang 2048–2057. PMLR, 2015. 3
- [45] Ning Xu, Brian Price, Scott Cohen, Jimei Yang và Thomas S Huang. Lựa chọn đối tượng tương tác sâu. Biên bản hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 373–381, 2016. 3
- [46] Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan và Shuai Ma. Kiểm soát chú thích hình ảnh theo không gian và thời gian. Trong Biên bản báo cáo của Hội nghị thường niên lần thứ 59 của Hiệp hội Ngôn ngữ học tính toán và Hội nghị chung quốc tế lần thứ 11 về Xử lý ngôn ngữ tự nhiên (Tập 1: Bài báo dài), trang 2014–2025, 2021. 3
- [47] Linjie Yang, Kevin Tang, Jianchao Yang và Li-Jia Li. Chú thích dày đặc với suy luận chung và ngữ cảnh trực quan. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 2193–2202, 2017. 2, 3
- [48] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang và Jing Shao. Bối cảnh và thuộc tính dựa trên chú thích dày đặc. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu, trang 6241–6250, 2019. 3
- [49] Zequn Zeng, Hao Zhang, Zhengjue Wang, Ruiying Lu, Dongsheng Wang và Bo Chen. Conzic: Chú thích hình ảnh zero-shot có thể kiểm soát bằng cách đánh bóng dựa trên mẫu. Bản in trước arXiv arXiv:2303.02437, 2023. 3
- [50] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Mở các mô hình ngôn ngữ biến đổi được đào tạo trước. Bản in trước arXiv arXiv:2205.01068, 2022. 3, 5
- [51] Wentian Zhao, Xinxiao Wu, và Xiaoxun Zhang. Memcap: Ghi nhớ kiến thức về phong cách để chú thích hình ảnh. Trong Biên bản báo cáo Hội nghị AAAI về Tri tuệ nhân tạo, tập 34, trang 12984–12992, 2020. 3