

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li Dongxu Li Silvio Savarese Steven Hoi
Salesforce Research

<https://github.com/salesforce/LAVIS/tree/main/projects/blip2>

Abstract

The cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models. This paper proposes BLIP-2, a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. BLIP-2 bridges the modality gap with a lightweight Querying Transformer, which is pre-trained in two stages. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen language model. BLIP-2 achieves state-of-the-art performance on various vision-language tasks, despite having significantly fewer trainable parameters than existing methods. For example, our model outperforms Flamingo80B by 8.7% on zero-shot VQAv2 with 54x fewer trainable parameters. We also demonstrate the model's emerging capabilities of zero-shot image-to-text generation that can follow natural language instructions.

1. Introduction

Vision-language pre-training (VLP) research has witnessed a rapid advancement in the past few years, where pre-trained models with increasingly larger scale have been developed to continuously push the state-of-the-art on various downstream tasks (Radford et al., 2021; Li et al., 2021; 2022; Wang et al., 2022a; Alayrac et al., 2022; Wang et al., 2022b). However, most state-of-the-art vision-language models incur a high computation cost during pre-training, due to end-to-end training using large-scale models and datasets.

Vision-language research sits at the intersection between vision and language, therefore it is naturally expected that vision-language models can harvest from the readily-available unimodal models from the vision and natural language communities. In this paper, we propose a *generic* and

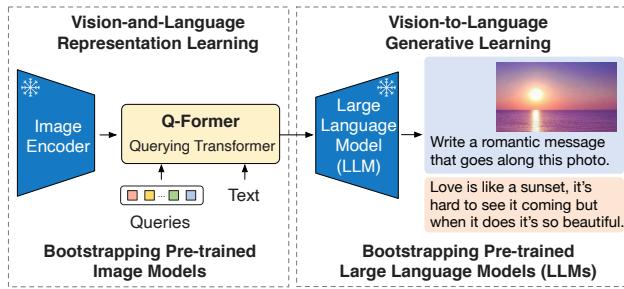


Figure 1. Overview of BLIP-2’s framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen language model. BLIP-2 achieves state-of-the-art performance on various vision-language tasks, despite having significantly fewer trainable parameters than existing methods. For example, our model outperforms Flamingo80B by 8.7% on zero-shot VQAv2 with 54x fewer trainable parameters. We also demonstrate the model’s emerging capabilities of zero-shot image-to-text generation that can follow natural language instructions.

compute-efficient VLP method by bootstrapping from off-the-shelf pre-trained vision models and language models. Pre-trained vision models offer high-quality visual representation. Pre-trained language models, in particular *large language models* (LLMs), offer strong language generation and zero-shot transfer abilities. To reduce computation cost and counteract the issue of catastrophic forgetting, the unimodal pre-trained models remain frozen during the pre-training.

In order to leverage pre-trained unimodal models for VLP, it is key to facilitate cross-modal alignment. However, since LLMs have not seen images during their unimodal pre-training, freezing them makes vision-language alignment in particular challenging. In this regard, existing methods (*e.g.* Frozen (Tsimpoukelli et al., 2021), Flamingo (Alayrac et al., 2022)) resort to an image-to-text generation loss, which we show is insufficient to bridge the modality gap.

To achieve effective vision-language alignment with frozen unimodal models, we propose a Querying Transformer (Q-Former) pre-trained with a new two-stage pre-training strategy. As shown in Figure 1, Q-Former is a lightweight transformer which employs a set of learnable query vectors to extract visual features from the frozen image encoder. It acts as an information bottleneck between the frozen image encoder and the frozen LLM, where it feeds the most useful

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa hình ảnh đóng băng và Mô hình ngôn ngữ lớn

Junnan Li Dongxu Li Silvio Savarese Steven Hoi
Nghiên cứu Salesforce

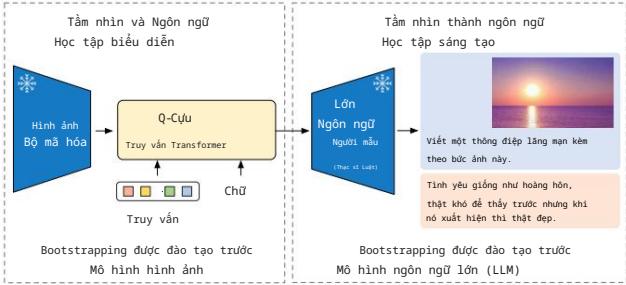
<https://github.com/salesforce/LAVIS/tree/main/projects/blip2>

Tóm tắt

Chi phí đào tạo trước về thị giác và ngôn ngữ đã trở nên ngày càng trả nêu cấm đoán do phải đào tạo toàn diện các mô hình quy mô lớn. Bài báo này đề xuất BLIP-2, một chiến lược tiền đào tạo chung và hiệu quả giúp khởi động ngôn ngữ thị giác đào tạo trước từ mô hình đóng băng có sẵn được đào tạo trước bộ mã hóa hình ảnh và các mô hình ngôn ngữ lớn đóng băng. BLIP-2 thu hẹp khoảng cách mô hình với Bộ chuyển đổi truy vấn nhẹ, được đào tạo trước trong hai giai đoạn. Giai đoạn đầu tiên khởi động việc học biểu diễn ngôn ngữ thị giác từ một bộ mã hóa hình ảnh đóng băng. Giai đoạn thứ hai bootstraps tầm nhìn-thành-nhân ngôn ngữ học tập tạo ra từ một mô hình ngôn ngữ đóng băng. BLIP-2 đạt được hiệu suất tiên tiến trên nhiều nhiệm vụ ngôn ngữ thị giác khác nhau, mặc dù có ít hơn đáng kể các thông số có thể đào tạo được hơn các phương pháp hiện có. Đối với Ví dụ, mô hình của chúng tôi vượt trội hơn Flamingo80B 8,7% trên VQAv2 không bắn với ít hơn 54 lần các thông số có thể đào tạo. Chúng tôi cũng chứng minh mô hình khả năng mới nổi của hình ảnh không cần chụp thành văn bản thế hệ có thể tuân theo các hướng dẫn bằng ngôn ngữ tự nhiên.

1. Giới thiệu

Nghiên cứu tiền đào tạo ngôn ngữ thị giác (VLP) đã chứng kiến một sự tiến bộ nhanh chóng trong vài năm qua, nơi được đào tạo trước các mô hình với quy mô ngày càng lớn hơn đã được phát triển liên tục đẩy công nghệ tiên tiến nhất trong nhiều nhiệm vụ hạ nguồn khác nhau (Radford và cộng sự, 2021; Li và cộng sự, 2021; 2022; Wang và cộng sự, 2022a; Alayrac và cộng sự, 2022; Wang và cộng sự, 2022b). Tuy nhiên, hầu hết các mô hình ngôn ngữ thị giác hiện đại đều phải chịu chi phí tính toán cao trong quá trình đào tạo trước, do đó nó được đào tạo toàn diện bằng cách sử dụng các mô hình và tập dữ liệu quy mô lớn. Nghiên cứu ngôn ngữ thị giác nằm ở giao điểm giữa tầm nhìn và ngôn ngữ, do đó nó được mong đợi một cách tự nhiên rằng các mô hình ngôn ngữ thị giác có thể thu thập từ các mô hình đơn thức có sẵn từ cộng đồng thị giác và ngôn ngữ tự nhiên. Trong bài báo này, chúng tôi đề xuất một mô hình chung và



Hình 1. Tổng quan về khuôn khổ BLIP-2. Chúng tôi đào tạo trước một Bộ chuyển đổi truy vấn nhẹ theo chiến lược hai giai đoạn để thu hẹp khoảng cách phương thức. Giai đoạn đầu tiên khởi động việc học biểu diễn ngôn ngữ thị giác từ bộ mã hóa hình ảnh đóng băng. Giai đoạn thứ hai khởi động quá trình học tập tạo ra từ tầm nhìn đến ngôn ngữ từ LLM đóng băng, cho phép tạo hình ảnh thành văn bản theo hướng dẫn không cần chụp (xem Hình 4 để biết thêm ví dụ).

phương pháp VLP hiệu quả về mặt tính toán bằng cách khởi động từ các mô hình ngôn ngữ và mô hình thị giác được đào tạo sẵn. Các mô hình thị giác được đào tạo trước cung cấp khả năng biểu diễn trực quan chất lượng cao. Các mô hình ngôn ngữ được đào tạo trước, đặc biệt là các mô hình ngôn ngữ lớn (LLM), cung cấp khả năng tạo ngôn ngữ mạnh mẽ và khả năng chuyển giao không cần bắt đầu. Để giảm chi phí tính toán và chống lại vấn đề quên lảng thâm khắc, đơn giản các mô hình được đào tạo trước vẫn bị đóng băng trong quá trình đào tạo trước.

Để tận dụng các mô hình đơn thức được đào tạo trước cho VLP, điều quan trọng là tạo điều kiện cho sự liên kết đa phương thức. Tuy nhiên, vì Các LLM không nhìn thấy hình ảnh trong quá trình đào tạo trước đơn phương thức của họ, việc đóng băng chúng làm cho sự liên kết giữa thị giác và ngôn ngữ đặc biệt là thách thức. Về vấn đề này, các phương pháp hiện có (ví dụ Frozen (Tsimpoukelli et al., 2021), Flamingo (Alayrac et al., 2022)) sử dụng phương pháp mắng mít từ hình ảnh sang văn bản, mà chúng tôi cho thấy là không đủ để thu hẹp khoảng cách phương thức.

Để đạt được sự liên kết ngôn ngữ-thị giác hiệu quả với đóng băng mô hình đơn thức, chúng tôi đề xuất một Bộ chuyển đổi truy vấn (Q-Former) được đào tạo trước với chiến lược đào tạo trước hai giai đoạn mới. Như thể hiện trong Hình 1, Q-Former là một bộ chuyển đổi nhẹ sử dụng một tập hợp các vectơ truy vấn có thể học được để trích xuất các đặc điểm trực quan từ bộ mã hóa hình ảnh đóng băng. Nó hoạt động như một nút thắt thông tin giữa hình ảnh đóng băng bộ mã hóa và LLM đóng băng, nơi nó cung cấp thông tin hữu ích nhất.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

visual feature for the LLM to output the desired text. In the first pre-training stage, we perform vision-language representation learning which enforces the Q-Former to learn visual representation most relevant to the text. In the second pre-training stage, we perform vision-to-language generative learning by connecting the output of the Q-Former to a frozen LLM, and trains the Q-Former such that its output visual representation can be interpreted by the LLM.

We name our VLP framework as BLIP-2: Bootstrapping Language-Image Pre-training with frozen unimodal models. The key advantages of BLIP-2 include:

- BLIP-2 effectively leverages both frozen pre-trained image models and language models. We bridge the modality gap using a Q-Former pre-trained in two-stages: representation learning stage and generative learning stage. BLIP-2 achieves state-of-the-art performance on various vision-language tasks including visual question answering, image captioning, and image-text retrieval.
- Powered by LLMs (e.g. OPT (Zhang et al., 2022), FlanT5 (Chung et al., 2022)), BLIP-2 can be prompted to perform zero-shot image-to-text generation that follows natural language instructions, which enables emerging capabilities such as visual knowledge reasoning, visual conversation, etc. (see Figure 4 for examples).
- Due to the use of frozen unimodal models and a lightweight Q-Former, BLIP-2 is more compute-efficient than existing state-of-the-arts. For example, BLIP-2 outperforms Flamingo (Alayrac et al., 2022) by 8.7% on zero-shot VQAv2, while using 54× fewer trainable parameters. Furthermore, our results show that BLIP-2 is a generic method that can harvest more advanced unimodal models for better VLP performance.

2. Related Work

2.1. End-to-end Vision-Language Pre-training

Vision-language pre-training aims to learn multimodal foundation models with improved performance on various vision-and-language tasks. Depending on the downstream task, different model architectures have been proposed, including the dual-encoder architecture (Radford et al., 2021; Jia et al., 2021), the fusion-encoder architecture (Tan & Bansal, 2019; Li et al., 2021), the encoder-decoder architecture (Cho et al., 2021; Wang et al., 2021b; Chen et al., 2022b), and more recently, the unified transformer architecture (Li et al., 2022; Wang et al., 2022b). Various pre-training objectives have also been proposed over the years, and have progressively converged to a few time-tested ones: image-text contrastive learning (Radford et al., 2021; Yao et al., 2022; Li et al., 2021; 2022), image-text matching (Li et al., 2021; 2022; Wang et al., 2021a), and (masked) language modeling (Li et al., 2021; 2022; Yu et al., 2022; Wang et al., 2022b).

Most VLP methods perform end-to-end pre-training using large-scale image-text pair datasets. As the model size keeps increasing, the pre-training can incur an extremely high computation cost. Moreover, it is inflexible for end-to-end pre-trained models to leverage readily-available unimodal pre-trained models, such as LLMs (Brown et al., 2020; Zhang et al., 2022; Chung et al., 2022).

2.2. Modular Vision-Language Pre-training

More similar to us are methods that leverage off-the-shelf pre-trained models and keep them frozen during VLP. Some methods freeze the image encoder, including the early work which adopts a frozen object detector to extract visual features (Chen et al., 2020; Li et al., 2020; Zhang et al., 2021), and the recent LiT (Zhai et al., 2022) which uses a frozen pre-trained image encoder for CLIP (Radford et al., 2021) pre-training. Some methods freeze the language model to use the knowledge from LLMs for vision-to-language generation tasks (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Chen et al., 2022a; Mañas et al., 2023; Tiong et al., 2022; Guo et al., 2022). The key challenge in using a frozen LLM is to align visual features to the text space. To achieve this, Frozen (Tsimpoukelli et al., 2021) finetunes an image encoder whose outputs are directly used as soft prompts for the LLM. Flamingo (Alayrac et al., 2022) inserts new cross-attention layers into the LLM to inject visual features, and pre-trains the new layers on billions of image-text pairs. Both methods adopt the language modeling loss, where the language model generates texts conditioned on the image.

Different from existing methods, BLIP-2 can effectively and efficiently leverage both frozen image encoders and frozen LLMs for various vision-language tasks, achieving stronger performance at a lower computation cost.

3. Method

We propose BLIP-2, a new vision-language pre-training method that bootstraps from frozen pre-trained unimodal models. In order to bridge the modality gap, we propose a Querying Transformer (Q-Former) pre-trained in two stages: (1) vision-language representation learning stage with a frozen image encoder and (2) vision-to-language generative learning stage with a frozen LLM. This section first introduces the model architecture of Q-Former, and then delineates the two-stage pre-training procedures.

3.1. Model Architecture

We propose Q-Former as the trainable module to bridge the gap between a frozen image encoder and a frozen LLM. It extracts a fixed number of output features from the image encoder, independent of input image resolution. As shown in Figure 2, Q-Former consists of two transformer submodules that share the same self-attention layers: (1) an image transformer that interacts with the frozen image encoder

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn

tính năng trực quan cho LLM để xuất ra văn bản mong muốn. Trong giai đoạn tiền đào tạo đầu tiên, chúng tôi thực hiện học biểu diễn ngôn ngữ thị giác, điều này buộc Q-Former phải học biểu diễn trực quan có liên quan nhất đến văn bản. Trong giai đoạn tiền đào tạo thứ hai, chúng tôi thực hiện học tạo thị giác sang ngôn ngữ bằng cách kết nối đầu ra của Q-Former với LLM đóng lạnh và đào tạo Q-Former sao cho biểu diễn trực quan đầu ra của nó có thể được LLM diễn giải.

Chúng tôi đặt tên cho khuôn khổ VLP của mình là BLIP-2: Khởi động đào tạo trước ngôn ngữ-hình ảnh với các mô hình đơn thức đóng băng. Những lợi thế chính của BLIP-2 bao gồm:

- BLIP-2 tận dụng hiệu quả cả mô hình hình ảnh được đào tạo trước đóng lạnh và mô hình ngôn ngữ. Chúng tôi thu hẹp khoảng cách phương thức bằng cách sử dụng Q-Former được đào tạo trước trong hai giai đoạn: giai đoạn học biểu diễn và giai đoạn học tạo sinh. BLIP-2 đạt hiệu suất tiên tiến nhất trong nhiều tác vụ ngôn ngữ thị giác bao gồm trả lời câu hỏi trực quan, chú thích hình ảnh và truy xuất văn bản hình ảnh.
- Được hỗ trợ bởi LLM (ví dụ OPT (Zhang và cộng sự, 2022), FlanT5 (Chung và cộng sự, 2022)), BLIP-2 có thể được nhắc thực hiện tạo hình ảnh thành văn bản không cần chụp theo hướng dẫn ngôn ngữ tự nhiên, cho phép các khả năng mới nổi như suy luận kiến thức trực quan, hội thoại trực quan, v.v. (xem Hình 4 để biết ví dụ).
- Do sử dụng các mô hình đơn thức đóng lạnh và Q-Former nhẹ, BLIP-2 hiệu quả hơn về mặt tính toán so với các công nghệ tiên tiến hiện có. Ví dụ, BLIP-2 vượt trội hơn Flamingo (Alayrac và cộng sự, 2022) 8.7% về VQAv2 không cần bắn, trong khi sử dụng ít hơn 54 lần các tham số có thể đào tạo. Hơn nữa, kết quả của chúng tôi cho thấy BLIP-2 là một phương pháp chung có thể thu thập các mô hình đơn thức tiên tiến hơn để có hiệu suất VLP tốt hơn.

2. Công trình liên quan

- 2.1. Tiền huấn luyện ngôn ngữ thị giác đầu cuối Tiền huấn luyện ngôn ngữ thị giác nhằm mục đích học các mô hình nền tảng đa phương thức với hiệu suất được cải thiện trên nhiều tác vụ thị giác và ngôn ngữ. Tùy thuộc vào tác vụ hạ nguồn, các kiến trúc mô hình khác nhau đã được đề xuất, bao gồm kiến trúc bộ mã hóa kép (Radford và cộng sự, 2021; Jia và cộng sự, 2021), kiến trúc bộ mã hóa hợp nhất (Tan & Bansal, 2019; Li và cộng sự, 2021), kiến trúc bộ mã hóa-giải mã (Cho và cộng sự, 2021; Wang và cộng sự, 2021b; Chen và cộng sự, 2022b) và gần đây hơn là kiến trúc bộ biến áp thống nhất (Li và cộng sự, 2022; Wang và cộng sự, 2022b). Nhiều mục tiêu tiền đào tạo khác nhau cũng đã được đề xuất trong những năm qua và dần dần hội tụ thành một số mục tiêu đã được kiểm nghiệm theo thời gian: học tương phản hình ảnh-văn bản (Radford và cộng sự, 2021; Yao và cộng sự, 2022; Li và cộng sự, 2021; 2022), so khớp hình ảnh-văn bản (Li và cộng sự, 2021; 2022; Wang và cộng sự, 2021a) và mô hình hóa ngôn ngữ (có che dấu) (Li và cộng sự, 2021; 2022; Yu và cộng sự, 2022; Wang và cộng sự, 2022b).

Hầu hết các phương pháp VLP đều thực hiện tiền đào tạo đầu cuối bằng cách sử dụng các tập dữ liệu cặp hình ảnh-văn bản quy mô lớn. Khi kích thước mô hình tiếp tục tăng, tiền đào tạo có thể phát sinh chi phí tính toán cực kỳ cao. Hơn nữa, các mô hình được đào tạo đầu cuối không linh hoạt để tận dụng các mô hình được đào tạo đơn phương thúc có sẵn, chẳng hạn như LLM (Brown và cộng sự, 2020; Zhang và cộng sự, 2022; Chung và cộng sự, 2022).

2.2. Đào tạo trước ngôn ngữ tầm nhìn mô-đun

Tương tự như chúng tôi là các phương pháp tận dụng các mô hình được đào tạo trước có sẵn và giữ chúng ở trạng thái đóng băng trong VLP. Một số phương pháp đóng băng bộ mã hóa hình ảnh, bao gồm công trình ban đầu sử dụng bộ phát hiện vật thể đóng băng để trích xuất các đặc điểm trực quan (Chen và cộng sự, 2020; Li và cộng sự, 2020; Zhang và cộng sự, 2021) và LiT gần đây (Zhai và cộng sự, 2022) sử dụng bộ mã hóa hình ảnh được đào tạo trước đóng băng cho quá trình đào tạo trước CLIP (Radford và cộng sự, 2021). Một số phương pháp đóng băng mô hình ngôn ngữ để sử dụng kiến thức từ LLM cho các tác vụ tạo thị giác thành ngôn ngữ (Tsimpoukelli và cộng sự, 2021; Alayrac và cộng sự, 2022; Chen và cộng sự, 2022a; Mañas và cộng sự, 2023; Tiong và cộng sự, 2022; Guo và cộng sự, 2022). Thách thức chính khi sử dụng LLM đóng lạnh là cần chỉnh các đặc điểm trực quan với không gian văn bản. Để đạt được điều này, Frozen (Tsimpoukelli và cộng sự, 2021) tinh chỉnh một bộ mã hóa hình ảnh có đầu ra được sử dụng trực tiếp làm lời nhắc mềm cho LLM. Flamingo (Alayrac và cộng sự, 2022) chèn các lớp chú ý chéo mới vào LLM để đưa các đặc điểm trực quan và đào tạo trước các lớp mới trên hàng tỷ cặp hình ảnh-văn bản. Cả hai phương pháp đều sử dụng mô hình ngôn ngữ để tạo ra văn bản dựa trên hình ảnh.

Khác với các phương pháp hiện có, BLIP-2 có thể tận dụng hiệu quả cả bộ mã hóa hình ảnh đóng lạnh và LLM đóng lạnh cho nhiều tác vụ ngôn ngữ thị giác khác nhau, đạt hiệu suất mạnh hơn với chi phí tính toán thấp hơn.

3. Phương pháp

Chúng tôi đề xuất BLIP-2, một phương pháp tiền đào tạo ngôn ngữ thị giác mới khởi động từ các mô hình đơn thức được đào tạo trước đóng lạnh. Để thu hẹp khoảng cách phương thức, chúng tôi đề xuất một Querying Transformer (Q-Former) được đào tạo trước trong hai giai đoạn: (1) giai đoạn học biểu diễn ngôn ngữ thị giác với bộ mã hóa hình ảnh đóng lạnh và (2) giai đoạn học tạo thị giác thành ngôn ngữ với LLM đóng lạnh. Phần này trước tiên giới thiệu kiến trúc mô hình của Q-Former, sau đó mô tả các quy trình tiền đào tạo hai giai đoạn.

3.1. Kiến trúc mô hình

Chúng tôi đề xuất Q-Former là mô-đun có thể đào tạo để thu hẹp khoảng cách giữa bộ mã hóa hình ảnh đóng lạnh và LLM đóng lạnh. Nó trích xuất một số lượng cố định các tính năng đầu ra từ bộ mã hóa hình ảnh, độc lập với độ phân giải hình ảnh đầu vào. Như thể hiện trong Hình 2, Q-Former bao gồm hai mô-đun con biến áp chia sẻ cùng một lớp tự chú ý: (1) một biến áp hình ảnh tương tác với bộ mã hóa hình ảnh đóng lạnh

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

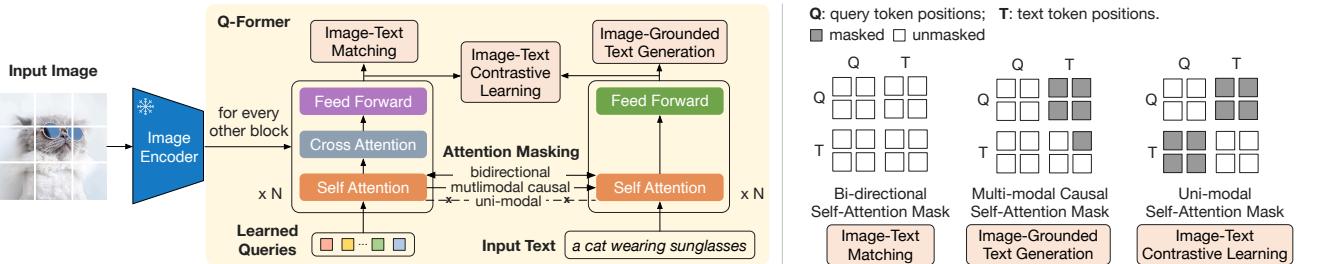


Figure 2. (Left) Model architecture of Q-Former and BLIP-2’s first-stage vision-language representation learning objectives. We jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text. (Right) The self-attention masking strategy for each objective to control query-text interaction.

for visual feature extraction, (2) a text transformer that can function as both a text encoder and a text decoder. We create a set number of learnable query embeddings as input to the image transformer. The queries interact with each other through self-attention layers, and interact with frozen image features through cross-attention layers (inserted every other transformer block). The queries can additionally interact with the text through the same self-attention layers. Depending on the pre-training task, we apply different self-attention masks to control query-text interaction. We initialize Q-Former with the pre-trained weights of BERT_{base} (Devlin et al., 2019), whereas the cross-attention layers are randomly initialized. In total, Q-Former contains 188M parameters. Note that the queries are considered as model parameters.

In our experiments, we use 32 queries where each query has a dimension of 768 (same as the hidden dimension of the Q-Former). We use Z to denote the output query representation. The size of Z (32×768) is much smaller than the size of frozen image features (e.g. 257×1024 for ViT-L/14). This bottleneck architecture works together with our pre-training objectives into forcing the queries to extract visual information that is most relevant to the text.

3.2. Bootstrap Vision-Language Representation Learning from a Frozen Image Encoder

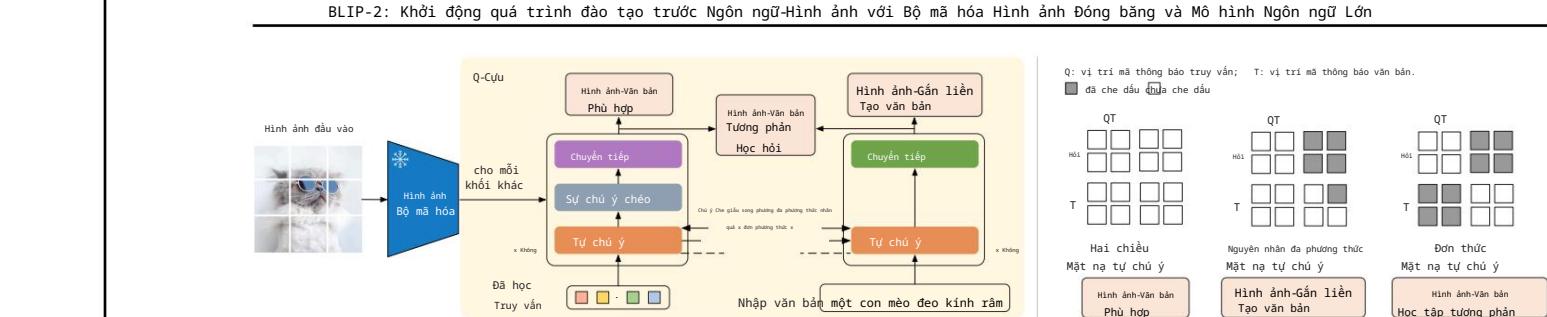
In the representation learning stage, we connect Q-Former to a frozen image encoder and perform pre-training using image-text pairs. We aim to train the Q-Former such that the queries can learn to extract visual representation that is most informative of the text. Inspired by BLIP (Li et al., 2022), we jointly optimize three pre-training objectives that share the same input format and model parameters. Each objective employs a different attention masking strategy between queries and text to control their interaction (see Figure 2).

Image-Text Matching (ITM) learns to align image representation and text representation such that their mutual information is maximized. It achieves so by contrasting the image-text similarity of a positive pair against those of negative pairs. We align the output query representation Z from the image transformer with the text representation

t from the text transformer, where t is the output embedding of the [CLS] token. Since Z contains multiple output embeddings (one from each query), we first compute the pairwise similarity between each query output and t , and then select the highest one as the image-text similarity. To avoid information leak, we employ a unimodal self-attention mask, where the queries and text are not allowed to see each other. Due to the use of a frozen image encoder, we can fit more samples per GPU compared to end-to-end methods. Therefore, we use in-batch negatives instead of the momentum queue in BLIP.

Image-grounded Text Generation (ITG) loss trains the Q-Former to generate texts, given input images as the condition. Since the architecture of Q-Former does not allow direct interactions between the frozen image encoder and the text tokens, the information required for generating the text must be first extracted by the queries, and then passed to the text tokens via self-attention layers. Therefore, the queries are forced to extract visual features that capture all the information about the text. We employ a multimodal causal self-attention mask to control query-text interaction, similar to the one used in UniLM (Dong et al., 2019). The queries can attend to each other but not the text tokens. Each text token can attend to all queries and its previous text tokens. We also replace the [CLS] token with a new [DEC] token as the first text token to signal the decoding task.

Image-Text Contrastive Learning (ITC) learns to align image representation and text representation such that their mutual information is maximized. It achieves so by contrasting the image-text similarity of a positive pair against those of negative pairs. We align the output query representation Z from the image transformer with the text representation



Hình 2. (Bên trái) Kiến trúc mô hình của Q-Former và các mục tiêu học ngôn ngữ thị giác giai đoạn đầu của BLIP-2. Chúng tôi cùng nhau tối ưu hóa ba mục tiêu thực thi các truy vấn (một tập hợp các nhúng có thể học được) để trích xuất biểu diễn trực quan có liên quan nhất đến văn bản. (Bên phải) Chiến lược che giấu sự chú ý của bản thân cho mỗi mục tiêu để kiểm soát tương tác truy vấn-văn bản.

để trích xuất đặc điểm trực quan, (2) một bộ chuyển đổi văn bản có thể hoạt động như cá bộ mã hóa văn bản và bộ giải mã văn bản. Chúng tôi tạo một số lượng nhúng truy vấn có thể học được làm đầu vào cho bộ chuyển đổi hình ảnh. Các truy vấn tương tác với nhau thông qua các lớp tự chú ý và tương tác với các đặc điểm hình ảnh bị đóng băng thông qua các lớp chú ý chéo (được chèn vào mọi khối bộ chuyển đổi khác). Các truy vấn cũng có thể tương tác với văn bản thông qua cùng các lớp tự chú ý. Tùy thuộc vào tác vụ đào tạo trước, chúng tôi áp dụng các mặt nạ tự chú ý khác nhau để kiểm soát tương tác truy vấn-văn bản. Chúng tôi khởi tạo Q-Former bằng cách trọng số được đào tạo trước của BERTbase (Devlin và cộng sự, 2019), trong khi các lớp chú ý chéo được khởi tạo ngẫu nhiên. Tổng cộng, Q-Former chứa 188 triệu tham số.

Lưu ý rằng các truy vấn được coi là tham số mô hình.

Trong các thí nghiệm của mình, chúng tôi sử dụng 32 truy vấn, trong đó mỗi truy vấn có kích thước là 768 (giống như kích thước ẩn của Q-Former). Chúng tôi sử dụng Z để biểu thị biểu diễn truy vấn đầu ra. Kích thước của Z (32×768) nhỏ hơn nhiều so với kích thước của các đặc điểm hình ảnh bị đóng băng (ví dụ: 257×1024 đối với ViT-L/14). Kiến trúc nút thất này hoạt động cùng với các mục tiêu đào tạo trước của chúng tôi để buộc các truy vấn trích xuất thông tin trực quan có liên quan nhất đến văn bản.

3.2. Học biểu diễn ngôn ngữ tầm nhìn Bootstrap từ bộ mã hóa hình ảnh đóng băng

Trong giai đoạn học biểu diễn, chúng tôi kết nối Q-Former với một bộ mã hóa hình ảnh đóng băng và thực hiện tiền đào tạo bằng cách sử dụng các cặp hình ảnh-văn bản. Chúng tôi hướng đến việc đào tạo Q-Former sao cho các truy vấn có thể học cách trích xuất biểu diễn trực quan cung cấp nhiều thông tin nhất cho văn bản. Lấy cảm hứng từ BLIP (Li và cộng sự, 2022), chúng tôi cùng nhau tối ưu hóa ba mục tiêu tiền đào tạo có chung định dạng đầu vào và các tham số mô hình. Mỗi mục tiêu sử dụng một chiến lược che giấu sự chú ý khác nhau giữa các truy vấn và văn bản để kiểm soát tương tác của chúng (xem

Hình 2). Nhúng truy vấn đầu ra Z do đó nắm bắt thông tin đa phương thức. Chúng tôi áp dụng chiến lược khai thác tiêu cực cứng từ Li et al. (2021; 2022) để tạo ra các cặp tiêu cực có thông tin.

t từ bộ chuyển đổi văn bản, trong đó t là nhúng đầu ra của mã thông báo [CLS]. Vì Z chứa nhiều nhúng đầu ra (một từ mỗi truy vấn), trước tiên chúng tôi tính toán độ tương đồng từng cặp giữa mỗi đầu ra truy vấn và t , sau đó chọn giá trị cao nhất làm độ tương đồng hình ảnh-văn bản. Để tránh rò rỉ thông tin, chúng tôi sử dụng mặt nạ tự chú ý đơn thức, trong đó các truy vấn và văn bản không được phép nhìn thấy nhau. Do sử dụng bộ mã hóa hình ảnh bị đóng băng, chúng tôi có thể phù hợp với nhiều mẫu hơn trên mỗi GPU so với các phương pháp đầu cuối. Do đó, chúng tôi sử dụng các âm bản trong lô thay vì hàng đợi động lượng trong BLIP.

Mắt mát Tạo văn bản dựa trên hình ảnh (ITG) đào tạo Q-Former để tạo văn bản, với hình ảnh đầu vào làm điều kiện. Vì kiến trúc của Q-Former không cho phép tương tác trực tiếp giữa bộ mã hóa hình ảnh đóng băng và mã thông báo văn bản, nên thông tin cần thiết để tạo văn bản trước tiên phải được trích xuất bởi các truy vấn, sau đó được truyền đến các mã thông báo văn bản thông qua các lớp tự chú ý. Do đó, các truy vấn buộc phải trích xuất các tính năng trực quan nắm bắt tất cả thông tin về văn bản. Chúng tôi sử dụng mặt nạ tự chú ý nhân quả đa phương thức để kiểm soát tương tác truy vấn-văn bản, tương tự như mặt nạ được sử dụng trong UniLM (Dong và cộng sự, 2019). Các truy vấn có thể chú ý đến nhau nhưng không phải các mã thông báo văn bản. Mỗi mã thông báo văn bản có thể chú ý đến tất cả các truy vấn và các mã thông báo văn bản trước đó của nó. Chúng tôi cũng thay thế mã thông báo [CLS] bằng mã thông báo [DEC] mới làm mã thông báo văn bản đầu tiên để báo hiệu tác vụ giải mã.

Để học hình ảnh-văn bản (ITM) nhằm mục đích tìm hiểu sự cân chỉnh chi tiết giữa biểu diễn hình ảnh và văn bản. Đây là nhiệm vụ phân loại nhị phân trong đó mô hình được yêu cầu dự đoán xem cặp hình ảnh-văn bản là tích cực (khớp) hay tiêu cực (không khớp). Chúng tôi sử dụng mặt nạ tự chú ý hai chiều trong đó tất cả các truy vấn và văn bản có thể chú ý đến nhau.

Nhúng truy vấn đầu ra Z do đó nắm bắt thông tin đa phương thức. Chúng tôi áp dụng chiến lược khai thác tiêu cực cứng (xem Hình 2) để có được logit và lấy trung bình logit trên tất cả các truy vấn làm điểm khớp đầu ra. Chúng tôi áp dụng chiến lược khai thác tiêu cực cứng từ Li et al. (2021; 2022) để tạo ra các cặp tiêu cực có thông tin.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

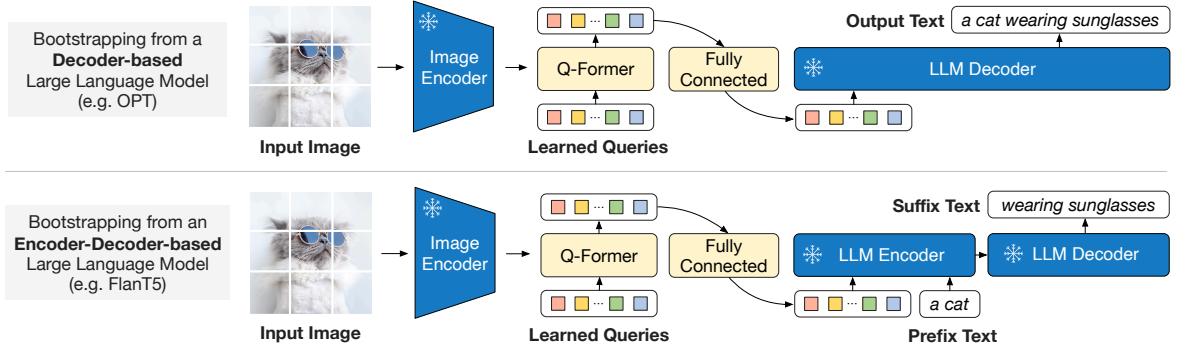


Figure 3. BLIP-2’s second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (Top) Bootstrapping a decoder-based LLM (e.g. OPT). (Bottom) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

3.3. Bootstrap Vision-to-Language Generative Learning from a Frozen LLM

In the generative pre-training stage, we connect Q-Former (with the frozen image encoder attached) to a frozen LLM to harvest the LLM’s generative language capability. As shown in Figure 3, we use a fully-connected (FC) layer to linearly project the output query embeddings Z into the same dimension as the text embedding of the LLM. The projected query embeddings are then prepended to the input text embeddings. They function as *soft visual prompts* that condition the LLM on visual representation extracted by the Q-Former. Since the Q-Former has been pre-trained to extract language-informative visual representation, it effectively functions as an information bottleneck that feeds the most useful information to the LLM while removing irrelevant visual information. This reduces the burden of the LLM to learn vision-language alignment, thus mitigating the catastrophic forgetting problem.

We experiment with two types of LLMs: decoder-based LLMs and encoder-decoder-based LLMs. For decoder-based LLMs, we pre-train with the language modeling loss, where the frozen LLM is tasked to generate the text conditioned on the visual representation from Q-Former. For encoder-decoder-based LLMs, we pre-train with the prefix language modeling loss, where we split a text into two parts. The prefix text is concatenated with the visual representation as input to the LLM’s encoder. The suffix text is used as the generation target for the LLM’s decoder.

3.4. Model Pre-training

Pre-training data. We use the same pre-training dataset as BLIP with 129M images in total, including COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), SBU (Ordonez et al., 2011), and 115M images from the LAION400M dataset (Schuhmann et al., 2021). We adopt the CapFilt method (Li et al., 2022) to create synthetic captions for the web images. Specifically, we generate 10

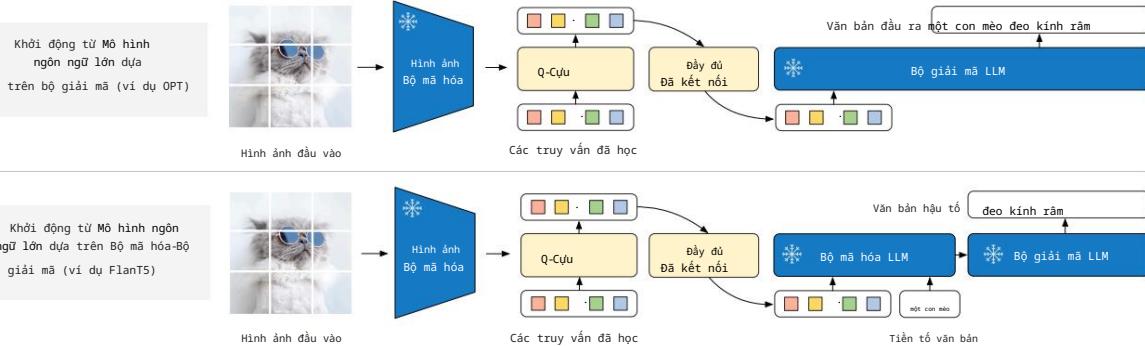
captions using the BLIP_{large} captioning model, and rank the synthetic captions along with the original web caption based on the image-text similarity produced by a CLIP ViT-L/14 model. We keep top-two captions per image as training data and randomly sample one at each pre-training step.

Pre-trained image encoder and LLM. For the frozen image encoder, we explore two state-of-the-art pre-trained vision transformer models: (1) ViT-L/14 from CLIP (Radford et al., 2021) and (2) ViT-g/14 from EVA-CLIP (Fang et al., 2022). We remove the last layer of the ViT and uses the second last layer’s output features, which leads to slightly better performance. For the frozen language model, we explore the unsupervised-trained OPT model family (Zhang et al., 2022) for decoder-based LLMs, and the instruction-trained FlanT5 model family (Chung et al., 2022) for encoder-decoder-based LLMs.

Pre-training settings. We pre-train for 250k steps in the first stage and 80k steps in the second stage. We use a batch size of 2320/1680 for ViT-L/ViT-g in the first stage and a batch size of 1920/1520 for OPT/FlanT5 in the second stage. During pre-training, we convert the frozen ViTs’ and LLMs’ parameters into FP16, except for FlanT5 where we use BFloat16. We found no performance degradation compared to using 32-bit models. Due to the use of frozen models, our pre-training is more computational friendly than existing large-scale VLP methods. For example, using a single 16-A100(40G) machine, our largest model with ViT-g and FlanT5-XXL requires less than 6 days for the first stage and less than 3 days for the second stage.

The same set of pre-training hyper-parameters are used for all models. We use the AdamW (Loshchilov & Hutter, 2017) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a weight decay of 0.05. We use a cosine learning rate decay with a peak learning rate of 1e-4 and a linear warmup of 2k steps. The minimum learning rate at the second stage is 5e-5. We use images of size 224×224, augmented with random resized cropping and horizontal flipping.

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn



Hình 3. Quá trình đào tạo trước tạo thị giác thành ngôn ngữ giải đoạn thứ hai của BLIP-2, khởi động từ các mô hình ngôn ngữ lớn đóng lạnh (LLM). (Trên) Khởi động LLM dựa trên bộ giải mã (ví dụ OPT). (Dưới) Khởi động LLM dựa trên bộ mã hóa-giải mã (ví dụ FlanT5). Lớp được kết nối dày dủ thích ứng từ chiều đầu ra của Q-Former sang chiều đầu vào của LLM đã chọn.

3.3. Bootstrap Vision-to-Language Học tập sáng tạo từ LLM đóng băng

Trong giai đoạn tiền đào tạo tạo sinh, chúng tôi kết nối Q-Former (có gắn bộ mã hóa hình ảnh đóng lạnh) với LLM đóng lạnh để thu thập khả năng ngôn ngữ tạo sinh của LLM.

Như thể hiện trong Hình 3, chúng tôi sử dụng một lớp được kết nối dày dủ (FC) để chiếu tuyển tính các nhúng truy vấn đầu ra Z vào cùng một chiều với nhúng văn bản của LLM. Các nhúng truy vấn được chiếu sau đó được thêm vào nhúng văn bản đầu vào. Chúng hoạt động như các lời nhắc trực quan mềm đặt điều kiện cho LLM về biểu diễn trực quan được trích xuất bởi Q-Former. Vì Q-Former đã được đào tạo trước để trích xuất biểu diễn trực quan cung cấp thông tin về ngôn ngữ, nên nó hoạt động hiệu quả như một nút thắt thông tin cung cấp thông tin hữu ích nhất cho LLM trong khi loại bỏ thông tin trực quan không liên quan. Điều này làm giảm gánh nặng của LLM trong việc học cách căn chỉnh ngôn ngữ-thị giác, do đó giảm thiểu vấn đề quên thăm khứ.

Chúng tôi thử nghiệm với hai loại LLM: LLM dựa trên bộ giải mã và LLM dựa trên bộ mã hóa-giải mã. Đối với LLM dựa trên bộ giải mã, chúng tôi đào tạo trước với mắt mát mô hình ngôn ngữ, trong đó LLM đóng lạnh được giao nhiệm vụ tạo văn bản được điều kiện hóa trên biểu diễn trực quan từ Q-Former. Đối với LLM dựa trên bộ mã hóa-giải mã, chúng tôi đào tạo trước với mắt mát mô hình ngôn ngữ tiền tố, trong đó chúng tôi chia văn bản thành hai phần. Văn bản tiền tố được nối với biểu diễn trực quan làm đầu vào cho bộ mã hóa LLM. Văn bản hậu tố được sử dụng làm mục tiêu tạo cho bộ giải mã LLM.

3.4. Mô hình tiền huấn luyện Dữ liệu tiền huấn luyện

liệu tiền huấn luyện. Chúng tôi sử dụng cùng một tập dữ liệu tiền huấn luyện như BLIP với tổng cộng 129 triệu hình ảnh, bao gồm COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), SBU (Ordonez et al., 2011) và 115 triệu hình ảnh từ tập dữ liệu LAION400M (Schuhmann et al., 2021). Chúng tôi áp dụng phương pháp CapFilt (Li et al., 2022) để tạo chủ đề tổng hợp cho hình ảnh trên web. Cụ thể, chúng tôi tạo 10

chú thích sử dụng mô hình chủ thích BLIPlarge và xếp hạng các chủ thích tổng hợp cùng với chú thích web gốc dựa trên độ tương đồng giữa hình ảnh và văn bản do mô hình CLIP ViT-L/14 tạo ra. Chúng tôi giữ hai chủ thích hàng đầu cho mỗi hình ảnh làm dữ liệu đào tạo và lấy mẫu ngẫu nhiên một chủ thích tại mỗi bước tiền đào tạo.

Bộ mã hóa hình ảnh được đào tạo trước và LLM. Đối với bộ mã hóa hình ảnh đóng băng, chúng tôi khám phá hai mô hình biến đổi thị giác được đào tạo trước tiên tiến: (1) ViT-L/14 từ CLIP (Radford và cộng sự, 2021) và (2) ViT-g/14 từ EVA-CLIP (Fang và cộng sự, 2022). Chúng tôi loại bỏ lớp cuối cùng của ViT và sử dụng các tính năng đầu ra của lớp thứ hai từ cuối, dẫn đến hiệu suất tốt hơn một chút. Đối với mô hình ngôn ngữ đóng băng, chúng tôi khám phá họ mô hình OPT được đào tạo không giám sát (Zhang và cộng sự, 2022) cho LLM dựa trên bộ giải mã và họ mô hình FlanT5 được đào tạo theo lệnh (Chung và cộng sự, 2022) cho LLM dựa trên bộ mã hóa-giải mã.

Cài đặt trước khi đào tạo. Chúng tôi đào tạo trước 250k bước trong giai đoạn đầu tiên và 80k bước trong giai đoạn thứ hai. Chúng tôi sử dụng kích thước lô là 2320/1680 cho ViT-L/ViT-g trong giai đoạn đầu tiên và kích thước lô là 1920/1520 cho OPT/FlanT5 trong giai đoạn thứ hai. Trong quá trình đào tạo trước, chúng tôi chuyển đổi các tham số ViTs và LLM đã đóng băng thành FP16, ngoại trừ FlanT5, nơi chúng tôi sử dụng BFloat16. Chúng tôi không thấy hiệu suất bị suy giảm so với khi sử dụng các mô hình 32 bit. Do sử dụng các mô hình đóng băng, quá trình đào tạo trước của chúng tôi thân thiện với tính toán hơn so với các phương pháp VLP quy mô lớn hiện có. Ví dụ, khi sử dụng một máy 16-A100(40G) duy nhất, mô hình lớn nhất của chúng tôi với ViT-g và FlanT5-XXL cần ít hơn 6 ngày cho giai đoạn đầu tiên và ít hơn 3 ngày cho giai đoạn thứ hai.

Cùng một bộ siêu tham số tiền huấn luyện được sử dụng cho tất cả các mô hình. Chúng tôi sử dụng trình tối ưu hóa AdamW (Loshchilov & Hutter, 2017) với $\beta_1 = 0.9$, $\beta_2 = 0.98$ và suy giảm trọng số là 0.05. Chúng tôi sử dụng suy giảm tốc độ học cosin với tốc độ học định là 1e-4 và khởi động tuyển tính là 2k bước. Tốc độ học tối thiểu ở giai đoạn thứ hai là 5e-5. Chúng tôi sử dụng hình ảnh có kích thước 224x224, được tăng cường bằng cắt xén thay đổi kích thước ngẫu nhiên và lật ngang.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

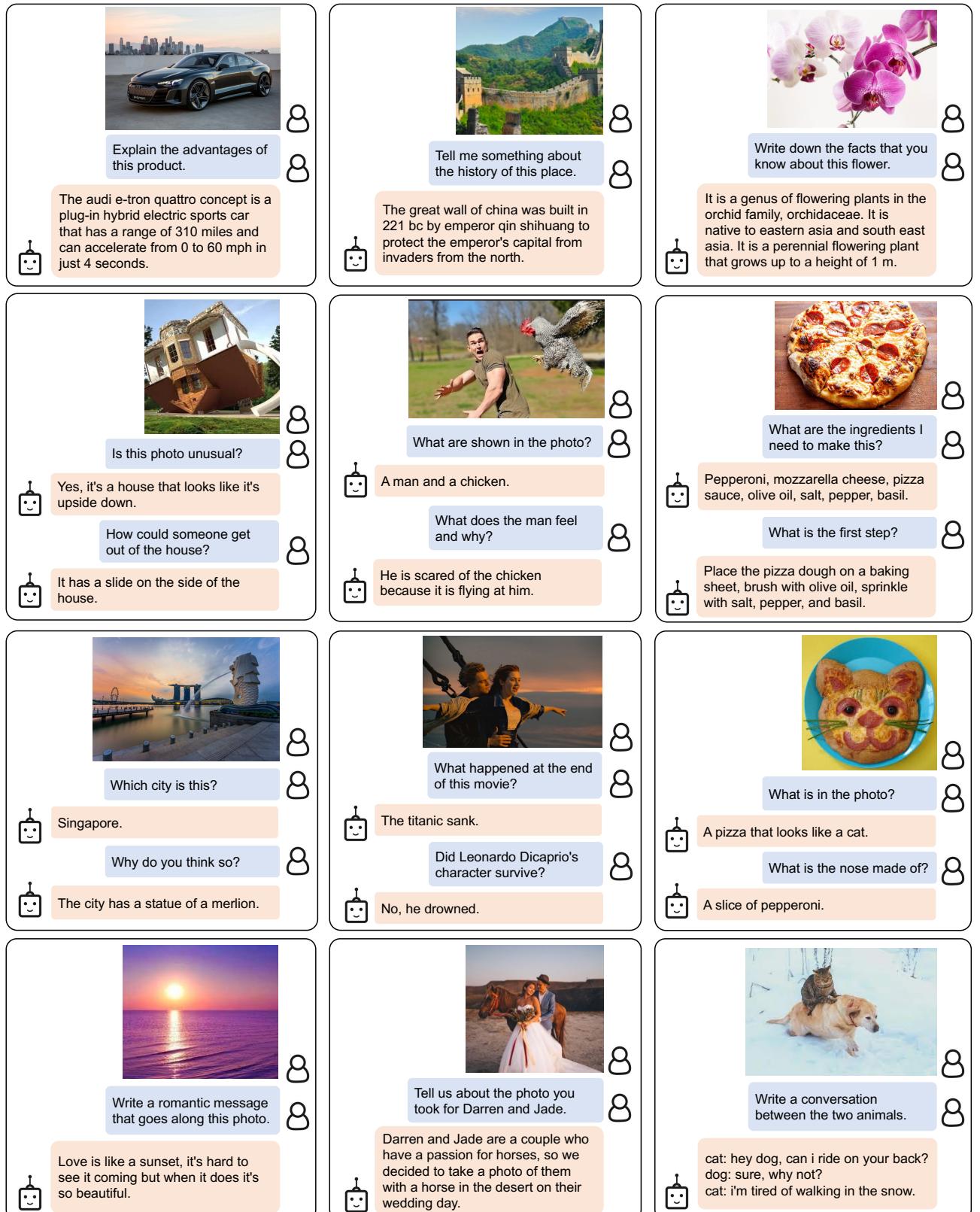
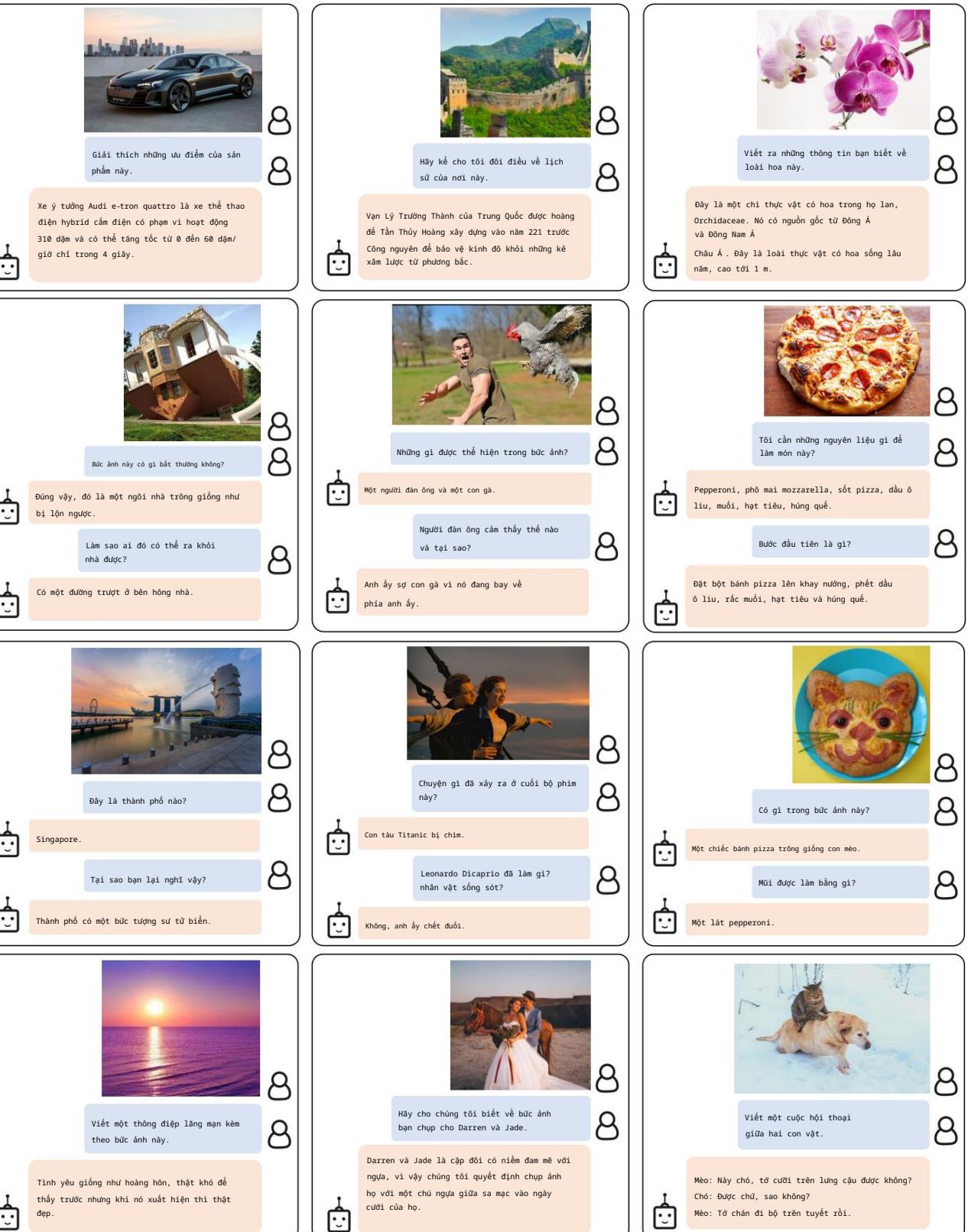


Figure 4. Selected examples of **instructed zero-shot image-to-text generation** using a BLIP-2 model w/ ViT-g and FlanT5XXL, where it shows a wide range of capabilities including visual conversation, visual knowledge reasoning, visual commonsense reasoning, storytelling, personalized image-to-text generation, etc.

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn



Hình 4. Các ví dụ được chọn về việc tạo hình ảnh thành văn bản không có hướng dẫn bằng cách sử dụng mô hình BLIP-2 với ViT-g và FlanT5XXL, trong đó thể hiện nhiều khả năng bao gồm giao tiếp trực quan, lý luận kiến thức trực quan, lý luận tương đồng trực quan, kể chuyện, tạo hình ảnh thành văn bản được cá nhân hóa, v.v.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Models	#Trainable Params	Open-sourced?	Visual Question Answering		Image Captioning		Image-Text Retrieval	
			VQAv2 (test-dev) VQA acc.	NoCaps (val) CIDEr	SPICE	TR@1	IR@1	
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7	
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-	
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5	
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-	
BLIP-2	188M	✓	65.0	121.6	15.8	97.6	89.7	

Table 1. Overview of BLIP-2 results on various **zero-shot** vision-language tasks. Compared with previous state-of-the-art models, BLIP-2 achieves the highest zero-shot performance while requiring the least number of trainable parameters during vision-language pre-training.

Models	#Trainable Params	#Total Params	VQAv2		OK-VQA test	GQA test-dev
			val	test-dev		
VL-T5 _{no-vqa}	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	50.6	-
BLIP-2 ViT-L OPT _{2.7B}	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT _{2.7B}	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT _{6.7B}	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5 _{XL}	103M	3.4B	62.6	62.3	39.4	<u>44.4</u>
BLIP-2 ViT-g FlanT5 _{XL}	107M	4.1B	<u>63.1</u>	<u>63.0</u>	40.7	44.2
BLIP-2 ViT-g FlanT5 _{XXL}	108M	12.1B	65.2	65.0	<u>45.9</u>	44.7

Table 2. Comparison with state-of-the-art methods on zero-shot visual question answering.

4. Experiment

Table 1 provides an overview of the performance of BLIP-2 on various zero-shot vision-language tasks. Compared to previous state-of-the-art models, BLIP-2 achieves improved performance while requiring substantially fewer number of trainable parameters during vision-language pre-training.

4.1. Instructed Zero-shot Image-to-Text Generation

BLIP-2 effectively enables a LLM to understand images while preserving its capability in following text prompts, which allows us to control image-to-text generation with instructions. We simply append the text prompt after the visual prompt as input to the LLM. Figure 4 shows examples to demonstrate a wide range of zero-shot image-to-text capabilities including visual knowledge reasoning, visual commonsense reasoning, visual conversation, personalized image-to-text generation, etc.

Zero-shot VQA. We perform quantitative evaluation on the zero-shot visual question answering task. For OPT models, we use the prompt “Question: {} Answer:”. For FlanT5 models, we use the prompt “Question: {} Short answer:”.

During generation, we use beam search with a beam width of 5. We also set the length-penalty to -1 which encourages shorter answers that align better with human annotation.

As shown in Table 2, BLIP-2 achieves state-of-the-art result on the VQAv2 (Goyal et al., 2017) and GQA (Hudson & Manning, 2019) datasets. It outperforms Flamingo80B by 8.7% on VQAv2, despite having 54x fewer trainable parameters. On the OK-VQA (Marino et al., 2019) dataset, BLIP-2 comes secondary to Flamingo80B. We hypothesis that this is because OK-VQA focuses more on open-world knowledge than visual understanding, and the 70B Chinchilla (Hoffmann et al., 2022) language model from Flamingo80B possesses more knowledge than the 11B FlanT5XXL.

We make a promising observation from Table 2: **a stronger image encoder or a stronger LLM both lead to better performance.** This observation is supported by several facts: (1) ViT-g outperforms ViT-L for both OPT and FlanT5. (2) Within the same LLM family, larger models outperform smaller ones. (3) FlanT5, an instruction-tuned LLM, outperforms the unsupervised-trained OPT on VQA. This observation validates BLIP-2 as a **generic vision-language pre-training method** that can efficiently harvest the rapid advances in vision and natural language communities.

Effect of Vision-Language Representation Learning.

The first-stage representation learning pre-trains the Q-Former to learn visual features relevant to the text, which reduces the burden of the LLM to learn vision-language alignment. Without the representation learning stage, Q-

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn

Mô hình	#Có thể huấn luyện Tham số	Mã nguồn mở?	Trả lời câu hỏi trực quan Chú thích hình ảnh Truy xuất hình ảnh-văn bản					
			Trả lời câu hỏi trực quan Chú thích hình ảnh Truy xuất hình ảnh-văn bản VQAv2 (kiểm thử-phát triển) Theo VQA	NoCaps (hợp lệ) CIDER SPICE TR@1 IR@1	Flickr (thử nghiệm)	CIDEr	SPICE	TR@1
BLIP (Li và cộng sự, 2022)	583 triệu	-	-	113,2	14.8	96,7	86,7	
SimVLM (Wang và cộng sự, 2021b)	1.4B BEiT-3	-	-	112,2	-	-	-	
(Wang và cộng sự, 2022b)	1,9B	-	-	-	-	94,9	81,5	
Hồng hạc (Alayrac và cộng sự, 2022)	10,2B	-	56,3	-	-	-	-	
BLIP-2	188 triệu	-	-	65,0	121,6	15,8	97,6	89,7

Bảng 1. Tổng quan về kết quả BLIP-2 trên nhiều nhiệm vụ ngôn ngữ thị giác zero-shot khác nhau. So sánh với các mô hình tiên tiến trước đây, BLIP-2 đạt được hiệu suất zero-shot cao nhất trong khi yêu cầu số lượng tham số có thể đào tạo ít nhất trong quá trình đào tạo trước ngôn ngữ thị giác.

Mô hình	#Có thể huấn luyện Tham số	#Tổng cộng Tham số	VQAv2		OK-VQA GQA	
			val	test-dev	ok-vqa test	thử nghiệm-phát triển
VL-T5no-vqa	224M	269 triệu	13,5	-	5,8	6,3
fewVLM (Jin và cộng sự, 2022)	740M	785 triệu	47,7	-	16,5	29,3
Đông lạnh (Tsimpoukelli và cộng sự, 2021)	40 triệu	7,1 tỷ 29,6	832	-	5,9	-
VLKD (Dai và cộng sự, 2022)	406M	triệu 42,6	3,2 tỷ	44,5	13,3	-
Flamingo3B (Alayrac và cộng sự, 2022)	1,4B	-	49,2	41,2	-	-
Flamingo9B (Alayrac và cộng sự, 2022)	1,8B	9,3B	-	51,8	44,7	-
Flamingo80B (Alayrac và cộng sự, 2022)	10,2B	80B	-	56,3	50,6	-
BLIP-2 ViT-L OPT _{2.7B}	104M	3.1B	50,1	49,7	30,2	33,9
ViT-g OPT _{2.7B}	107M	3.8B	53,5	52,3	31,7	34,6
OPT _{6.7B}	108M	7.8B	54,3	52,6	36,4	36,4
FlanT5 _{XL} BLIP-2 ViT-g	103M	3.4B	62,6	62,3	39,4	<u>44,4</u>
FlanT5 _{XL} BLIP-2 ViT-g	107M	4.1B	<u>63,1</u>	<u>63,0</u>	40,7	44,2
FlanT5 _{XXL}	108M	12.1B	65,2	65,0	45,9	44,7

Bảng 2. So sánh với các phương pháp hiện đại về trả lời câu hỏi trực quan không cần nhìn.

4. Thí nghiệm

Bảng 1 cung cấp tổng quan về hiệu suất của BLIP-2 trên nhiều nhiệm vụ ngôn ngữ tầm nhìn không cần nhìn. So với các mô hình tiên tiến trước đây, BLIP-2 đạt được cải tiến hiệu suất trong khi yêu cầu số lượng ít hơn đáng kể các thông số có thể đào tạo được trong quá trình đào tạo trước về ngôn ngữ thị giác.

4.1. Hướng dẫn tạo ảnh thành văn bản không cần chụp

BLIP-2 cho phép LLM hiểu hình ảnh một cách hiệu quả trong khi vẫn giữ nguyên khả năng theo dõi các lời nhắc văn bản, cho phép chúng ta kiểm soát việc tạo hình ảnh thành văn bản bằng hướng dẫn. Chúng tôi chỉ cần thêm lời nhắc văn bản sau lời nhắc trực quan như đầu vào cho LLM. Hình 4 cho thấy các ví dụ để chứng minh một loạt các hình ảnh-thành-văn bản không-cảnh khả năng bao gồm lý luận kiến thức trực quan, trực quan lý luận tương xứng, cuộc trò chuyện trực quan, cá nhân hóa tạo hình ảnh thành văn bản, v.v.

VQA Zero-shot. Chúng tôi thực hiện đánh giá định lượng về nhiệm vụ trả lời câu hỏi trực quan không cần nhìn. Đối với các mô hình OPT, chúng tôi sử dụng lời nhắc “Câu hỏi: {} Trả lời:”. Đối với FlanT5 mô hình, chúng tôi sử dụng lời nhắc “Câu hỏi: {} Câu trả lời ngắn:”. Trong quá trình tạo, chúng tôi sử dụng tìm kiếm chum tia với chiều rộng chum tia của 5. Chúng tôi cũng đặt hình phạt chiều dài thành -1 để khuyến khích câu trả lời ngắn hơn phù hợp hơn với chủ đề của con người.

Như thể hiện trong Bảng 2, BLIP-2 đạt được kết quả tiên tiến trên VQAv2 (Goyal và cộng sự, 2017) và GQA (Hudson & Manning, 2019) bộ dữ liệu. Nó vượt trội hơn Flamingo80B bởi 8,7% trên VQAv2, mặc dù có ít hơn 54 lần các tham số có thể đào tạo. Trên tập dữ liệu OK-VQA (Marino et al., 2019), BLIP-2 thử yếu so với Flamingo80B. Chúng tôi đưa ra giả thuyết rằng đây là vì OK-VQA tập trung nhiều hơn vào kiến thức thế giới mở hơn là hiểu biết trực quan, và mô hình ngôn ngữ Chinchilla 70B (Hoffmann và cộng sự, 2022) từ Flamingo80B sở hữu nhiều kiến thức hơn so với FlanT5XXL 11B.

Chúng tôi đưa ra một quan sát đầy hứa hẹn từ Bảng 2: một bộ mã hóa hình ảnh hoặc LLM mạnh hơn đều dẫn đến hiệu suất tốt hơn. Quan sát này được hỗ trợ bởi một số sự kiện:

(1) ViT-g vượt trội hơn ViT-L đối với cả OPT và FlanT5. (2) Trong cùng một họ LLM, các mô hình lớn hơn có hiệu suất vượt trội hơn những cái nhỏ hơn. (3) FlanT5, một LLM được điều chỉnh theo hướng dẫn, thực hiện tốt hơn OPT được đào tạo không có giám sát trên VQA.

Quan sát này xác nhận BLIP-2 là một ngôn ngữ thị giác chung phương pháp đào tạo trước có thể thu hoạch hiệu quả nhanh chóng tiến bộ trong tầm nhìn và cộng đồng ngôn ngữ tự nhiên.

Hiệu ứng của việc học biểu diễn ngôn ngữ thị giác.

Việc học biểu diễn giai đoạn đầu tiên sẽ đào tạo trước Q-Former để học các đặc điểm trực quan có liên quan đến văn bản, giảm bớt gánh nặng của LLM trong việc học ngôn ngữ thị giác cẩn chỉnh. Không có giai đoạn học biểu diễn, Q-

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Models	#Trainable Params	NoCaps Zero-shot (validation set)								COCO Fine-tuned	
		in-domain		near-domain		out-domain		overall		Karpathy test	B@4
		C	S	C	S	C	S	C	S		C
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	80.9	11.3	37.4	127.8
VinVL (Zhang et al., 2021)	345M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP (Li et al., 2022)	446M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7
OFA (Wang et al., 2022a)	930M	-	-	-	-	-	-	-	-	43.9	145.3
Flamingo (Alayrac et al., 2022)	10.6B	-	-	-	-	-	-	-	-	138.1	
SimVLM (Wang et al., 2021b)	~1.4B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP-2 ViT-g OPT _{2.7B}	1.1B	123.0	15.8	117.8	15.4	123.4	15.1	119.7	15.4	43.7	145.8
BLIP-2 ViT-g OPT _{6.7B}	1.1B	123.7	15.8	119.2	15.3	124.4	14.8	121.0	15.3	43.5	145.2
BLIP-2 ViT-g FlanT5 _{XL}	1.1B	123.7	16.3	120.2	15.9	124.8	15.1	121.6	15.8	42.4	144.5

Table 3. Comparison with state-of-the-art image captioning methods on NoCaps and COCO Caption. All methods optimize the cross-entropy loss during finetuning. C: CIDEr, S: SPICE, B@4: BLEU@4.

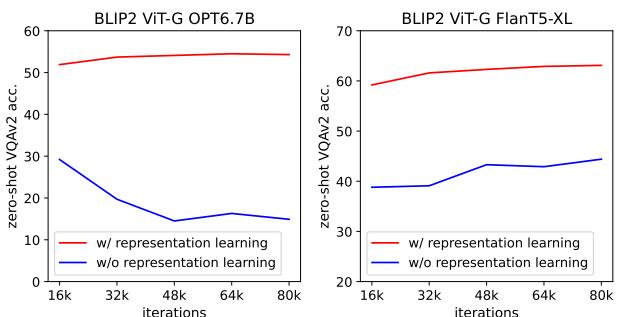


Figure 5. Effect of vision-language representation learning on vision-to-language generative learning. Without representation learning, the Q-Former fails the bridge the modality gap, leading to significantly lower performance on zero-shot VQA.

Former relies solely on the vision-to-language generative learning to bridge the modality gap, which is similar to the Perceiver Resampler in Flamingo. Figure 5 shows the effect of representation learning on generative learning. Without representation learning, both types of LLMs give substantially lower performance on zero-shot VQA. In particular, OPT suffers from catastrophic forgetting where performance drastically degrades as training proceeds.

4.2. Image Captioning

We finetune BLIP-2 models for the image captioning task, which asks the model to generate a text description for the image’s visual content. We use the prompt “a photo of” as an initial input to the LLM and trains the model to generate the caption with the language modeling loss. We keep the LLM frozen during finetuning, and updates the parameters of the Q-Former together with the image encoder. We experiment with ViT-g and various LLMs. Detailed hyperparameters can be found in the appendix. We perform finetuning on COCO, and evaluate on both COCO test set and zero-shot transfer to NoCaps (Agrawal et al., 2019) validation set.

The results are shown in Table 3. BLIP-2 achieves state-

Models	#Trainable Params	VQAv2	
		test-dev	test-std
<i>Open-ended generation models</i>			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-g FlanT5_{XL}	1.2B	81.55	81.66
BLIP-2 ViT-g OPT_{2.7B}	1.2B	81.59	81.74
BLIP-2 ViT-g OPT_{6.7B}	1.2B	82.19	82.30
<i>Closed-ended classification models</i>			
VinVL	345M	76.52	76.60
SimVLM (Wang et al., 2021b)	~1.4B	80.03	80.34
CoCa (Yu et al., 2022)	2.1B	82.30	82.30
BEIT-3 (Wang et al., 2022b)	1.9B	84.19	84.03

Table 4. Comparison with state-of-the-art models fine-tuned for visual question answering.

of-the-art performance with significant improvement on NoCaps over existing methods, demonstrating strong generalization ability to out-domain images.

4.3. Visual Question Answering

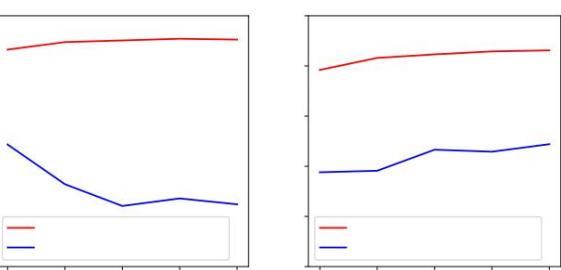
Given annotated VQA data, we finetune the parameters of the Q-Former and the image encoder while keeping the LLM frozen. We finetune with the open-ended answer generation loss, where the LLM receives Q-Former’s output and the question as input, and is asked to generate the answer. In order to extract image features that are more relevant to the question, we additionally condition Q-Former on the question. Specifically, the question tokens are given as input to the Q-Former and interact with the queries via the self-attention layers, which can guide the Q-Former’s cross-attention layers to focus on more informative image regions.

Following BLIP, our VQA data includes the training and validation splits from VQAv2, as well as training samples from Visual Genome. Table 4 demonstrates the state-of-the-art results of BLIP-2 among open-ended generation models.

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đông bằng và Mô hình Ngôn ngữ Lớn

Mô hình	#Có thể huấn luyện Tham số	NoCaps Zero-shot (bộ xác thực) gần miền				tổng thể C	COCO tinh chỉnh Kiểm tra Karpathy C
		trong miền C	ngoài miền SCSCS	tổng thể C	SB@4		
OSCAR (Li và cộng sự, 2020)	345 triệu	-	-	-	-	80,9	11,3
VinVL (Zhang và cộng sự, 2021)	345 triệu	103,1	14,2	96,1	13,8	88,3	95,5
BLIP (Li và cộng sự, 2022)	446M	114,9	15,2	112,1	14,9	115,3	13,5
OFA (Wang và cộng sự, 2022a)	930M	-	-	-	-	-	-
Flamingo (Alayrac và cộng sự, 2022)	10.6B	-	-	-	-	-	-
SimVLM (Wang và cộng sự, 2021b)	~1.4B	113,7	-	110,9	-	115,2	-
BLIP-2 ViT-g OPT_{2.7B}	1.1B	123,0	15,8	117,8	15,4	123,4	15,1
BLIP-2 ViT-g OPT_{6.7B}	1.1B	123,7	15,8	119,2	15,3	124,4	14,8
BLIP-2 ViT-g FlanT5_{XL}	1.1B	123,7	16,3	120,2	15,9	124,8	15,1

Bảng 3. So sánh với các phương pháp chú thích hình ảnh hiện đại trên NoCaps và COCO Caption. Tất cả các phương pháp đều tối ưu hóa tần số thất entroy chéo trong quá trình tinh chỉnh. C: CIDEr, S: SPICE, B@4: BLEU@4.



Hình 5. Tác động của việc học biểu diễn ngôn ngữ thị giác lên học tập tạo ra từ tầm nhìn đến ngôn ngữ. Không có sự biểu diễn học tập, cả hai loại LLM đều cho hiệu suất thấp hơn đáng kể trên VQA không có lằn kiểm tra nào. Đặc biệt, OPT bị lãng quên thảm khốc nơi hiệu suất giảm sút nghiêm trọng khi quá trình đào tạo diễn ra.

Trước đây chỉ dựa vào sự tạo ra tầm nhìn thành ngôn ngữ học cách thu hẹp khoảng cách phương thức, tương tự như Bộ lấy mẫu lại bộ thu phát trong Flamingo. Hình 5 cho thấy hiệu ứng của việc học biểu diễn trên việc học tạo ra. Không có học biểu diễn, cả hai loại LLM đều cho hiệu suất thấp hơn đáng kể trên VQA không có lằn kiểm tra nào. Đặc biệt, OPT bị lãng quên thảm khốc nơi hiệu suất giảm sút nghiêm trọng khi quá trình đào tạo diễn ra.

4.2. Chú thích hình ảnh

Chúng tôi tinh chỉnh các mô hình BLIP-2 cho nhiệm vụ chú thích hình ảnh, yêu cầu mô hình tạo ra một mô tả văn bản cho nội dung trực quan của hình ảnh. Chúng tôi sử dụng lời nhắc “một bức ảnh của” như một đầu vào ban đầu cho LLM và đào tạo mô hình để tạo ra chú thích với mô hình ngôn ngữ mắt mèo. Chúng tôi giữ LLM đông lạnh trong quá trình tinh chỉnh và cập nhật các thông số của Q-Former cùng với bộ mã hóa hình ảnh. Chúng tôi thử nghiệm với ViT-g và nhiều LLM khác nhau. Siêu tham số chi tiết có thể được tìm thấy trong phần phụ lục. Chúng tôi thực hiện tinh chỉnh trên COCO và đánh giá trên cả bộ thử nghiệm COCO và zero-shot chuyển sang bộ xác thực NoCaps (Agrawal và cộng sự, 2019).

Kết quả được thể hiện trong Bảng 3. BLIP-2 đạt được trạng thái

Mô hình	#Có thể huấn luyện Tham số	VQAv2	
		kiểm tra-phát triển kiểm tra	kiểm tra
<i>Mô hình thế hệ mới</i>			
ALBEF (Li và cộng sự, 2021)	314M	75,84	76,04
BLIP (Li và cộng sự, 2022)	385M	78,25	78,32
OFA (Wang và cộng sự, 2022a)	930M	82,00	82,00
Flamingo80B (Alayrac và cộng sự, 2022)	10.6B	82,00	82,10
BLIP-2 ViT-g FlanT5 _{XL}	1.2B	81,55	81,66
OPT2.7B 1.2B BLIP-2 ViT-g OPT _{6.7B} 1.2B	81,59	81,74	82,19
OPT2.7B 1.2B BLIP-2 ViT-g OPT _{2.7B} 1.2B	81,59	81,74	82,19
<i>Mô hình phân loại đóng</i>			
VinVL 345M SimVLM (Wang và cộng sự, 2021b)	1,4B CoCa (Yu và cộng sự, 2022)	76,52	76,60
1,4B CoCa (Yu và cộng sự, 2022)	2.1B	80,03	80,34
BEIT-3 (Wang và cộng sự, 2022b			

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
<i>Dual-encoder models</i>													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	100.0	81.5	95.6	97.8	84.8	96.5	98.3	67.2	87.7	92.8
<i>Fusion-encoder models</i>													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
<i>Dual encoder + Fusion encoder reranking</i>													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	100.0	100.0	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	100.0	100.0	88.6	97.6	98.9	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-g	1.2B	97.6	100.0	100.0	89.7	98.1	98.9	85.4	97.0	98.5	68.3	87.7	92.6

Table 5. Comparison with state-of-the-art image-text retrieval methods, finetuned on COCO and zero-shot transferred to Flickr30K.

COCO finetuning objectives	Image → Text		Text → Image		Hình ảnh	Văn bản						
	R@1	R@5	R@1	R@5								
ITC + ITM	84.5	96.2	67.2	87.1								
ITC + ITM + ITG	85.4	97.0	68.3	87.7								

Table 6. The image-grounded text generation (ITG) loss improves image-text retrieval performance by enforcing the queries to extract language-relevant visual features.

4.4. Image-Text Retrieval

Since image-text retrieval does not involve language generation, we directly finetune the first-stage-pretrained model w/o LLM. Specifically, we finetune the image encoder together with Q-Former on COCO using the same objectives (*i.e.* ITC, ITM, and ITG) as pre-training. We then evaluate the model for both image-to-text retrieval and text-to-image retrieval on COCO and Flickr30K (Plummer et al., 2015) datasets. During inference, we follow Li et al. (2021; 2022) which first select $k = 128$ candidates based on the image-text feature similarity, followed by a re-ranking based on pairwise ITM scores. We experiment with both ViT-L and ViT-g as the image encoder. Detailed hyperparameters can be found in the appendix.

The results are shown in Table 5. BLIP-2 achieves state-of-the-art performance with significant improvement over existing methods on zero-shot image-text retrieval.

The ITC and ITM losses are essential for image-text retrieval as they directly learn image-text similarity. In Table 6, we show that the ITG (image-grounded text generation) loss is also beneficial for image-text retrieval. This result supports our intuition in designing the representation learning objectives: the ITG loss enforces the queries to extract visual features most relevant to the text, thus improving vision-language alignment.

5. Limitation

Recent LLMs can perform in-context learning given few-shot examples. However, our experiments with BLIP-2 do not observe an improved VQA performance when providing the LLM with in-context VQA examples. We attribute the lack of in-context learning capability to our pre-training dataset, which only contains a single image-text pair per sample. The LLMs cannot learn from it the correlation among multiple image-text pairs in a single sequence. The same observation is also reported in the Flamingo paper, which uses a close-sourced interleaved image and text dataset (M3W) with multiple image-text pairs per sequence. We aim to create a similar dataset in future work.

BLIP-2’s image-to-text generation could have unsatisfactory results due to various reasons including inaccurate knowledge from the LLM, activating the incorrect reasoning path, or not having up-to-date information about new image content (see Figure 7). Furthermore, due to the use of frozen models, BLIP-2 inherits the risks of LLMs, such as outputting offensive language, propagating social bias, or leaking private information. Remediation approaches include using instructions to guide model’s generation or training on a filtered dataset with harmful content removed.

6. Conclusion

We propose BLIP-2, a generic and compute-efficient method for vision-language pre-training that leverages frozen pre-trained image encoders and LLMs. BLIP-2 achieves state-of-the-art performance on various vision-language tasks while having a small amount of trainable parameters during pre-training. BLIP-2 also demonstrates emerging capabilities in zero-shot instructed image-to-text generation. We consider BLIP-2 as an important step towards building a multimodal conversational AI agent.

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn

Người mẫu	#Có thể huấn luyện Tham số	Flickr30K Zero-shot (bộ thử nghiệm 1K)						COCO Fine-tuned (bộ kiểm tra 5K)					
		Hình ảnh R@1	Văn bản R@5	Hình ảnh R@10	Văn bản R@5	Hình ảnh R@10	Văn bản R@1	Hình ảnh R@10	Văn bản R@5	Hình ảnh R@10	Văn bản R@1	Hình ảnh R@10	Văn bản R@5
<i>Mô hình mã hóa kép</i>													
CLIP (Radford và cộng sự, 2021)	428M	88,0	98,7	99,4	68,7	90,6	95,2	88,6	98,7	99,7	-	-	-
ALIGN (Jia và cộng sự, 2021)	820M	75,7	93,8	96,8	77,0	93,5	96,9	59,9	93,3	89,8	-	-	-
FILIP (Yao và cộng sự, 2022)	417M	89,8	99,2	99,8	75,0	93,4	96,3	78,9	94,4	97,4	61,2	84,3	90,6
Florence (Yuan và cộng sự, 2021)	893M	90,9	99,1	-	76,7	93,6	-	63,2	85,7	-	81,8	95,2	-
BEIT-3(Wang và cộng sự, 2022b)	1,9B	94,9	99,9	100.0	81,5	95,6	97,8	84,8	96,5	98,3	67,2	87,7	92,8
<i>Mô hình mã hóa hợp nhất</i>													
UNITER (Chen và cộng sự, 2020)	303M	83,6	95,7	97,7	68,7	89,2	93,9	65,7	88,6	93,8	52,9	79,9	88,0
OSCAR (Li và cộng sự, 2020)	345M	-	-	-	-	-	-	-	-	-	70,0	91,1	95,5
VinVL (Zhang và cộng sự, 2021)	345M	-	-	-	-	-	-	-	-	-	75,4	92,9	96,2
<i>Xếp hạng lại bộ mã hóa kép + bộ mã hóa Fusion</i>													
ALBEF (Li và cộng sự, 2021)	233M	94,1	99,5	99,7	82,8	96,3	98,1	77,6	94,3	97,2	60,7	84,3	90,5
BLIP (Li và cộng sự, 2022)	446M	96,7	100.0	100.0	86,7	97,3	98,7	82,4	95,4	97,9	65,1	86,3	91,8
BLIP-2 ViT-L	474M	96,9	100.0	100.0	88,6	97,6	98,9	83,5	96,0	98,0	66,3	8	

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

References

- Agrawal, H., Anderson, P., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., and Lee, S. nocaps: novel object captioning at scale. In *ICCV*, pp. 8947–8956, 2019.

Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *NeurIPS*, 2020.

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

Chen, J., Guo, H., Yi, K., Li, B., and Elhoseiny, M. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *CVPR*, pp. 18009–18019, 2022a.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A. J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B. K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., and Soricut, R. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022b.

Chen, Y., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. UNITER: universal image-text representation learning. In *ECCV*, volume 12375, pp. 104–120, 2020.

Cho, J., Lei, J., Tan, H., and Bansal, M. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*, 2021.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S.,

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn

Tài liệu tham khảo

- Agrawal, H., Anderson, P., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., và Lee, S. nocaps: chú thích đối tượng mới lạ ở quy mô lớn. Trong ICCV, trang 8947–8956, 2019.

Narang, S., Mishra, G., Yu, A., Zhao, VY, Huang, Y., Dai, AM, Yu, H., Petrov, S., Chi, EH, Dean, J., Devlin, J., Roberts, A., Chu, D., Le, Q., và Wei, J. Chia tý lệ mô hình ngôn ngữ được tinh chỉnh theo hướng dẫn. bản in trước arXiv arXiv:2210.11416, 2022.

Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Has-son, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., và Simonyan, K. Flamingo: mô hình ngôn ngữ hình ảnh dành cho một số ít - học bẩn. bản in trước arXiv arXiv:2204.14198, 2022.

Brown, TB., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, DM, Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., và Amodei, D. Các mô hình ngôn ngữ là những người học ít lần. Trong Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., và Lin, H. (eds.), *NeurIPS*, 2020.

Changpinyo, S., Sharma, P., Ding, N., và Soricut, R. Khái niệm
12M: Đầy mạnh quá trình đào tạo trước hình ảnh-văn bản quy mô
web để nhận dạng các khái niệm trực quan đuôi dài. Trong CVPR, 2021.

Chen, J., Guo, H., Yi, K., Li, B., và Elhoseiny, M. Visu-algpt: Thích ứng hiếu quả dữ liệu của các mô hình ngôn ngữ được đào tạo trước để chú thích hình ảnh. Trong CVPR, trang 18009-18019, 2022a.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, AJ,
Padlewski, P., Salz, D., Goodman, S., Grycner, A.,
Mustafa, B., Beyer, L., Kolesnikov , A., Puigcerver, J.,
Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L.,
Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M.,
Jia, C., Ayan, BK, Riquelme, C., Steiner, A., Angelova ,
A., Zhai, X., Houlsby , N., và Soricut, R. Pali: Một mô
hình hình ảnh-ngôn ngữ đa ngôn ngữ được chia tỷ lệ
chung. bản in trước arXiv arXiv:2209.06794, 2022b.

Chen, Y., Li, L., Yu, L., Kholby, AE, Ahmed, F., Gan, Z., Cheng, Y., và Liu, J. UNITER: học biểu diễn hình ảnh-văn bản phổ quát. Trong ECCV, tập 12375, trang 104-120, 2020.

Cho, J., Lei, J., Tan, H. và Bansal, M. Thống nhất các nhiệm vụ về thị giác và ngôn ngữ thông qua việc tạo văn bản. Bản in trước arXiv arXiv:2102.02779, 2021.

Chung, HW, Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E.,
Wang, X., Dehghani, M., Brahma, S.,

son, A., Gu, SS., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A.,
Khang, S., Mishra, G., Yu, A., Zhao, VY., Huang, Y., Dai, AM., Yu, H.,
Trov, S., Chi, EH., Dean, J., Devlin, J., Roberts, A., Chu, D., Le, Q.,
Wei, J. Chia tý lệ mô hình ngôn ngữ được tính chính theo hướng dẫn.
in in trước arXiv arXiv:2210.11416, 2022.

W., Hou, L., Shang, L., Jiang, X., Liu, Q., và Fung, P. Kích hoạt khả năng tạo đa phương thức trên CLIP thông qua chất lọc kiến thức ngôn ngữ thị giác. Ở Muresan, S., Nakov, P., và Villavicencio, A. (eds.), *nhát hiện ACL*, trang 2383-2395 , 2022.

n., J., Chang, M., Lee, K., và Toutanova, K. BERT: đào tạo trước các biến đổi song hướng sâu để hiểu ngôn ngữ. Trong Burstein, J., Doran, và Solorio, T. (biên tập), NAACL, trang 4171-4186, 2019.

L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, , và Hon, H. Đào tạo trước mô hình ngôn ngữ thống nhất để hiểu và tạo ngôn ngữ tự nhiên. Trong Wallach, HM, Larochelle, H., Beygelzimer, , d'Alche-Buc, F., Fox, EB và Garnett, R. (biên tập), NeurIPS, trang 042-13054, 2019.

Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, , và Cao, Y. Eva: Khám phá giới hạn của việc học biểu diễn hình ảnh mặt nạ ở quy mô lớn. Bản in trước arXiv arXiv:2211.07636, 2022.

, Y., Khot, T., Summers-Stay, D., Batra, D., và Parikh, D. Làm cho Vũ V trong VQA trở nên quan trọng: Nâng cao vai trò của việc hiểu hình ảnh trong việc trả lời câu hỏi trực quan. Trong CVPR, trang 6325-6334, 2017.

J., Li, J., Li, D., Tiong, AMH, Li, B., Tao, D., và
bi, SCH Từ hình ảnh đến lời nhắc văn bản: VQA Zero-shot
với các mô hình ngôn ngữ lớn bị đóng băng. Trong CVPR, 2020.

ann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T.,
therford, E., Casas, D. d. L., Hendricks, LA, Welbl, J., Clark, A.,
nnigan, T., Noland, E., Millican, K., Driessche, G. vd, Damoc, B.,
y, A., Osindero, S., Simonyan, K., Elsen, E., Rae, JW, Vinyals, O. V.
lfre, L. Đào tạo các mô hình ngôn ngữ lớn tối ưu tính toán. bản in
cục arXiv arXiv:2203.15556, 2022.

on, DA và Manning, CD GQA: Một tập dữ liệu mới cho lý luận trực quan trong thế giới thực và trả lời câu hỏi áng tác. Trong CVPR, trang 6700-6709, 2019.

C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, QV, Yang, Y., Li, Z., và Duerig, T. Mở rộng quy mô học tập biểu diễn ngôn ngữ thi giác và ngôn ngữ thi giác với sự giám sát văn bản có nhiều. Bản trướcc arXiv arXiv:2102.05918, 2021.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

- Jin, W., Cheng, Y., Shen, Y., Chen, W., and Ren, X. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *ACL*, pp. 2763–2775, 2022.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.
- Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.
- Li, J., Li, D., Xiong, C., and Hoi, S. C. H. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900, 2022.
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pp. 121–137, 2020.
- Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *ECCV*, volume 8693, pp. 740–755, 2014.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Mañas, O., Rodríguez, P., Ahmadi, S., Nematzadeh, A., Goyal, Y., and Agrawal, A. MAPL: parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *EACL*, 2023.
- Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.
- Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *NIPS*, pp. 1143–1151, 2011.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
- Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Gurevych, I. and Miyao, Y. (eds.), *ACL*, pp. 2556–2565, 2018.
- Tan, H. and Bansal, M. LXMERT: learning cross-modality encoder representations from transformers. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *EMNLP*, pp. 5099–5110, 2019.
- Tiong, A. M. H., Li, J., Li, B., Savarese, S., and Hoi, S. C. H. Plug-and-play VQA: zero-shot VQA by conjoining large pretrained models with zero training. In *EMNLP Findings*, 2022.
- Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S. M. A., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *NeurIPS*, pp. 200–212, 2021.
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *ICML*, pp. 23318–23340, 2022a.
- Wang, W., Bao, H., Dong, L., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021a.
- Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022b.
- Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021b.
- Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. FILIP: fine-grained interactive language-image pre-training. In *ICLR*, 2022.
- Jin, W., Cheng, Y., Shen, Y., Chen, W., và Ren, X. Một lời nhắc tốt đáng giá hàng triệu tham số: Học tập dựa trên lời nhắc ít tài nguyên cho các mô hình ngôn ngữ thị giác. Trong Muresan, S., Nakov, P., và Villavicencio, A. (biên tập), *ACL*, trang 2763–2775, 2022.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D., Bernstein, M. S., và Fei-Fei, L. Bộ gen thị giác: Kết nối ngôn ngữ và thị giác bằng cách sử dụng chủ thích hình ảnh dày đặc do cộng đồng đóng góp. *IJCV*, 123(1):32–73, 2017.
- Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., và Hoi, S. C. H. Cân chỉnh trước khi kết hợp: Học biểu diễn ngôn ngữ và thị giác với chung cất động lượng. Trong *NeurIPS*, 2021.
- Li, J., Li, D., Xiong, C., và Hoi, S. C. H. BLIP: khởi động quá trình đào tạo trước ngôn ngữ-hình ảnh để hiểu và tạo ngôn ngữ-thị giác thống nhất. Trong *ICML*, trang 12888–12900, 2022.
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., và Gao, J. Oscar: Đào tạo trước theo nghĩa đối tượng cho các nhiệm vụ ngôn ngữ thị giác. Trong *ECCV*, trang 121–137, 2020.
- Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollar, P., và Zitnick, C. L. Microsoft COCO: các đối tượng phổ biến trong ngữ cảnh. Trong Häm đội, D. J., Pajdla, T., Schiele, B., và Tuytelaars, T. (eds.), *ECCV*, tập 8693, trang 740–755, 2014.
- Loshchilov, I., và Hutter, F. Chính quy hóa suy giảm trọng lượng tách biệt. Bản in trước *arXiv arXiv:1711.05101*, 2017.
- Manas, O., Rodríguez, P., Ahmadi, S., Nematzadeh, A., Goyal, Y., và Agrawal, A. MAPL: sự điều chỉnh hiệu quả về tham số của các mô hình được đào tạo trước đơn thức cho lời nhắc ngôn ngữ thị giác ít lần. Trong *EACL*, 2023.
- Marino, K., Rastegari, M., Farhadi, A., và Mottaghi, R. Ok-vqa: Một chuẩn mực trả lời câu hỏi trực quan đòi hỏi kiến thức bên ngoài. Trong *CVPR*, 2019.
- Ordonez, V., Kulkarni, G., và Berg, T. L. Im2text: Mô tả hình ảnh bằng 1 triệu bức ảnh có chủ thích. Trong Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., và Weinberger, K. Q. (eds.), *NIPS*, trang 1143–1151, 2011.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., và Lazebnik, S. Flickr30k thực thể: Thu thập sự tương ứng giữa vùng và cụm từ để có các mô hình hình ảnh và câu phong phú hơn. Trong *ICCV*, trang 2641–2649, 2015.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên. Bản in trước *arXiv arXiv:2103.00020*, 2021.
- Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., và Komatsuzaki, A. Laion-400m: Bộ dữ liệu mở gồm 400 triệu cặp hình ảnh-văn bản được lọc theo clip. Bản in trước của *arXiv arXiv:2111.02114*, 2021.
- Sharma, P., Ding, N., Goodman, S., và Soricut, R. 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 Gurevych, I., và Miyao, Y. (biên tập), *ACL*, trang 2556–2565, 2018.
- Tan, H., và Bansal, M. LXMERT: học các biểu diễn mã hóa đa phương thức từ các bộ biến đổi. Trong Inui, K., Jiang, J., Ng, V., và Wan, X. (biên tập), *EMNLP*, trang 5099–5110, 2019.
- Tiong, A. M. H., Li, J., Li, B., Savarese, S., và Hoi, S. C. H. Plug-and-play VQA: VQA zero-shot bằng cách kết hợp các mô hình được đào tạo trước lớn với đào tạo bằng không. Trong *EMNLP Findings*, 2022.
- Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S. M. A., Vinyals, O., và Hill, F. Học tập đa phương thức ít lần với các mô hình ngôn ngữ đóng băng. Trong Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., và Vaughan, J. W. (biên tập), *NeurIPS*, trang 200–212, 2021.
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., và Yang, H. OFA: thống nhất các kiến trúc, nhiệm vụ và phương thức thông qua một khuôn khổ học tập trình tự-trình tự đơn giản. Trong Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., và Sabato, S. (biên tập), *ICML*, trang 23318–23340, 2022a.
- Wang, W., Bao, H., Dong, L., và Wei, F. Vlmo: Đào tạo trước ngôn ngữ thị giác thống nhất với các chuyên gia kết hợp nhiều phương thức. Bản in trước *arXiv arXiv:2111.02358*, 2021a.
- Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., và Wei, F. Hình ảnh như một ngôn ngữ nước ngoài: Đào tạo trước Beit cho tất cả các nhiệm vụ thị giác và ngôn ngữ thị giác. Bản in trước *arXiv arXiv:2208.10442*, 2022b.
- Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., và Cao, Y. Simvlm: Đào tạo trước mô hình ngôn ngữ hình ảnh đơn giản với khả năng giám sát yếu. Bản in trước *arXiv arXiv:2108.10904*, 2021b.
- Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., và Xu, C. FILIP: đào tạo trước hình ảnh-nhận ngôn ngữ tương tác chi tiết. Trong *ICLR*, 2022.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yuan, L., Chen, D., Chen, Y., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pp. 18102–18112, 2022.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M. T., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. OPT: open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., và Wu, Y. Coca: Các偏好 là mô hình nền tăng hình ảnh-văn bản. Bản in trước arXiv arXiv:2205.01917, 2022.

Yuan, L., Chen, D., Chen, Y., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Chu, L. và Zhang, P. Florence: Mô hình nền tăng mới cho thị giác máy tính. Bản in trước arXiv arXiv:2111.11432, 2021.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., và Beyer, L. Lit: Chuyển giao không ảnh với điều chỉnh văn bản hình ảnh bị khóa. Trong *CVPR*, trang 18102-18112, 2022.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., và Gao, J. Vinvl: Làm cho các biểu diễn trực quan trở nên quan trọng trong các mô hình ngôn ngữ thị giác. Bản in trước arXiv arXiv:2101.00529, 2021.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, MT, Li, X., Lin, XV, Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, PS, Sridhar, A., Wang, T. và Zettlemoyer, L. OPT: mô hình ngôn ngữ biến đổi được đào tạo trước mờ. Bản in trước arXiv arXiv:2205.01068, 2022.

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

LLM	FlanT5XL	OPT _{2.7B}	OPT _{6.7B}
Fine-tuning epochs	5		
Warmup steps	1000		
Learning rate	1e-5		
Batch size	256		
AdamW β	(0.9,0.999)		
Weight decay	0.05		
Drop path	0		
Image resolution	364		
Prompt	“a photo of”		
Inference beam size	5		
Layer-wise learning rate decay for ViT	1	1	0.95

Table 7. Hyperparameters for fine-tuning BLIP-2 with ViT-g on COCO captioning.

LLM	FlanT5XL	OPT _{2.7B}	OPT _{6.7B}
Fine-tuning epochs	5		
Warmup steps	1000		
Learning rate	1e-5		
Batch size	128		
AdamW β	(0.9,0.999)		
Weight decay	0.05		
Drop path	0		
Image resolution	490		
Prompt	“Question: {} Answer:”		
Inference beam size	5		
Layer-wise learning rate decay for ViT	0.95	0.95	0.9

Table 8. Hyperparameters for fine-tuning BLIP-2 with ViT-g on VQA.

Image Encoder	ViT-L/14	ViT-g/14
Fine-tuning epochs	5	
Warmup steps	1000	
Learning rate	5e-6	1e-5
Batch size	224	
AdamW β	(0.9,0.98)	(0.9,0.999)
Weight decay	0.05	
Drop path	0	
Image resolution	364	
Layer-wise learning rate decay for ViT	1	0.95

Table 9. Hyperparameters for fine-tuning BLIP-2 on COCO image-text retrieval.

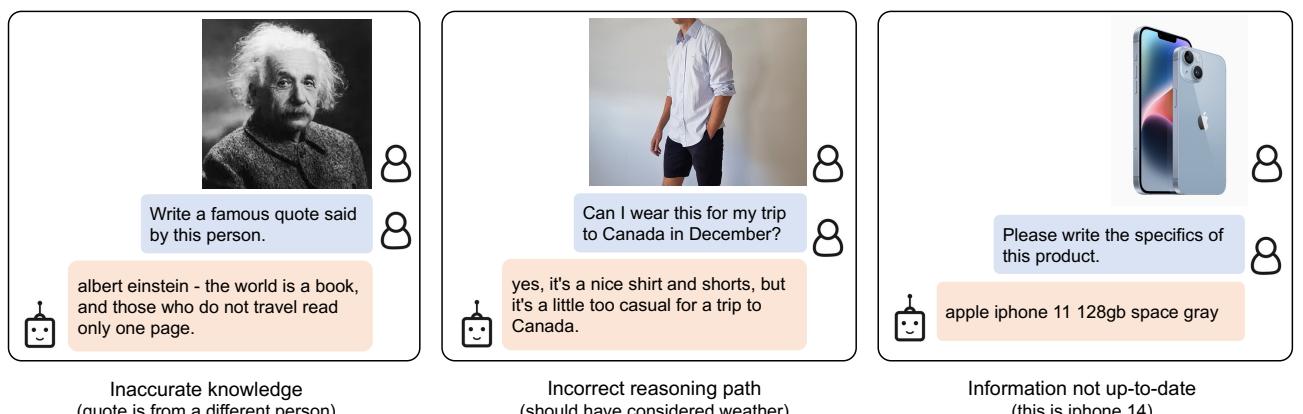


Figure 6. Incorrect output examples for instructed zero-shot image-to-text generation using a BLIP-2 model w/ ViT-g and FlanT5XXL.

BLIP-2: Khởi động quá trình đào tạo trước Ngôn ngữ-Hình ảnh với Bộ mã hóa Hình ảnh Đóng băng và Mô hình Ngôn ngữ Lớn

Thực tế Luật	FlanT5XL TÙY CHỌN2.7B TÙY CHỌN6.7B
Tinh chỉnh các ký nguyên	5
Các bước khởi động	1000
Tỷ lệ học tập	1e-5
Kích thước lô hàng	256
AdamW β	(0.9,0.999)
Giảm cân	0.05
Đường dẫn thả	0
Độ phân giải hình ảnh	364
Nhắc nhở	"một bức ảnh của"
Kích thước chùm suy luận	5
Tốc độ học tập theo từng lớp giảm dần cho ViT	1 1 0.95

Bảng 7. Siêu tham số để tinh chỉnh BLIP-2 với ViT-g trên chủ đề COCO.

Thực tế Luật	FlanT5XL TÙY CHỌN2.7B TÙY CHỌN6.7B
Tinh chỉnh các ký nguyên	5
Các bước khởi động	1000
Tỷ lệ học tập	1e-5
Kích thước lô hàng	128
AdamW β	(0.9,0.999)
Giảm cân	0.05
Đường dẫn thả	0
Độ phân giải hình ảnh	490
Nhắc nhở	"Câu hỏi: {} Trả lời:"
Kích thước chùm suy luận	5
Tốc độ học tập theo từng lớp giảm dần cho ViT	0.95 0.95 0.9

Bảng 8. Siêu tham số để tinh chỉnh BLIP-2 với ViT-g trên VQA.

Bộ mã hóa hình ảnh	ViT-L/14	ViT-g/14
Tinh chỉnh các ký nguyên	5	
Các bước khởi động	1000	
Tỷ lệ học tập	5e-6	1e-5
Kích thước lô hàng	224	
AdamW β	(0.9,0.98) (0.9,0.999)	
Giảm cân	0.05	
Đường dẫn thả	0	
Độ phân giải hình ảnh	364	
Tốc độ học tập theo từng lớp giảm dần cho ViT	1	0.95

Bảng 9. Siêu tham số để tinh chỉnh BLIP-2 khi truy xuất hình ảnh-văn bản COCO.



Hình 6. Ví dụ đầu ra không chính xác cho việc tạo hình ảnh thành văn bản không có hướng dẫn bằng cách sử dụng mô hình BLIP-2 với ViT-g và FlanT5XXL.

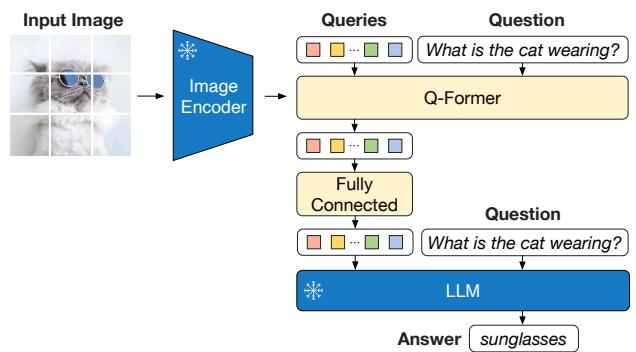
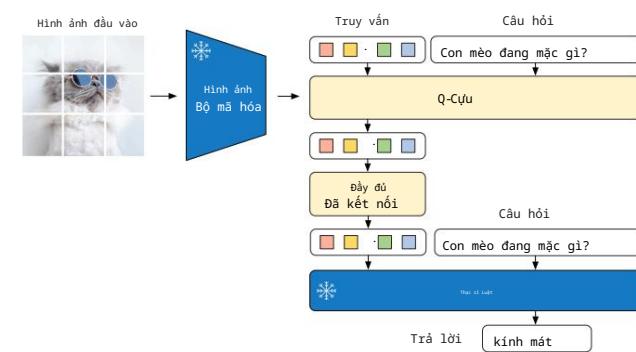


Figure 7. Model architecture for VQA finetuning, where the LLM receives Q-Former's output and the question as input, then predicts answers. We also provide the question as a condition to Q-Former, such that the extracted image features are more relevant to the question.



Hình 7. Kiến trúc mô hình cho tinh chỉnh VQA, trong đó LLM nhận đầu ra của Q-Former và câu hỏi làm đầu vào, sau đó dự đoán câu trả lời. Chúng tôi cũng cung cấp câu hỏi như một điều kiện cho Q-Former, sao cho các đặc điểm hình ảnh được trích xuất có liên quan hơn đến câu hỏi.