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 *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.

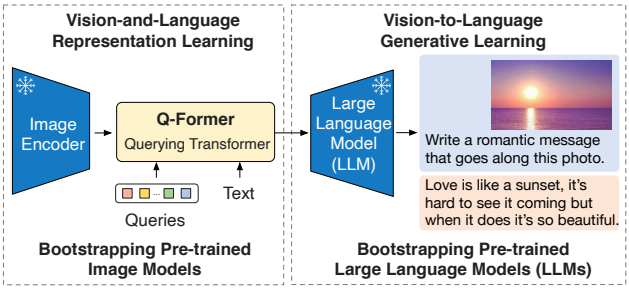

Figure 1. Overview of BLIP-2's framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen LLM, which enables zero-shot instructed image-to-text generation (see Figure 4 for more examples).

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：使用 Frozen Image 编码器和大型语言模型进行 Bootstrapping Language-Image Pre-training

李俊楠 李东旭 李 Silvio Savarese 史蒂文·许
Salesforce 研究
https://github.com/salesforce/LAVIS/tree/main/projects/blip2

抽象

由于大规模模型的端到端训练，视觉和语言预训练的成本变得越来越高。本文提出了 BLIP-2，这是一种通用且高效的预训练策略，它从现成的冻结预训练图像编码器和冻结的大型语言模型中引导视觉语言预训练。BLIP-2 通过轻量级 Querying Transformer 弥合了模态差距，该 Querying Transformer 分两个阶段进行预训练。第一阶段从冻结的图像编码器中引导视觉语言表示学习。第二阶段从冻结的语言模型引导视觉到语言的生成学习。BLIP-2 在各种视觉语言任务上实现了最先进的性能，尽管可训练参数明显少于现有方法。例如，我们的模型在零样本 VQAv2 上的性能比 Flamingo80B 高 8.7%，而可训练参数减少了 54 倍。我们还演示了该模型新兴的零镜头图像到文本生成功能，可以遵循自然语言指令。

1. 引言

视觉语言预训练 （VLP） 研究在过去几年中取得了快速发展，其中开发了规模越来越大的预训练模型，以不断推动各种下游任务的最新技术（Radford 等人，2021 年;Li et al.， 2021;2022;Wang et al.， 2022a;Alayrac et al.， 2022;Wang et al.， 2022b）。然而，由于使用大规模模型和数据集进行端到端训练，大多数最先进的视觉语言模型在预训练期间会产生高计算成本。

视觉语言研究位于视觉和语言之间的交叉点，因此自然而然地期望视觉语言模型可以从视觉和自然语言社区现成的单模态模型中收获。在本文中，我们提出了一种通用的

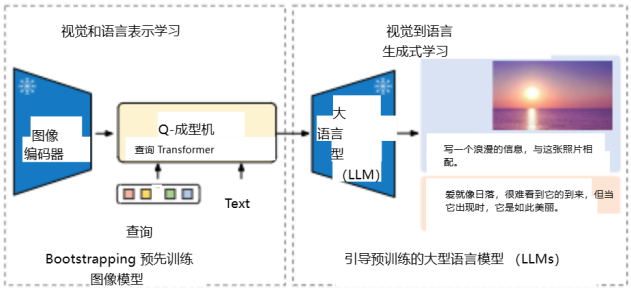

图 1.BLIP-2 框架概述。我们按照两阶段策略预先训练了一个轻量级的 Querying Transformer 来弥合模态差距。第一阶段引导从冻结图像编码器学习的 visionlanguage 表示。第二阶段从 frozen LLM 中引导视觉到语言的生成学习，从而实现零镜头定向图像到文本的生成（更多示例见图 4）。

通过从现成的预训练视觉模型和语言模型进行引导，实现高效的 VLP 方法。预先训练的视觉模型提供高质量的视觉表示。预先训练的语言模型，特别是大型语言模型 （），LLMs提供强大的语言生成和零样本迁移功能。为了降低计算成本并抵消灾难性遗忘的问题，单峰预训练模型在预训练期间保持冻结状态。

为了将预先训练的单峰模型用于 VLP，促进跨模态对齐是关键。然而，由于LLMs在单峰预训练期间没有看到图像，因此冻结它们会使视觉-语言对齐特别具有挑战性。在这方面，现有方法（例如 Frozen （Tsimpoukelli et al.， 2021）、Flamingo （Alayrac et al.， 2022））诉诸于图像到文本的生成损失，我们表明这不足以弥合模态差距。

为了实现与冻结单模态模型的有效视觉-语言对齐，我们提出了一个查询转换器 （QFormer），它使用新的两阶段预训练策略进行预训练。如图 1 所示，Q-Former 是一个轻量级转换器，它使用一组可学习的查询向量从冻结图像编码器中提取视觉特征。它充当了 freeze 图像编码器和 frozen LLM之间的信息瓶颈，在这里它馈送了最有用的

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 exisiting 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

Visual 功能输出LLM所需的文本。在第一个预训练阶段，我们进行视觉语言表示学习，强制 Q-Former 学习与文本最相关的视觉表示。在第二个预训练阶段，我们通过将 Q-Former 的输出连接到冻结LLM的 来执行视觉到语言的生成学习，并训练 Q-Former，使其输出的视觉表示可以被 LLM 解释。

我们将 VLP 框架命名为 BLIP-2： Bootstrapping Language-Image Pre-training with frozen unimodal models。BLIP-2 的主要优势包括：

- BLIP-2 有效地利用了冻结的预训练图像模型和语言模型。我们使用 Q-Former 来弥合模态差距，该 Q-Former 分两个阶段进行预训练：表征学习阶段和生成学习阶段。BLIP-2 在各种视觉语言任务上实现了最先进的性能，包括视觉问答、图像描述和图像文本检索。
- 在LLMs（例如 OPT（Zhang et al.， 2022）、FlanT5 （Chung et al.， 2022））的支持下，可以提示 BLIP-2 执行遵循自然语言指令的零镜头图像到文本生成，从而实现视觉知识推理、视觉对话等新兴功能（示例见图 4）。
- 由于使用了冻结的单峰模型和轻量级的 Q-Former，BLIP-2 的计算效率高于现有的最先进的模型。例如，BLIP-2 在零镜头 VQAv2 上比 Flamingo（Alayrac et al.， 2022）高出 8.7%，同时使用的可训练参数少 54×。此外，我们的结果表明，BLIP-2 是一种通用方法，可以收获更先进的单峰模型以获得更好的 VLP 性能。

2. 相关工作

2.1. 端到端视觉-语言预训练

视觉语言预训练旨在学习多模态基础模型，提高各种视觉和语言任务的性能。根据下游任务，已经提出了不同的模型架构，包括双编码器架构（Radford et al.， 2021;Jia et al.， 2021）、融合编码器架构（Tan & Bansal， 2019;Li et al.， 2021）、编码器-解码器架构（Cho et al.， 2021;Wang et al.， 2021b;Chen et al.， 2022b），以及最近的统一变压器架构（Li et al.， 2022;Wang et al.， 2022b）。多年来，还提出了各种训练前目标，并逐渐收敛为一些久经考验的目标：图像-文本对比学习（Radford et al.， 2021;Yao et al.， 2022;Li et al.， 2021;2022 年）、图像文本匹配（Li et al.， 2021;2022;Wang et al.， 2021a） 和（掩蔽的）语言建模 （Li et al.， 2021;2022;Yu et al.， 2022;Wang et al.， 2022b）。

大多数 VLP 方法使用大规模图像-文本对数据集执行端到端预训练。随着模型大小的不断增加，预训练可能会产生极高的计算成本。此外，端到端预训练模型利用现成的单峰预训练模型是不灵活的，例如LLMs（Brown et al.， 2020;Zhang et al.， 2022;Chung et al.， 2022）。

2.2. 模块化视觉-语言预训练

与我们更相似的是利用现成的预训练模型并在 VLP 期间将其冻结的方法。一些方法可以冻结图像编码器，包括早期采用冻结对象检测器提取视觉特征的工作（Chen et al.， 2020;Li et al.， 2020;Zhang等人，2021 年），以及最近的 LiT（Zhai 等人，2022 年），它使用冻结的预训练图像编码器进行 CLIP（Radford等人，2021 年）预训练。一些方法冻结语言模型以将知识LLMs用于视觉到语言的生成任务（Tsimpoukelli et al.， 2021;Alayrac et al.， 2022;Chen et al.， 2022a;马 nas 等人，2023 年;Tiong等人，2022 年;Guo et al.， 2022）。使用 frozen LLM 的关键挑战是将视觉特征与文本空间对齐。为了实现这一目标，Frozen（Tsimpoukelli et al.， 2021）微调了一个图像编码器，其输出直接用作 LLM.Flamingo （Alayrac et al.， 2022） 在 中LLM插入新的交叉注意力层以注入视觉特征，并在数十亿个图像文本对上预训练新层。两种方法都采用语言建模损失，即语言模型以图像为条件生成文本。

与现有方法不同，BLIP-2 可以有效且高效地利用冻结图像编码器和冻结LLMs进行各种视觉语言任务，以较低的计算成本实现更强的性能。

3. 方法

我们提出了 BLIP-2，一种新的视觉语言预训练方法，它从冻结的预训练单峰模型中引导而来。为了弥合模态差距，我们提出了一个分两个阶段进行预训练的 Querying Transformer （Q-Former）： （1） 带有冻结图像编码器的视觉-语言表示学习阶段和 （2） 带有冻结LLM的视觉到语言生成学习阶段。本节首先介绍了 Q-Former 的模型架构，然后描述了两个阶段的预训练过程。

3.1. 模型架构

我们建议 Q-Former 作为可训练模块，以弥合冻结图像编码器和冻结LLM图像编码器之间的差距。它从图像编码器中提取固定数量的输出特征，与输入图像分辨率无关。如图 2 所示，Q-Former 由两个 transformer 子模块组成，它们共享相同的自注意力层：（1） 一个与冻结图像编码器交互的图像 transformer

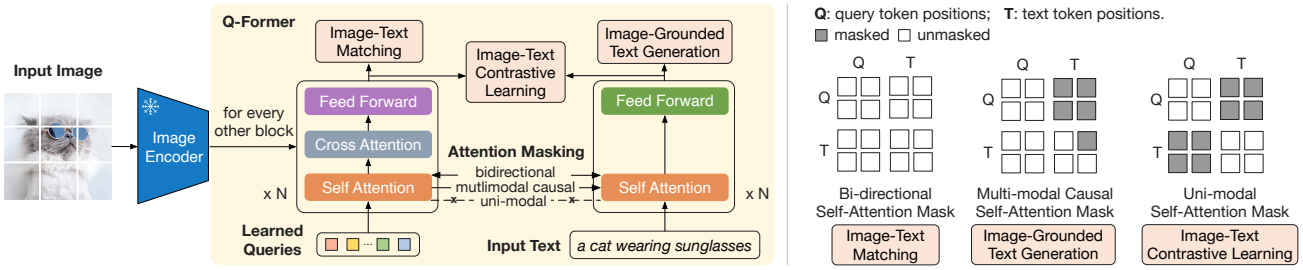

Figure 2. (**Left**) Model architecture of Q-Former and BLIP-2's first-stage vision-language representation learning objectives. We jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text. (**Right**) The self-attention 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 $\text{BERT}_{\text{base}}$ (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 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 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 Matching (ITM) aims to learn fine-grained alignment between image and text representation. It is a binary classification task where the model is asked to predict whether an image-text pair is positive (matched) or negative (unmatched). We use a bi-directional self-attention mask where all queries and texts can attend to each other. The output query embeddings Z thus capture multimodal information. We feed each output query embedding into a two-class linear classifier to obtain a logit, and average the logits across all queries as the output matching score. We adopt the hard negative mining strategy from Li et al. (2021; 2022) to create informative negative pairs.

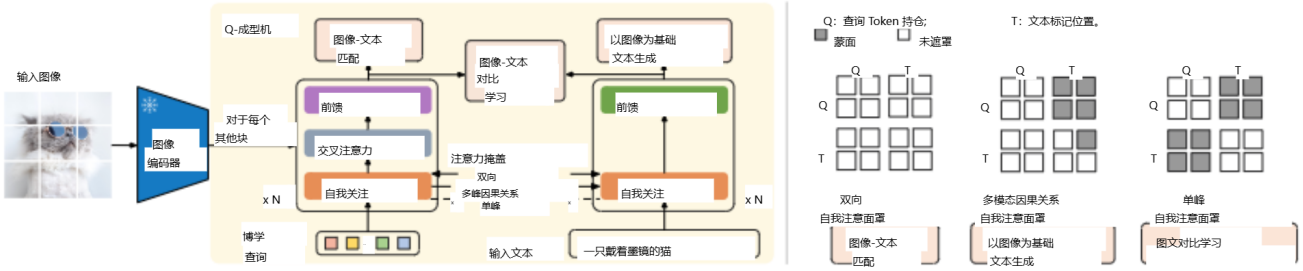

图 2.（左）Q-Former 和 BLIP-2 的第一阶段视觉-语言表示学习目标的模型架构。我们共同优化了三个目标，这些目标强制执行查询（一组可学习的嵌入）以提取与文本最相关的视觉表示。（右）用于控制查询文本交互的每个目标的自注意力掩码策略。

对于视觉特征提取，（2） 一个文本转换器，它既可以用作文本编码器，也可以用作文本解码器。我们创建一组可学习的查询嵌入作为 image transformer 的输入。查询通过自注意力层相互交互，并通过交叉注意力层（每隔一个 transformer 块插入一次）与冻结的图像特征交互。查询还可以通过相同的自我注意层与文本交互。根据预训练任务，我们应用不同的自我注意掩码来控制查询-文本交互。我们使用 BERT 的预训练权重初始化 QFormer （Devlin et al.， 2019），而交叉注意力层是随机初始化的。Q-Former 总共包含 188M 个参数。

请注意，查询被视为模型参数。

在我们的实验中，我们使用 32 个查询，其中每个查询的维度为 768（与 Q-Former 的隐藏维度相同）。我们使用 Z 来表示输出查询表示形式。Z 的大小（32 × 768）远小于冻结图像特征的大小（例如 ViT-L/14 为 257 × 1024）。这种瓶颈架构与我们的预训练目标一起工作，以强制查询提取与文本最相关的视觉信息。

3.2. Bootstrap Vision-Language Representation 从 Frozen Image 编码器中学习

在表示学习阶段，我们将 Q-Former 连接到冻结的图像编码器，并使用图像-文本对进行预训练。我们的目标是训练 Q-Former，以便查询可以学习提取对文本信息量最大的视觉表示。受 BLIP （Li et al.， 2022） 的启发，我们共同优化了三个具有相同输入格式和模型参数的预训练目标。每个目标在查询和文本之间采用不同的注意力掩蔽策略来控制它们的交互（参见图 2）。

图像-文本对比学习 （ITC） 学习对齐图像表示和文本表示，以便最大限度地利用它们的共同信息。它通过对比正对与负对的图像-文本相似性来实现这一点

我们将图像转换器的输出查询表示 Z 与文本转换器的文本表示 t 对齐，其中 t 是 [CLS] 标记的输出嵌入。由于 Z 包含多个输出嵌入（每个查询一个），我们首先计算每个查询输出与 t 之间的成对相似性，然后选择最高的一个作为图像-文本相似度。为避免信息泄露，我们采用了单模态自注意力掩码，其中 queries 和 text 不允许相互看到。由于使用了冻结图像编码器，与端到端方法相比，我们可以在每个 GPU 上容纳更多的样本。因此，我们在 BLIP 中使用批量内负数而不是动量队列。

图像基于的文本生成 （ITG） 损失会训练 Q-Former 生成文本，以输入图像为条件。由于 Q-Former 的架构不允许冻结图像编码器和文本令牌之间的直接交互，因此必须首先由查询提取生成文本所需的信息，然后通过自注意力层传递给文本令牌。因此，查询被迫提取捕获有关文本的所有信息的视觉特征。我们采用多模态因果自我注意掩码来控制查询-文本交互，类似于 UniLM 中使用的掩码 （Dong et al.， 2019）。查询可以相互处理，但不能处理文本标记。每个文本标记都可以处理所有查询及其以前的文本标记。我们还将 [CLS] 令牌替换为新的 [DEC] 令牌，作为向解码任务发出信号的第一个文本令牌。

图像-文本匹配 （ITM） 旨在学习图像和文本表示之间的细粒度对齐。这是一项二元分类任务，要求模型预测图像-文本对是正（匹配）还是负（不匹配）。我们使用双向的自我注意掩码，其中所有查询和文本都可以相互关注。因此，嵌入 Z 的输出查询会捕获多模式信息。我们将每个输出查询嵌入馈送到一个两类线性分类器中，以获得 logit，并将所有查询的 logit 平均为输出匹配分数。我们采用 Li 等人 （2021 年;2022 年）的硬负挖掘策略来创建信息丰富的负对。

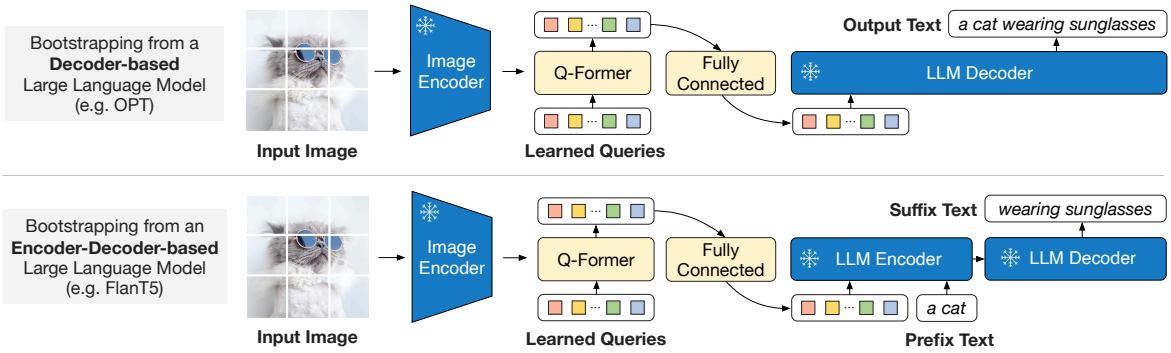

Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the 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 $\text{BLIP}_{\text{large}}$ 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 $\beta_1 = 0.9$, $\beta_1 = 0.98$, 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.

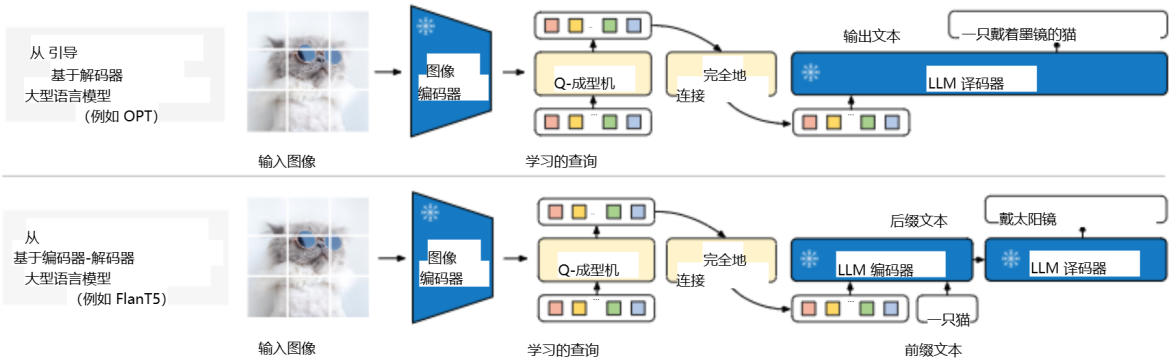

图 3.BLIP-2 的第二阶段视觉到语言生成式预训练，从冻结的大型语言模型 （）LLMs 引导。（页首）引导基于LLM解码器的 （例如 OPT）。（下）引导基于LLM编码器-解码器的 （例如 FlanT5）。全连接层从 Q-Former 的输出维度适应所选 LLM的输入维度。

3.3. 从冻结LLM开始引导视觉到语言生成学习

在生成式预训练阶段，我们将 QFormer（连接了冻结图像编码器）连接到冻结LLM项，以获得 LLM的生成语言能力。如图 3 所示，我们使用全连接 （FC） 层将输出查询嵌入 Z 线性投影到与 LLM.然后，投影的查询嵌入将添加到输入文本嵌入的前面。它们充当柔和的视觉提示，以调节 Q-Former 提取的LLM视觉表示。由于 Q-Former 已经经过预先训练以提取语言信息丰富的视觉表示，因此它有效地充当了信息瓶颈，将最有用的信息提供给 Q-FormerLLM，同时删除了不相关的视觉信息。这减轻了学习视觉-语言对齐的 LLM负担，从而减轻了灾难性的遗忘问题。

我们试验了两种类型的LLMs：基于LLMs解码器和基于 LLMs编码器-解码器。对于基于 decoderbased LLMs，我们使用语言建模损失进行预训练，其中 frozen LLM 的任务是生成以 Q-Former 的视觉表示为条件的文本。对于基于 LLMsencoder-decoder 的 ，我们使用前缀 language modeling loss 进行预训练，将文本分成两部分。前缀文本与视觉表示形式连接在一起，作为 LLM编码器的输入。后缀文本用作 LLM的 解码器的生成目标。

我们使用 BLIPcaptioning 模型生成 10 个字幕，并根据 CLIP ViT-L/14 模型生成的图像文本相似性对合成字幕和原始 Web 字幕进行排名。我们将每张图像的前两个字幕作为训练数据，并在每个预训练步骤中随机采样一个。

预训练图像编码器和 LLM.对于冻结的 im-年龄编码器，我们探索了两种最先进的预训练视觉转换器模型：（1） 来自 CLIP 的 ViT-L/14（Radford等人，2021 年）和 （2） 来自 EVA-CLIP 的 ViT-g/14（Fang等人，2022 年）。我们删除了 ViT 的最后一层，并使用倒数第二层的输出功能，这导致性能略好。对于冻结语言模型，我们探索了基于LLMs解码器的无监督训练的 OPT 模型族（Zhang et al.， 2022），以及基于编码器-解码器LLMs的指令训练的 FlanT5 模型族（Chung et al.， 2022）。

训练前设置。我们在第一阶段预训练 250k 步，在第二阶段预训练 80k 步。我们在第一阶段对 ViT-L/ViT-g 使用 2320/1680 的批次大小，在第二阶段对 OPT/FlanT5 使用 1920/1520 的批次大小。在预训练期间，我们将冻结的 ViTs' 和 LLMs' 参数转换为 FP16，但 FlanT5 除外，我们使用 BFloat16。我们发现与使用 32 位模型相比，性能没有下降。由于使用了冻结模型，我们的预训练比现有的大规模 VLP 方法更具计算友好性。例如，使用一台 16-A100 （40G） 机器，我们最大的型号配备 ViT-g 和 FlanT5-XXL 的第一级需要不到 6 天，第二级需要不到 3 天。

3.4. 模型预训练

预训练数据。我们使用与 BLIP 相同的预训练数据集，总共有 129M 张图像，包括 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） 和来自 LAION400M 数据集的 115M 图像（Schuhmann et al.， 2021）。我们采用 CapFilt 方法（Li等人，2022 年）为 Web 图像创建合成标题。具体说来

所有模型都使用同一组预训练超参数。我们使用 AdamW （Loshchilov & Hutter， 2017） 优化器，β= 0.9，β= 0.98，权重衰减为 0.05。我们使用余弦学习率衰减，峰值学习率为 1e-4，线性预热为 2k 步。第二阶段的最低学习率为 5e-5。我们使用大小为 224×224 的图像，并通过随机调整大小的裁剪和水平翻转进行增强。

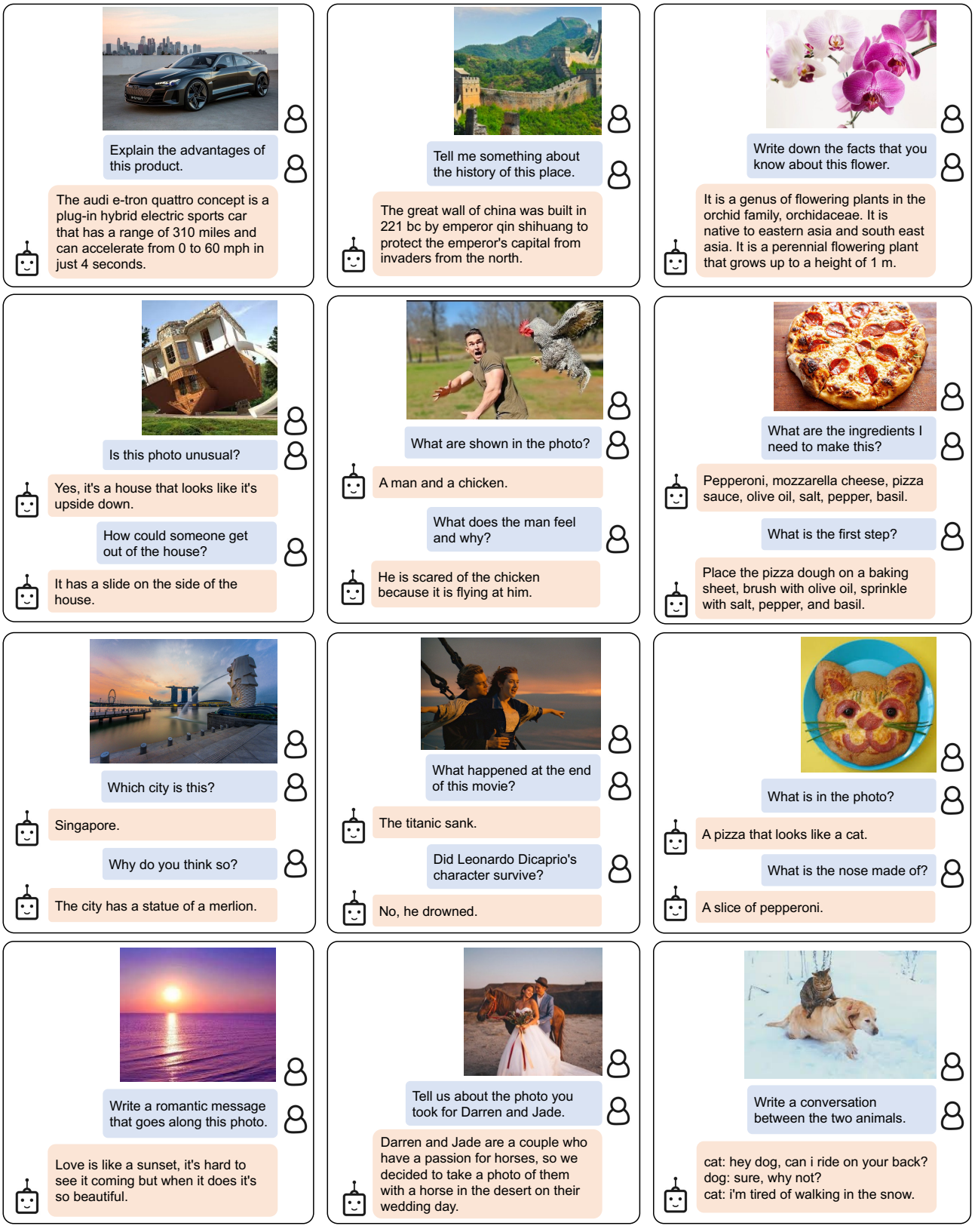

Figure 4. Selected examples of **instructed zero-shot image-to-text generation** using a BLIP-2 model w/ ViT-g and FlanT5$_{XXL}$, where it shows a wide range of capabilities including visual conversation, visual knowledge reasoning, visual commensense reasoning, storytelling, personalized image-to-text generation, etc.

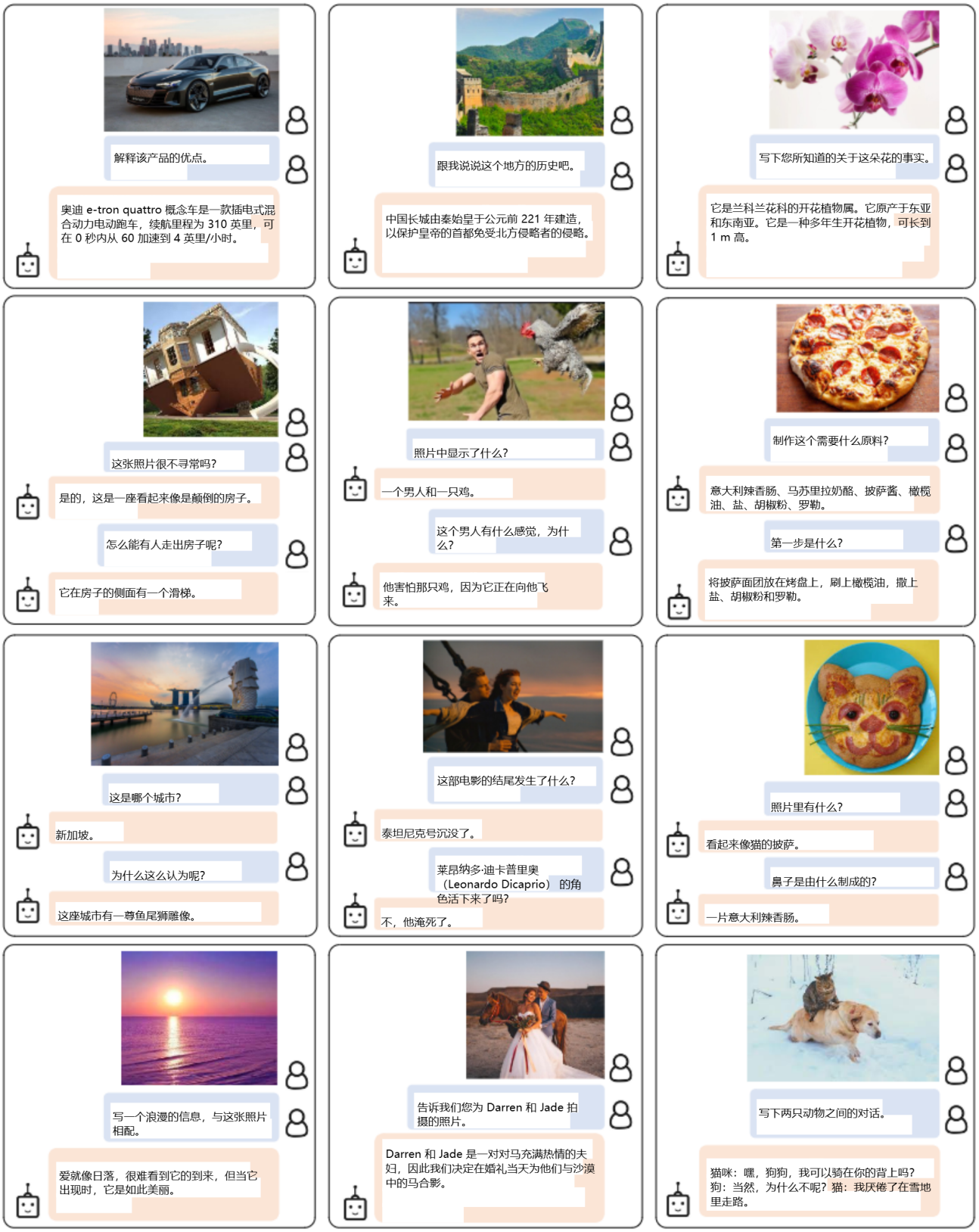

图 4.使用带有 ViT-g 和 FlanT5 的 BLIP-2 模型进行定向零镜头图像到文本生成的精选示例，其中它展示了广泛的功能，包括视觉对话、视觉知识推理、视觉共感推理、讲故事、个性化图像到文本生成等。

Models	#Trainable Params	Open-sourced?	Visual Question Answering VQAv2 (test-dev) VQA acc.	Image Captioning NoCaps (val) CIDEr	SPICE	Image-Text Retrieval Flickr (test) 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 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{\text{no-vqa}}$	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	44.4
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**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 commensense 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 FlanT5$_{XXL}$.

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-

模型	#Trainable 参数	打开-来源?	视觉问答 图像字幕 图像文本检索 VQAv2 (test-dev) NoCaps (val) Flickr (test) VQA acc. 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 火烈鸟 (Alayrac et al., 2022) 10.2B ✗ 56.3 - - - - -			
BLIP-2	188M ✓		65.0 121.6 15.8 97.6 89.7

表 1.各种零镜头视觉语言任务的 BLIP-2 结果概述。与以前的先进型号相比。BLIP-2 实现了最高的零镜头性能，同时在视觉语言预训练期间需要最少数量的可训练参数。

模型	#Trainable 参数	#Total 参数	VQAv2 OK-VQA GQA val test-dev test test-dev
VL-T5no-vqa 224M 269M 13.5 - 5.8 6.3 FewVLM（Jin 等人，2022 年） 740M 785M 47.7 - 16.5 29.3 冷冻（Tsimpoukelli 等人，2021 年） 40M 7.1B 29.6 - 5.9 VLKD（Dai 等人，2022 年） 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 OPT104M 3.1B 50.1 49.7 30.2 33.9 BLIP-2 ViT-g OPT107M 3.8B 53.5 52.3 31.7 34.6			
BLIP-2 ViT-g OPT108M 7.8B 54.3 52.6 36.4 36.4 BLIP-2 ViT-L FlanT5103M 3.4B 62.6 62.3 39.4 44.4 BLIP-2 ViT-g FlanT5107M 4.1B 63.1 63.0 40.7 44.2 BLIP-2 ViT-g FlanT5108M 12.1B 65.2 65.0 45.9 44.7			

表 2.与最先进的零镜头视觉问答方法的比较。

4. 实验

表 1 概述了 BLIP-2 在各种零镜头视觉语言任务上的性能。与以前最先进的模型相比，BLIP-2 实现了更高的性能，同时在视觉语言预训练期间需要的可训练参数数量大大减少。

4.1. 指示的零样本图像到文本生成

BLIP-2 有效地使 a LLM 能够理解图像，同时保留其在遵循文本提示方面的能力，这使我们能够通过指令控制图像到文本的生成。我们只需要在视觉提示之后附加文本提示作为输入到 LLM.图 4 显示了一些示例，展示了各种零样本图像到文本功能，包括视觉知识推理、视觉共感推理、视觉对话、个性化图像到文本生成等。

零散发 VQA。我们对零镜头视觉问答任务进行定量评估。对于 OPT 模型，我们使用提示 “Question： {} Answer：”。对于 FlanT5 模型，我们使用提示 “Question： {} Short answer：”。在生成过程中，我们使用光束宽度为 5 的光束搜索。我们还将 length-penalty 设置为 -1，这鼓励更短的答案，这些答案更符合人工注释。

如表 2 所示。BLIP-2在VQAv2（Goyal等人，2017）和GQA（Hudson & Manning，2019）数据集上取得了最先进的结果。它在 VQAv2 上的性能比 Flamingo80B 高出 8.7%，尽管可训练参数少了 54 倍。在 OK-VQA (Marino et al., 2019) 数据集上，BLIP-2 仅次于 Flamingo80B。我们假设这是因为 OK-VQA 更关注开放世界知识而不是视觉理解，而 Flamingo80B 的 70B Chinchilla（Hoffmann 等人，2022 年）语言模型比 11B FlanT5 拥有更多的知识。

我们从表 2 中得出了一个有希望的观察结果：更强 Image Encoder 或更强LLM的 Image Encoder 都会导致更好的 Per- 形式。这一观察结果得到了几个事实的支持：（1） ViT-g 在 OPT 和 FlanT5 上都优于 ViT-L。（2） 在同一LLM系列中，较大的模型优于较小的模型。（3） FlanT5 是一种指令调整 LLM的，在 VQA 上优于无监督训练的 OPT。这一观察结果验证了 BLIP-2 是一种通用的视觉语言预训练方法，可以有效地收获视觉和自然语言社区的快速发展。

视觉-语言表征学习的效果。

第一阶段的表示学习预先训练 QFormer 学习与文本相关的视觉特征，从而减轻了学习视觉-语言对齐的LLM负担。没有表征学习阶段，Q-

Models	#Trainable Params	NoCaps Zero-shot (validation set) in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Fine-tuned Karpathy test B@4	COCO Fine-tuned Karpathy test 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.

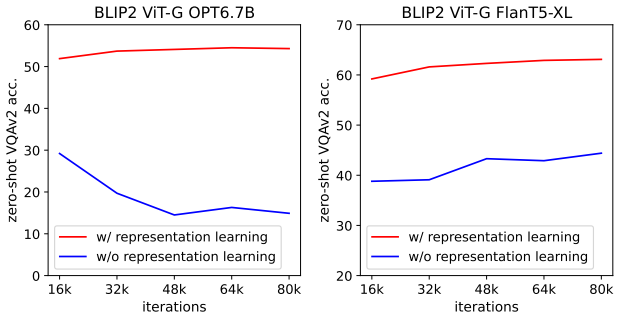

Figure 5. Effect of vision-language representation learning on vision-to-language generative learning. Without representation learning, the Q-Former fails the bridge the modality gap, leading to significantly lower performance on zero-shot VQA.

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

Models	#Trainable Params	VQAv2 test-dev	VQAv2 test-std
Open-ended generation models			
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**
Closed-ended classification models			
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.

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.

模型	#Trainable 参数	COCO 微调域内近域域外 整体 Karpathy 测试 C S C S C C S B@4 C
奥斯卡（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 - - - - 2021b）~1.4B 113.7 - 110.9 - 115.2 - 112.2 - 40.6 143.3		
BLIP-2 ViT-g OPT1.1B 123.0 15.8 117.8 15.4 123.4 15.1 119.7 15.4 43.7 145.8 BLIP-2 ViT-g OPT1.1B 123.7 15.8 119.2 15.3 124.4 14.8 121.0 15.3 43.5 145.2 BLIP-2 ViT-g FlanT51.1B 123.7 16.3 120.2 15.9 124.8 15.1 121.6 15.8 42.4 144.5		

表 3.与 NoCaps 和 COCO Caption 上最先进的图像字幕方法的比较。所有方法都优化了微调期间的交叉熵损失。C： CIDEr， S： 香料，B@4： BLEU@4.

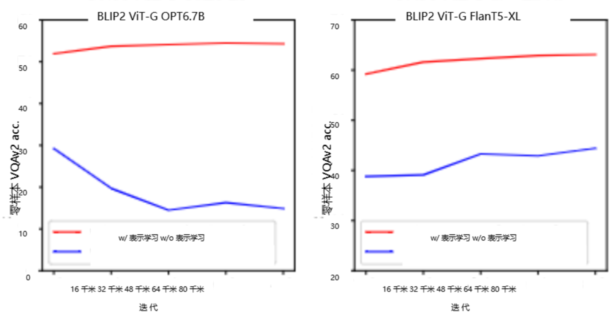

图 5.视觉-语言表征学习对视觉到语言生成学习的影响。如果没有表示学习，Q-Former 无法弥合模态差距，导致零样本 VQA 的性能显着降低。

模型	#Trainable 参数	VQAv2 版本 测试开发 test-std
开放式代系模型		
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 FlanT51.2B 81.55 81.66 BLIP-2 ViT-g OPT1.2B 81.59 81.74 BLIP-2 ViT-g OPT1.2B 82.19 82.30		
封闭式分类模型		
VinVL 345M 76.52 76.60 SimVLM（Wang 等人，2021b）~1.4B 80.03 80.34 CoCa（Yu 等人，2022）2.1B 82.30 82.30 BEIT-3（Wang 等人，2022b）1.9B 84.19 84.03		

表 4.与针对视觉问答进行微调的最先进的模型进行比较。

Former 完全依靠视觉到语言的生成学习来弥合模态差距，这类似于 Flamingo 中的 Perceiver Resampler。图 5 显示了表征学习对生成学习的影响。如果没有表示学习，这两种类型的 在 zero-shot VQA 上LLMs的性能都会大大降低。特别是，OPT 遭受了灾难性的遗忘，随着训练的进行，性能会急剧下降。

4.2. 图像字幕

我们为图像描述任务微调 BLIP-2 模型，该任务要求模型为图像的视觉内容生成文本描述。我们使用提示 “a photo of” 作为初始输入，LLM并训练模型生成带有语言建模损失的标题。我们在微调过程中保持LLM冻结状态，并与图像编码器一起更新 Q-Former 的参数。我们试验了 ViT-g 和各种 LLMs.详细的超参数可以在附录中找到。我们对 COCO 进行微调，并评估 COCO 测试集和零镜头转移到 NoCaps（Agrawal等人，2019 年） 验证集。

结果如表 3 所示。BLIP-2 达到 state-

与现有方法相比，NoCaps 具有最先进的性能，表现出对域外图像的强大泛化能力。

4.3. 视觉问答

给定带注释的 VQA 数据，我们在保持LLM冻结的同时微调 Q-Former 和图像编码器的参数。我们对开放式答案生成损失进行微调，其中LLM接收 Q-Former 的输出和问题作为输入，并被要求生成答案。为了提取与问题更相关的图像特征，我们还在问题上对 Q-Former 进行了条件设置。具体来说，问题标记作为 Q-Former 的输入，并通过自我注意层与查询交互，这可以引导 Q-Former 的交叉注意力层专注于信息量更大的图像区域。

在 BLIP 之后，我们的 VQA 数据包括 VQAv2 的训练和验证拆分，以及 Visual Genome 的训练样本。表 4 展示了 BLIP-2 在开放式生成模型中的最新结果。

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
Dual-encoder models													
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**
Fusion-encoder models													
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
Dual encoder + Fusion encoder reranking													
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	
	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.

型	#Trainable 参数	Flickr30K 零点（1K 测试集） 图像→文本 文本 → 图像			COCO 微调（5K 测试集） 图像→文本 文本 → 图像			图像 → 文本 文本 → 图像					
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
双编码器型号													
CLIP（Radford 等人，2021 年）	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	佛罗伦萨（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			
Fusion-encoder 模型													
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						
双编码器 + Fusion 编码器重新排序													
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

表 5.与最先进的图像文本检索方法进行比较，在 COCO 上进行了微调，并转移到 Flickr30K 的零镜头。

COCO 微调目标	图像→文本文本 → 图像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

表 6.图像基于的文本生成 （ITG） 损失通过强制查询来提取与语言相关的视觉特征，从而提高了图像文本检索性能。

4.4. 图像文本检索

由于图像文本检索不涉及语言生成，因此我们直接对第一阶段预训练模型进行微调，而无需 .LLM具体来说，我们使用与预训练相同的目标（即 ITC、ITM 和 ITG）在 COCO 上对图像编码器和 Q-Former 进行微调。然后，我们在 COCO 和 Flickr30K （Plummer et al.， 2015） 数据集上评估了该模型的图像到文本检索和文本到图像检索。在推理过程中，我们遵循 Li 等人（2021;2022 年）首先根据图像文本特征相似性选择 k = 128 个候选人，然后根据成对 ITM 分数重新排序。我们实验了 ViT-L 和 ViT-g 作为图像编码器。详细的超参数可以在附录中找到。

结果如表 5 所示。BLIP-2 实现了最先进的性能，与现有方法相比，在零样本图像文本检索方面有了显著改进。

ITC 和 ITM 损失对于图像文本检索至关重要，因为它们直接学习图像文本相似性。在表 6 中，我们表明 ITG（图像基于文本生成）损失也有利于图像文本检索。这个结果支持了我们设计表征学习目标的直觉： ITG 损失强制查询提取与文本最相关的视觉特征，从而改善视觉语言对齐。

5. 限制

Recent LLMs 可以在给定 fewshot 示例的情况下进行上下文学习。然而，在提供LLM上下文 VQA 示例时，我们对 BLIP-2 的实验并没有观察到 VQA 性能的改进。我们将缺乏上下文学习能力归因于我们的预训练数据集，每个样本仅包含一个图像-文本对。无法LLMs从中学习单个序列中多个图像-文本对之间的相关性。Flamingo 论文中也报告了相同的观察结果，该论文使用闭源交错图像和文本数据集 （M3W），每个序列具有多个图像-文本对。

我们的目标是在未来的工作中创建一个类似的数据集。

由于各种原因，BLIP-2 的图像到文本生成可能会产生不令人满意的结果LLM，包括来自不准确的知识、激活不正确的推理路径或没有关于新图像内容的最新信息（参见图 7）。此外，由于使用了冻结模型，BLIP-2 继承了 LLMs等风险，例如输出冒犯性语言、传播社会偏见或泄露私人信息。修复方法包括使用说明来指导模型在已删除有害内容的筛选数据集上生成或训练。

6. 总结

我们提出了 BLIP-2，这是一种用于视觉语言预训练的通用且计算高效的方法，它利用冻结的预训练图像编码器和 LLMs.BLIP-2 在各种视觉语言任务上实现了最先进的性能，同时在预训练期间具有少量的可训练参数。BLIP-2 还展示了零样本定向图像到文本生成的新兴功能。我们认为 BLIP-2 是构建多模式对话式 AI 代理的重要一步。

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., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V. Y., Huang, Y., Dai, A. M., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q., and Fung, P. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *ACL Findings*, pp. 2383–2395, 2022.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *NAACL*, pp. 4171–4186, 2019.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H. Unified language model pre-training for natural language understanding and generation. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *NeurIPS*, pp. 13042–13054, 2019.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6325–6334, 2017.

Guo, J., Li, J., Li, D., Tiong, A. M. H., Li, B., Tao, D., and Hoi, S. C. H. From images to textual prompts: Zero-shot VQA with frozen large language models. In *CVPR*, 2022.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

引用

Agrawal，H.，Anderson，P.，Desai，K.，Wang，Y.，Chen，X.，Jain，R.，Johnson，M.，Batra，D.，Parikh，D.，Lee，S.，和 Anderson，P. nocaps：大规模的新奇对象标题。在 ICCV 中，第 1 页。8947–8956, 2019.

阿莱拉克，J.，多纳休，J.，卢克，P.，米奇，A.，巴尔，I.，哈森，Y.，伦茨，K.，门施，A.，米利肯，K.，雷诺兹，M.，林，R.，卢瑟福，E.，卡比，S.，韩，T.，龚，Z.，萨曼古伊，S.，蒙泰罗，M.，梅尼克，J.，博尔格，S.，布洛克，A.，内马扎德，A.，沙里夫扎德，S.，宾科夫斯基，M.，巴雷拉，R.，文亚尔斯，O.，齐塞尔曼，A.和西蒙尼扬，K.火烈鸟：用于 Fewshot 学习的视觉语言模型。arXiv 预印本 arXiv：2204.14198,2022 年。

布朗，TB，曼恩，B.，莱德，N.，苏比亚，M.，卡普兰，J.，达里瓦尔，P.，尼拉坎坦，A.，夏姆，P.，萨斯特里，G.，阿斯凯尔，A.，阿加瓦尔，S.，赫伯特-沃斯，A.，克鲁格，G.，亨尼根，T.，柴尔德，R.，拉梅什，A.，齐格勒，DM，吴，J.，温特，C.，黑塞，C.，陈，M.，西格勒，E.，利特文，M.，格雷，S.，国际象棋，B.，克拉克，J.，伯纳，C.，McCandlish，S.，Radford，A.，Sutskever，I.，和 Amodei，D. 语言模型是 Few-shot 学习器。在 Larochelle，H.、Ranzato，M.、Hadsell，R.、Balcan，M. 和 Lin，H.（编辑）中，NeurIPS，2020 年。

Changpinyo，S.，Sharma，P.，Ding，N.，和 Soricut，R. 概念 12M：推动网络规模的图像文本预训练以识别长尾视觉概念。在 CVPR，2021 年。

Chen，J.，Guo，H.，Yi，K.，Li，B.，and Elhoseiny，M. Visualgpt：用于图像描述的预训练语言模型的数据高效适应。在 CVPR 中，第 18009-18019 页，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. 和 Soricut，R. Pali：联合缩放的多语言语言图像模型。arXiv 预印本 arXiv：2209.06794，2022b。

Chen，Y.，Li，L.，Yu，L.，Kholy，AE，Ahmed，F.，Gan，Z.，Cheng，Y.，and Liu，J. UNITER：通用图像文本表示学习。ECCV，第 12375 卷，第 1 页。104–120, 2020.

Cho，J.，Lei，J.，Tan，H.，and Bansal，M. 通过文本生成统一视觉和语言任务。arXiv 预印本 arXiv：2102.02779,2021 年。

钟，HW，侯，L.，隆普雷，S.，佐夫，B.，泰，Y.，费杜斯，W.，李，E.，王，X.，Dehghani，M.，梵

韦伯森，A.，顾，S.S.，戴，Z.，苏兹贡，M.，陈，X.，乔杜里，A.，纳朗，S.，米什拉，G.，于，A.，赵，V.Y.，黄，Y.，戴，A.M.，于，H.，彼得罗夫，S.，池，E.H.，迪恩，J.，德夫林，J.，罗伯茨，A.，周，D.，Le，Q.V.和 Wei，J.缩放指令微调语言模型。arXiv 预印本 arXiv：2210.11416,2022 年。

戴伟文，侯立文，尚立文，江，X.，刘琦，Q. 和 Fung，P. 通过视觉语言知识蒸馏在 CLIP 上实现多模态生成。在 Muresan，S.、Nakov，P. 和 Villavicencio，A.（编辑）中，ACL 调查结果，第 2383-2395 页，2022 年。

Devlin，J.，Chang，M.，Lee，K.，和 Toutanova，K. BERT：用于语言理解的深度双向转换器的预训练。在 Burstein，J.、Doran，C. 和 Solorio，T.（编辑）中，NAACL，第 4171-4186 页，2019 年。

董玲玲、杨南、王伟、魏芬、刘晓明、王耀、高杰、周明和韩浩 统一语言模型自然语言理解和生成的预训练。在 Wallach，HM、Larochelle，H.、Beygelzimer，A.、d'Alch 'e-Buc，F.、Fox，EB 和 Garnett，R.（编辑）中，NeurIPS，第 13042-13054 页，2019 年。

Fang，Y.，Wang，W.，Xie，B.，Sun，Q.，Wu，L.，Wang，X.，Huang，T.，Wang，X.，and Cao，Y. Eva：探索大规模掩蔽视觉表示学习的极限。arXiv 预印本 arXiv：2211.07636,2022 年。

Goyal，Y.，Khot，T.，Summers-Stay，D.，Batra，D.，和 Parikh，D. 让 VQA 中的 V 很重要：提升图像理解在视觉问答中的作用。在 CVPR，第 6325-6334 页，2017 年。

Guo，J.，Li，J.，Li，D.，Tiong，A. M. H.，Li，B.，Tao，D.，和 Hoi，S. C. H.从图像到文本提示：使用冻结的大型语言模型进行零样本 VQA。在 CVPR 中，2022 年。

Hoffmann，J.，Borgeaud，S.，Mensch，A.，Buchatskaya，E.，Cai，T.，Rutherford，E.，Casas，DDL，Hendricks，L. A.，Welbl，J.，Clark，A.，Hennigan，T.，Noland，E.，Millican，K.，Driessche，GVD，Damoc，B.，Guy，A.，Osindero，S.，Simonyan，K.，Elsen，E.，Rae，JW，Vinyals，O.，and Sifre，L. 训练计算最优大型语言模型。arXiv 预印本 arXiv：2203.15556,2022 年。

Hudson，DA 和 Manning，CDGQA：用于真实世界视觉推理和作文问答的新数据集。在 CVPR，第 6700-6709 页，2019 年。

Jia，C.，Yang，Y.，Xia，Y.，Chen，Y.-T.，Parekh，Z.，Pham，H.，Le，Q. V.，Sung，Y.，Li，Z.，and Duerig，T. 扩大视觉和视觉语言表征学习嘈杂的文本监督。arXiv 预印本 arXiv：2102.05918，2021.

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.

金文、程文、沈英、陈文和任 X。一个好的提示值数百万个参数：用于视觉语言模型的基于低资源提示的学习。在 Muresan， S.、Nakov， P. 和 Villavicencio， A.（编辑），ACL，第 2763-2775 页，2022 年。

克里希纳，R.，朱，Y.，格罗斯，O.，约翰逊，J.，哈塔，K.，克拉维茨，J.，陈，S.，卡兰蒂迪斯，Y.，李，L.，沙玛，

DA， Bernstein， MS 和 Fei-Fei， L. 视觉基因组：使用众包密集图像注释连接语言和视觉。IJCV，123（1）：32-73 2017 年

Li， J.， Selvaraju， RR， Gotmare， AD， Joty， S.， Xiong， C.， and Hoi， S. 融合前对齐：使用动量蒸馏进行视觉和语言表示学习。在 NeurIPS 中，2021 年。

Li， J.， Li， D.， Xiong， C.， and Hoi， SCH BLIP：用于统一视觉语言理解和生成的引导语言图像预训练。在 ICML 中，pp.

12888–12900, 2022.

李晓东， 尹晓东， 李俊杰， 张平， 胡晓明， 张丽玲， 王俊杰，

L.， 胡， H.， Dong， L.， Wei， F.， Choi， Y.， 和 Gao， J. Oscar： 视觉语言任务的对象语义对齐预训练。在 ECCV，第 121-137 页，2020 年

Lin， T.， Maire， M.， Belongie， SJ， Hays， J.， Perona， P.， Ramanan， D.， Doll 'ar， P.， and Zitnick， CL Microsoft COCO：上下文中的常见对象。在 Fleet， DJ、Pajdla， T.、Schiele， B. 和 Tuytelaars， T.（编辑）中，ECCV，第 8693 卷，第 740-755 页，2014 年。

Loshchilov， I. 和 Hutter， F. 解耦的权重衰减正则化。arXiv 预印本 arXiv：1711.05101,2017 年。

马 nas， O.， Rodr'ıguez， P.， Ahmadi， S.， Nematzadeh， A.， Goyal， Y.， 和 Agrawal， A. MAPL： 单峰预训练模型的参数高效适应视觉语言少数镜头提示。在 EACL，2023 年。

Marino， K.、Rastegari， M.、Farhadi， A. 和 Mottaghi， R. Okvqa：需要外部知识的视觉问答基准。在 CVPR，2019 年。

Ordonez， V.、Kulkarni， G. 和 Berg， TL Im2text：使用 100 万张带标题的照片描述图像。在

Shawe-Taylor， J.、Zemel， R. S.、Bartlett， P. L.、Pereira， F. C. N. 和 Weinberger， KQ（编辑），NIPS，第 1143-1151 页，2011 年

Plummer， BA， Wang， L.， 塞万提斯， CM， 凯塞多， JC， Hockenmaier， J. 和 Lazebnik， S. Flickr30k 实体：为更丰富的图像到句子模型收集区域到短语的对应关系。在 ICCV，第 2641-2649 页中

2015.

Radford， A.， Kim， JW， Hallacy， C.， Ramesh， A.， Goh， G.， Agarwal， S.， Sastry， G.， Askell， A.， Mishkin， P.， Clark， J.， et al 从自然中学习可转移语言监督。arXiv 预印本 arXiv：2103.00020,2021 年。

Schuhmann， C.， Vencu， R.， Beaumont， R.， Kaczmarczyk， R.， Mullis， C.， Katta， A.， Coombes， T.， Jitsev， J.， and Komatsuzaki， A. Laion-400m：剪辑过滤的 4 亿个图像文本对的开放数据集。arXiv 预印本 arXiv：2111.02114,2021 年。

Sharma， P.， Ding， N.， Goodman， S.， 和 Soricut， R. 概念字幕：用于自动图像字幕的清理、上位词状图像替代文本数据集。在 Gurevych， I.

和 Miyao， Y.（eds.），ACL，第 2556-2565 页，2018 年。

Tan， H. 和 Bansal， M. LXMERT：从变压器学习跨模态编码器表示。Inui， K.， 江， J.， Ng， V.， 和 Wan， X.（编辑），EMNLP，第 5099-5110 页，2019 年。

Tiong， A. M. H.， Li， J.， Li， B.， Savarese， S. 和 Hoi， S.

CH 即插即用 VQA：通过将大型预训练模型与零训练相结合来实现零样本 VQA。在 EMNLP 调查结果中，2022 年

Tsimpoukelli， M.， Menick， J.， Cabi， S.， Eslami， S. M. A.， Vinyals， O.， 和 Hill， F. 多模态小样本学习使用冻结的语言模型。在 Ranzato， M.、Beygelzimer， A.、Dauphin， YN、Liang， P. 和 Vaughan， JW 中。（编辑），NeurIPS，第 200-212 页，2021 年。

Wang， P.， Yang， A.， Men， R.， Lin， J.， Bai， S.， Li， Z.， 马， J.， 周， C.， 周， J.， and Yang， H. OFA： 通过简单的序列到序列学习框架统一架构、任务和模式。在 Chaudhuri， K.、Jegelka， S.、Song， L.、Szepesv 'ari， C.、Niu， G. 和 Sabato， S.（编辑）中，ICML，第 23318-23340 页，2022a。

Wang， W.， Bao， H.， Dong， L.， 和 Wei， F. Vlmo：与模态混合专家进行统一视觉语言预训练。arXiv 预印本 arXiv：2111.02358,2021a。

Wang， W.， Bao， H.， Dong， L.， Bjorck， J.， Peng， Z.， Liu， Q.， Aggarwal， K.， Mohammed， OK， Singhal， S.， Som， S.， and Wei， F. 图像作为外语：所有视觉和视觉语言任务的 Beit 预训练。arXiv 预印本 arXiv：2208.10442， 2022b.

Wang， Z.， Yu， J.， Yu， AW， Dai， Z.， Tsvetkov， Y.， and Cao， Y. Simvlm： 简单的视觉语言模型预训练

监督薄弱。arXiv 预印本 arXiv：2108.10904,2021b。

姚丽玲， 黄玲玲， 侯玲玲， 卢玲玲， 牛玲玲， 徐晓玲， 梁晓玲， 李玲玲， 江晓晓， 徐玲： 细粒度交互式语言图像预训练 在 ICLR，2022 年

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.

Yu，J.，Wang，Z.，Vasudevan，V.，Yeung，L.，Seyedhosseini，M.，和 Wu，Y. Coca：对比字幕是图像文本基础模型。arXiv 预印本 arXiv：2205.01917,2022 年。

袁林，陈大，陈，陈，科德拉，N.，戴，X.，高，J.，胡，H.，黄，X.，李 B.，李，C.，刘 C.，刘 M.，刘 Z.，卢 Y.，石 Y.，王 L.，王 J.，肖 B.，肖 Z.，杨 J.，曾 M.，周 L.，和张 P.

Florence：计算机视觉的新基础模型。arXiv 预印本 arXiv：2111.11432,2021 年。

Zhai，X.，Wang，X.，Mustafa，B.，Steiner，A.，Keysers，D.，Kolesnikov，A.，and Beyer，L. Lit：使用锁定图像文本调整的零镜头传输。在 CVPR，第 18102-18112 页，2022 年。

Zhang，P.，Li，X.，胡，X.，Yang，J.，Zhang，L.，Wang，L.，Choi，Y.，and Gao，J. Vinvl：使视觉表示在视觉语言模型中变得重要。arXiv 预印本 arXiv：2101.00529,2021 年。

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.，and Zettlemoyer，L. OPT：开放式预训练转换器语言模型。

arXiv 预印本 arXiv：2205.01068,2022 年。

LLM	FlanT5$_{XL}$	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	FlanT5$_{XL}$	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 FlanT5$_{XXL}$.

LLM	FlanT5OPTOPT
微调时期 5 预热步骤 1000 学习率 1e-5 批量大小 256 AdamW β （0.9,0.999） 权重衰减 0.05 放置路径 0 图像分辨率 364 提示 “照片” 推理光束大小 5 ViT 的逐层学习率衰减 1	
1	0.95

表 7.用于在 COCO 字幕上使用 ViT-g 微调 BLIP-2 的超参数。

LLM	FlanT5OPTOPT
微调时期 5 预热步骤 1000 学习率 1e-5 批量大小 128 AdamW β （0.9,0.999） 权重衰减 0.05 放置路径 0 图像分辨率 490 提示 “问题： {} 答案：” 推理光束大小 5 ViT 的逐层学习率衰减 0.95 0.95 0.9 0.9	

表 8.用于在 VQA 上使用 ViT-g 微调 BLIP-2 的超参数。

图像编码器	ViT-L/14 ViT-g/14
微调时期 5 预热步骤 1000 学习率 5e-6 1e-5 批量大小 224 AdamW β （0.9,0.98） （0.9,0.999） 权重衰减 0.05 放置路径 0 图像分辨率 364 ViT 的逐层学习率衰减 1 0.95	

表 9.用于在 COCO 图像文本检索上微调 BLIP-2 的超参数。

图 6.使用 BLIP-2 模型（含 ViT-g 和 FlanT5）生成指示的零样本图像到文本的输出示例不正确。

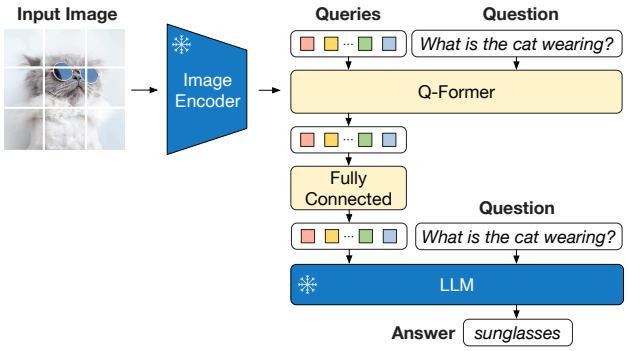

Figure 7. Model architecture for VQA finetuning, where the LLM receives Q-Former's output and the question as input, then predicts answers. We also provide the question as a condition to Q-Former, such that the extracted image features are more relevant to the question.

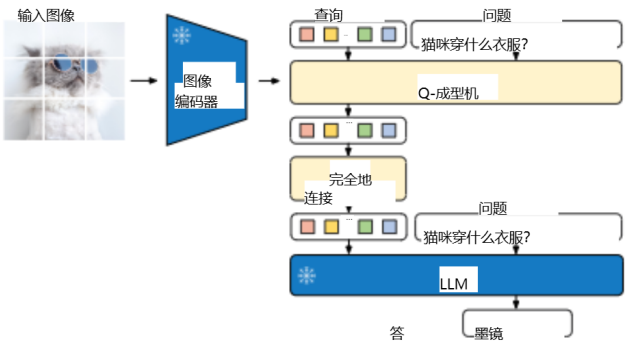

图 7.VQA 微调的模型架构，其中LLM接收 Q-Former 的输出和问题作为输入，然后预测答案。我们还将问题作为条件提供给 Q-Former，以便提取的图像特征与问题更相关。