InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning

Wenliang Dai[†1,2*] **Junnan Li**[†,✉,1] **Dongxu Li**[1] **Anthony Meng Huat Tiong**[1,3]
Junqi Zhao[3] **Weisheng Wang**[3] **Boyang Li**[3] **Pascale Fung**[2] **Steven Hoi**[✉,1]
[1]Salesforce Research [2]Hong Kong University of Science and Technology
[3]Nanyang Technological University, Singapore
https://github.com/salesforce/LAVIS/tree/main/projects/instructblip
[†]Equal contribution [✉]Corresponding authors: {junnan.li,shoi@salesforce.com}

Abstract

Large-scale pre-training and instruction tuning have been successful at creating general-purpose language models with broad competence. However, building general-purpose vision-language models is challenging due to the rich input distributions and task diversity resulting from the additional visual input. Although vision-language pretraining has been widely studied, vision-language instruction tuning remains under-explored. In this paper, we conduct a systematic and comprehensive study on vision-language instruction tuning based on the pretrained BLIP-2 models. We gather 26 publicly available datasets, covering a wide variety of tasks and capabilities, and transform them into instruction tuning format. Additionally, we introduce an instruction-aware Query Transformer, which extracts informative features tailored to the given instruction. Trained on 13 held-in datasets, InstructBLIP attains state-of-the-art zero-shot performance across all 13 held-out datasets, substantially outperforming BLIP-2 and larger Flamingo models. Our models also lead to state-of-the-art performance when finetuned on individual downstream tasks (e.g., 90.7% accuracy on ScienceQA questions with image contexts). Furthermore, we qualitatively demonstrate the advantages of InstructBLIP over concurrent multimodal models. All InstructBLIP models are open-sourced.

1 Introduction

A longstanding aspiration of Artificial Intelligence (AI) research is to build a single model that can solve arbitrary tasks specified by the user. In natural language processing (NLP), instruction tuning [46, 7] proves to be a promising approach toward that goal. By finetuning a large language model (LLM) on a wide range of tasks described by natural language instructions, instruction tuning enables the model to follow arbitrary instructions. Recently, instruction-tuned LLMs have also been leveraged for vision-language tasks. For example, BLIP-2 [20] effectively adapts frozen instruction-tuned LLMs to understand visual inputs and exhibits preliminary capabilities to follow instructions in image-to-text generation.

Compared to NLP tasks, vision-language tasks are more diverse in nature due to the additional visual inputs from various domains. This poses a greater challenge to a unified model that is supposed to generalize to diverse vision-language tasks, many unseen during training. Most previous work can be grouped into two approaches. The first approach, multitask learning [6, 27], formulates various vision-language tasks into the same input-output format. However, we empirically find multitask learning without instructions (Table 4) does not generalize well to unseen datasets and tasks. The

[*]Work done during internship at Salesforce.

Preprint. Under review.

InstructBLIP：通过指令调优实现通用视觉语言模型

戴文亮 李俊楠 张俊琪 赵伟生 王博阳 [B] 李佩斯卡尔 冯Steven Hoi B

[1] Salesforce Research 香港科技大学
[3] 新加坡南洋理工大学
https://github.com/salesforce/LAVIS/tree/main/projects/instructblip
[†] 平等贡献 [B] 通讯作者： {junnan.li,shoi@salesforce.com}

抽象

大规模的预训练和指令调整已经成功地创建了具有广泛能力的通用语言模型。然而，构建通用视觉语言模型具有挑战性，因为额外的视觉输入会导致丰富的输入分布和任务多样性。尽管视觉语言预训练已被广泛研究，但视觉语言教学调整仍未得到充分探索。在本文中，我们基于预训练的 BLIP-2 模型对视觉语言教学调整进行了系统而全面的研究。我们收集了 26 个公开可用的数据集，涵盖各种任务和功能，并将它们转换为指令调优格式。此外，我们还引入了一个指令感知 Query Transformer，它可以提取针对给定指令量身定制的信息特征。InstructBLIP 在 13 个保留数据集上进行了训练，在所有 13 个保留数据集中都获得了最先进的零镜头性能，大大优于 BLIP-2 和更大的 Flamingo 模型。当对单个下游任务进行微调时，我们的模型还可以带来最先进的性能（例如，在图像上下文中的 ScienceQA 问题上，准确率为 90.7%）。此外，我们定性地证明了 InstructBLIP 相对于并发多模态模型的优势。所有 InstructBLIP 模型都是开源的。

1 介绍

人工智能 （AI） 研究的一个长期愿望是构建一个可以解决用户指定的任意任务的单一模型。在自然语言处理 （NLP） 中，指令调优 [46， 7] 被证明是实现这一目标的一种很有前途的方法。通过对自然语言指令描述的各种任务对大型语言模型 （）LLM 进行微调，指令优化使模型能够遵循任意指令。最近，指令调整LLMs也被用于视觉语言任务。例如，BLIP-2 [20] 有效地调整了 frozen instructiontuned LLMs 以理解视觉输入，并展示了在图像到文本生成中遵循指令的初步能力。

与 NLP 任务相比，由于来自各个领域的额外视觉输入，视觉语言任务在本质上更加多样化。这对一个统一的模型构成了更大的挑战，该模型应该推广到不同的视觉语言任务，其中许多任务在训练过程中是看不见的。以前的大多数工作可以分为两种方法。第一种方法，多任务学习 [6， 27]，将各种视觉语言任务表述为相同的输入-输出格式。然而，我们凭实证发现，没有指令的多任务学习（表 4）并不能很好地推广到看不见的数据集和任务。这

[*] 在 Salesforce 实习期间完成的工作。

预印本。审核中。

What could have happened based on the current scene?

Based on the current scene in the image, it is possible that a hurricane or severe weather event caused significant damage to the buildings and infrastructure in the area. The person walking through the debris-covered area suggests that they are inspecting the damage caused by the disaster. The presence of palm trees indicates that the location may be in a tropical or sub-tropical region, which can be more susceptible to hurricanes and other natural disasters.

Can yon tell me about this image in detail?

The image depicts a collection of various vegetables including carrots, cucumbers, tomatoes, and nuts arranged on a stone surface. There are several jars filled with different types of ingredients, such as peanuts, cashews, sunflower seeds, and pumpkin seeds. These ingredients are likely to be part of a healthy meal or snack.

Can you give me an example of such meal?

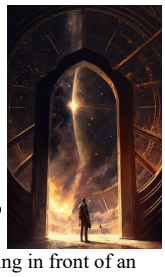
An example of a healthy meal or snack that incorporates some of the ingredients in the image could be a salad made with mixed greens, carrots, cucumbers, tomatoes, and nuts. The salad can be dressed with a homemade vinaigrette made with olive oil, lemon juice, salt, and pepper. Additionally, the salad can be topped with pumpkin seeds or sunflower seeds for added crunch and nutritional value.

Sounds good! Show me steps of making this salad.

Sure! Here are the steps to make this healthy salad:
1. Wash and chop the mixed greens, carrots, cucumbers, and tomatoes into bite-sized pieces.
2. In a small bowl, whisk together olive oil, lemon juice, salt, and pepper to create a homemade vinaigrette dressing.
3. Place the chopped vegetables in a large bowl and toss them with the vinaigrette dressing.
4. Sprinkle pumpkin seeds or sunflower seeds on top of the salad for added crunch and nutritional value.
5. Serve the salad immediately or store it in the refrigerator until ready to enjoy.

Introduce me this painting in detail.

The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.



Can you describe this image in detail?

The image depicts a person standing in front of an open doorway, looking out into space. They are wearing a jacket and appears to be contemplating the vastness of the universe beyond the doorway. In the background, there are multiple stars visible, creating a stunning celestial scene. The person's gaze is directed towards the center of the image, where a bright star can be seen shining brightly. There are also smaller stars scattered throughout the scene, adding to the overall cosmic atmosphere.

What if they jump over the doorway?

If the person were to jump over the doorway, they would be leaving the confines of their current location and venturing into the vastness of space. This could be a metaphorical representation of stepping out of one's comfort zone or embarking on a new adventure. However, it is important to note that such an action would be extremely dangerous and potentially life-threatening due to the harsh conditions of space.

What makes this dog special?

The dog in the image is wearing futuristic armor and has a glowing green eye, giving it a unique and futuristic appearance. This combination of elements makes the dog stand out from other dogs and adds an element of fantasy or sci-fi to the image. Additionally, the dog's posture and facial expression convey a sense of confidence and determination, further emphasizing its unique features.

Figure 1: A few qualitative examples generated by our InstructBLIP Vicuna model. Here, a range of its diverse capabilities are demonstrated, including complex visual scene understanding and reasoning, knowledge-grounded image description, multi-turn visual conversation, etc.

图 1：我们的 InstructBLIP Vicuna 模型生成的一些定性示例。在这里，展示了其一系列多样化的能力，包括复杂的视觉场景理解和推理、基于知识的图像描述、多回合视觉对话等。

second approach [20, 4] extends a pre-trained LLM with additional visual components, and trains the visual components with image caption data. Nevertheless, such data are too limited to allow broad generalization to vision-language tasks that require more than visual descriptions.

To address the aforementioned challenges, this paper presents InstructBLIP, a vision-language instruction tuning framework that enables general-purpose models to solve a wide range of visual-language tasks through a unified natural language interface. InstructBLIP uses a diverse set of instruction data to train a multimodal LLM. Specifically, we initialize training with a pre-trained BLIP-2 model consisting of an image encoder, an LLM, and a Query Transformer (Q-Former) to bridge the two. During instruction tuning, we finetune the Q-Former while keeping the image encoder and LLM frozen. Our paper makes the following key contributions:

- We perform a comprehensive and systematic study on vision-language instruction tuning. We transform 26 datasets into the instruction tuning format and group them into 11 task categories. We use 13 held-in datasets for instruction tuning and 13 held-out datasets for zero-shot evaluation. Moreover, we withhold four entire task categories for zero-shot evaluation at the task level. Exhaustive quantitative and qualitative results demonstrate the effectiveness of InstructBLIP on vision-language zero-shot generalization.
- We propose instruction-aware visual feature extraction, a novel mechanism that enables flexible and informative feature extraction according to the given instructions. Specifically, the textual instruction is given not only to the frozen LLM, but also to the Q-Former, so that it can extract instruction-aware visual features from the frozen image encoder. Also, we propose a balanced sampling strategy to synchronize learning progress across datasets.
- We evaluate and open-source a suite of InstructBLIP models using two families of LLMs: 1) FlanT5 [7], an encoder-decoder LLM finetuned from T5 [34]; 2) Vicuna [2], a decoder-only LLM finetuned from LLaMA [41]. The InstructBLIP models achieve state-of-the-art zero-shot performance on a wide range of vision-language tasks. Furthermore, InstructBLIP models lead to state-of-the-art finetuning performance when used as the model initialization on individual downstream tasks.

2 Vision-Language Instruction Tuning

InstructBLIP aims to address the unique challenges in vision-language instruction tuning and provide a systematic study on the models' improved generalization ability to unseen data and tasks. In this section, we first introduce the construction of instruction-tuning data, followed by the training and evaluation protocols. Next, we delineate two techniques to improve instruction-tuning performance from the model and data perspectives, respectively. Lastly, we present the implementation details.

2.1 Tasks and Datasets

To ensure the diversity of instruction tuning data while considering their accessibility, we gather comprehensive set of publicly available vision-language datasets, and transform them into the instruction tuning format. As shown in Figure 2, the final collection covers 11 task categories and 26 datasets, including image captioning [23, 3, 51], image captioning with reading comprehension [38], visual reasoning [16, 24, 29], image question answering [11, 12], knowledge-grounded image question answering [30, 36, 28], image question answering with reading comprehension [31, 39], image question generation (adapted from the QA datasets), video question answering [47, 49], visual conversational question answering [8], image classification [18], and LLaVA-Instruct-150K [25]. We include detailed descriptions and statistics of each dataset in Appendix C.

For every task, we meticulously craft 10 to 15 distinct instruction templates in natural language. These templates serve as the foundation for constructing instruction tuning data, which articulates the task and the objective. For public datasets inherently favoring short responses, we use terms such as *short* and *briefly* into some of their corresponding instruction templates to reduce the risk of the model overfitting to always generating short outputs. For the LLaVA-Instruct-150K dataset, we do not incorporate additional instruction templates since it is naturally structured in the instruction format. The full list of instruction templates can be found in Appendix D.

第二种方法 [20 ， 4] 使用额外的视觉组件扩展了预训练LLM，并使用图像标题数据训练视觉组件。然而，这些数据太有限了，无法广泛推广到需要视觉描述以外的视觉语言任务。

为了解决上述挑战，本文提出了 InstructBLIP，这是一个视觉语言指令调优框架，它使通用模型能够通过统一的自然语言界面解决各种视觉语言任务。InstructBLIP 使用一组不同的指令数据来训练多模态 LLM。具体来说，我们使用一个预先训练的 BLIP-2 模型来初始化训练，该模型由一个图像编码器、一个LLM和一个 Query Transformer （Q-Former） 组成，以桥接两者。在指令调整期间，我们在保持图像编码器和LLM冻结状态的同时对 Q-Former 进行微调。我们的论文做出了以下主要贡献：

- 我们对视觉-语言教学调整进行了全面而系统的研究。我们将 26 个数据集转换为指令调整格式，并将它们分为 11 个任务类别。我们使用 13 个保留数据集进行指令调整，使用 13 个保留数据集进行零样本评估。此外，我们保留了四个完整的任务类别，以便在任务级别进行零镜头评估。详尽的定量和定性结果表明 InstructBLIP 在视觉语言零镜头泛化方面的有效性。

- 我们提出了指令感知视觉特征提取，这是一种新颖的机制，可以根据给定的指令进行灵活和信息丰富的特征提取。具体来说，文本指令不仅提供给 冻结 LLM，还提供给 Q-Former，以便它可以从冻结图像编码器中提取指令感知视觉特征。此外，我们还提出了一种平衡的采样策略来同步跨数据集的学习进度。

- 我们使用两个系列评估并开源了一套 InstructBLIP 模型LLMs：1） FlanT5 [7]，一个从 T5 [34] 微调而来的编码器-解码器LLM;2） Vicuna [2]，一种仅由 LLaMA [41] 微调而成的解码LLM器。InstructBLIP 模型在各种视觉语言任务中实现了最先进的零镜头性能。此外，InstructBLIP 模型在用作单个下游任务的模型初始化时，可实现最先进的微调性能。

2 视觉语言教学调优

InstructBLIP 旨在解决视觉语言教学调整中的独特挑战，并对模型对看不见的数据和任务的泛化能力的改进进行系统研究。在本节中，我们首先介绍指令调整数据的构造，然后是训练和评估协议。接下来，我们分别从模型和数据角度描述了两种提高指令调优性能的技术。最后，我们介绍实现细节。

2.1 任务和数据集

为了确保指令调优数据的多样性，同时考虑其可访问性，我们收集了一整套公开可用的视觉语言数据集，并将它们转换为指令调优格式。如图 2 所示，最终集合涵盖了 11 个任务类别和 26 个数据集，包括图像字幕 [23 ， 3 ， 51]、图像字幕与阅读理解 [38]、视觉推理 [16 ， 24 ， 29]、图像问答 [11 ， 12]、基于知识的图像问答 [30 ， 36 ， 28]、图像问答与阅读理解 [31 ， 39]、图像问题生成（改编自 QA 数据集）、 视频问答 [47 ， 49]、视觉对话问答 [8]、图像分类 [18] 和 LLaVA-Instruct-150K [25]。我们在附录 C 中包含了每个数据集的详细说明和统计数据。

对于每项任务，我们都会用自然语言精心制作 10 到 15 个不同的教学模板。这些模板是构建指令调优数据的基础，这些数据阐明了任务和目标。对于本质上偏爱短响应的公共数据集，我们在它们的一些相应指令模板中使用了 short 和 brief 等术语，以降低模型过拟合的风险，从而始终生成短输出。对于 LLaVA-Instruct-150K 数据集，我们没有合并额外的指令模板，因为它自然地以指令格式构建。指令模板的完整列表可在附录 D 中找到。

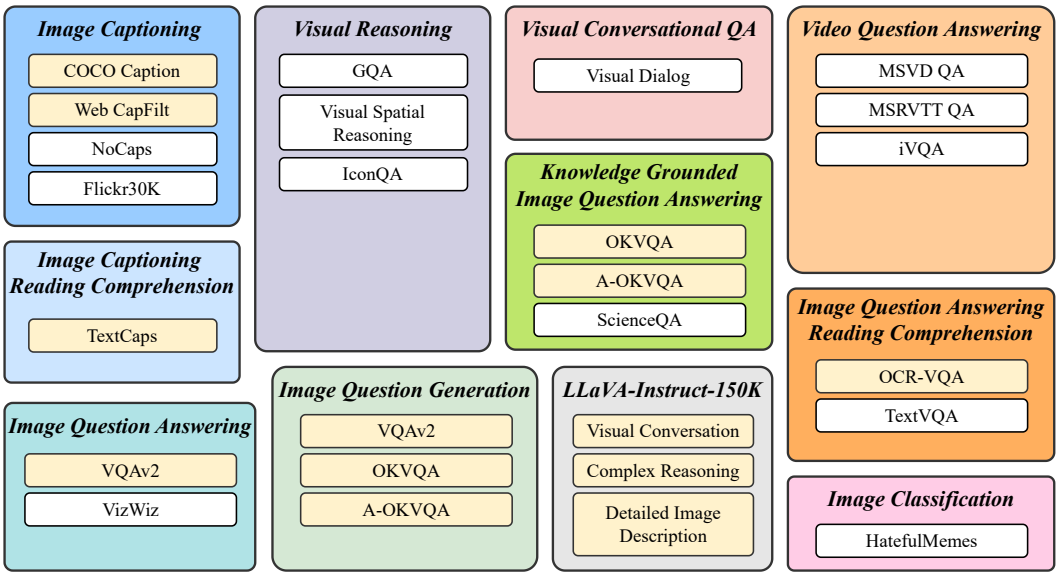


Figure 2: Tasks and their corresponding datasets used for vision-language instruction tuning. The held-in datasets are indicated by yellow and the held-out datasets by white.

2.2 Training and Evaluation Protocols

To ensure sufficient data and tasks for training and zero-shot evaluation, we divide the 26 datasets into 13 held-in datasets and 13 held-out datasets, indicated by yellow and white respectively in Figure 2. We employ the training sets of the held-in datasets for instruction tuning and their validation or test sets for held-in evaluation.

For held-out evaluation, our aim is to understand how instruction tuning improves the model's zero-shot performance on unseen data. We define two types of held-out data: 1) datasets not exposed to the model during training, but whose tasks are present in the held-in cluster; 2) datasets and their associated tasks that remain entirely unseen during training. Addressing the first type of held-out evaluation is nontrivial due to the data distribution shift between held-in and held-out datasets. For the second type, we hold out several tasks completely, including visual reasoning, video question answering, visual conversational QA, and image classification.

To avoid data contamination, datasets are selected carefully so that no evaluation data appear in the held-in training cluster across different datasets. During instruction tuning, we mix all the held-in training sets and sample instruction templates uniformly for each dataset. The models are trained with the standard language modeling loss to directly generate the response given the instruction. Furthermore, for datasets that involve scene texts, we add OCR tokens in the instruction as supplementary information.

2.3 Instruction-aware Visual Feature Extraction

Existing zero-shot image-to-text generation methods, including BLIP-2, take an instruction-agnostic approach when extracting visual features. That results in a set of static visual representations being fed into the LLM, regardless of the task. In contrast, an instruction-aware vision model can adapt to the task instruction and produce visual representations most conducive to the task at hand. This is clearly advantageous if we expect the task instructions to vary considerably for the same input image.

We show the architecture of InstructBLIP in Figure 3. Similarly to BLIP-2 [20], InstructBLIP utilizes a Query Transformer, or Q-Former, to extract visual features from a frozen image encoder. The input to the Q-Former contains a set of K learnable query embeddings, which interact with the image encoder's output through cross attention. The output of the Q-Former consists of K encoded visual vectors, one per query embedding, which then go through a linear projection and are fed to the frozen LLM. As in BLIP-2, the Q-Former is pretrained in two stages using image-caption data

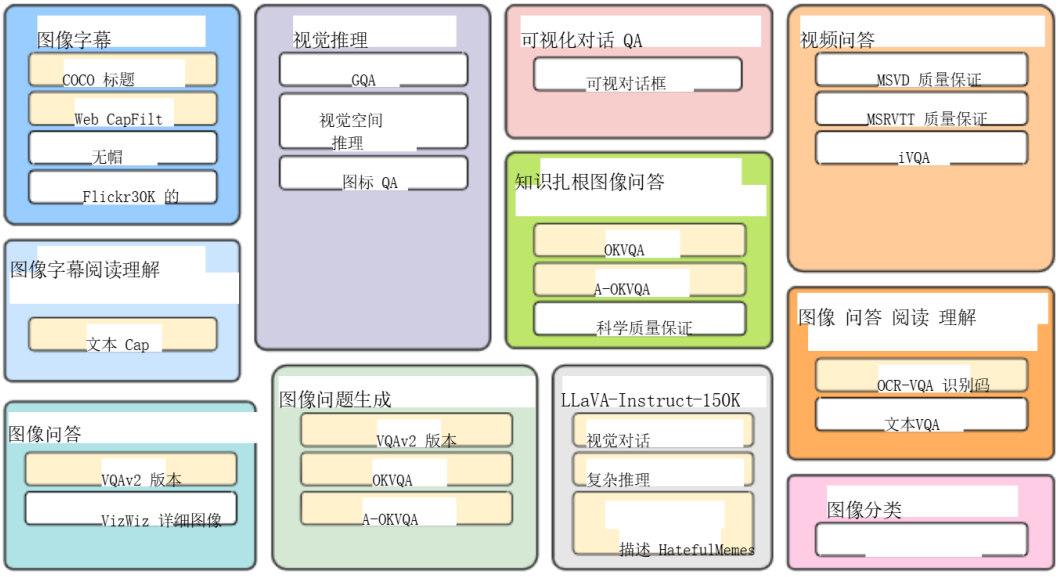


图 2：用于视觉语言教学调整的任务及其相应的数据集。保留的数据集用黄色表示，保留的数据集用白色表示。

2.2 训练和评估方案

为了确保有足够的数据和任务进行训练和零镜头评估，我们将 26 个数据集分为 13 个保留数据集和 13 个保留数据集，在图 2 中分别用黄色和白色表示。我们使用 hold-in 数据集的训练集进行指令调整，并使用它们的验证或测试集进行 hold-in 评估。

对于保留评估，我们的目标是了解指令调优如何提高模型对看不见的数据的零镜头性能。我们定义了两种类型的保留数据：1） 在训练期间未暴露给模型，但其任务存在于保留集群中的数据集;2） 在训练过程中完全不可见的数据集及其相关任务。解决第一种类型的保留评估并非易事，因为保留数据集和保留数据集之间的数据分布发生了变化。对于第二种类型，我们完全承担了几项任务，包括视觉推理、视频问答、视觉对话 QA 和图像分类。

为避免数据污染，我们会仔细选择数据集，以便不会在跨不同数据集的保留训练集群中出现评估数据。在指令调整期间，我们为每个数据集统一混合所有保留的训练集和示例指令模板。这些模型使用标准语言建模 loss 进行训练，以直接生成给定指令的响应。此外，对于涉及场景文本的数据集，我们在指令中添加 OCR 标记作为补充信息。

2.3 指令感知视觉特征提取

现有的零样本图像到文本生成方法（包括 BLIP-2）在提取视觉特征时采用与指令无关的方法。这会导致一组静态视觉表示被馈送到 LLM中，而不管任务是什么。相比之下，指令感知视觉模型可以适应任务指令并产生最有利于手头任务的视觉表示。如果我们预计同一输入图像的任务指令会有很大差异，这显然是有利的。

我们在图 3 中展示了 InstructBLIP 的架构。与 BLIP-2 [20] 类似，InstructBLIP 利用查询转换器或 Q-Former 从冻结图像编码器中提取视觉特征。Q-Former 的输入包含一组 K 个可学习的查询嵌入，这些嵌入通过交叉注意与图像编码器的输出进行交互。Q-Former 的输出由 K 个编码的视觉向量组成，每个查询嵌入一个，然后通过线性投影并馈送到 frozen LLM.与 BLIP-2 一样，Q-Former 使用图像标题数据分两个阶段进行预训练

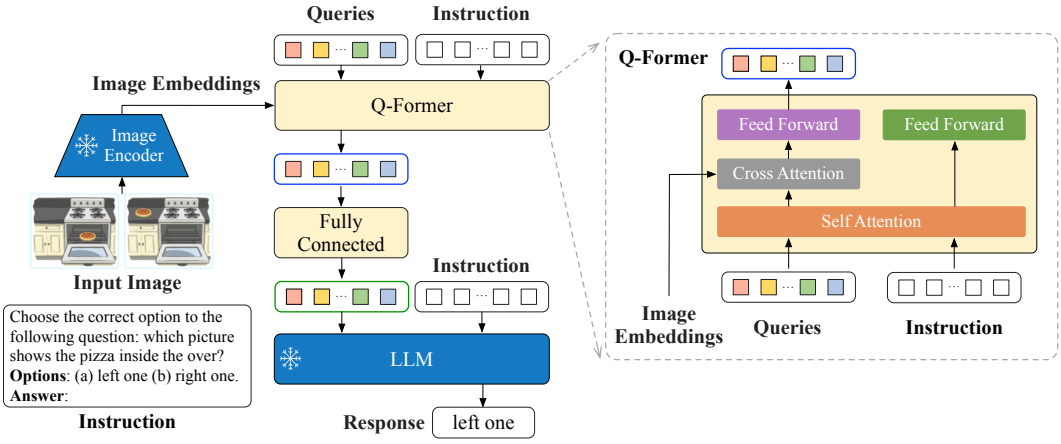


Figure 3: Model architecture of InstructBLIP. The Q-Former extracts instruction-aware visual features from the output embeddings of the frozen image encoder, and feeds the visual features as soft prompt input to the frozen LLM. We instruction-tune the model with the language modeling loss to generate the response.

before instruction tuning. The first stage pretrains the Q-Former with the frozen image encoder for vision-language representation learning. The second stage adapts the output of Q-Former as soft visual prompts for text generation with a frozen LLM . After pretraining, we finetune the Q-Former with instruction tuning, where the LLM receives as input the visual encodings from the Q-Former and the task instruction.

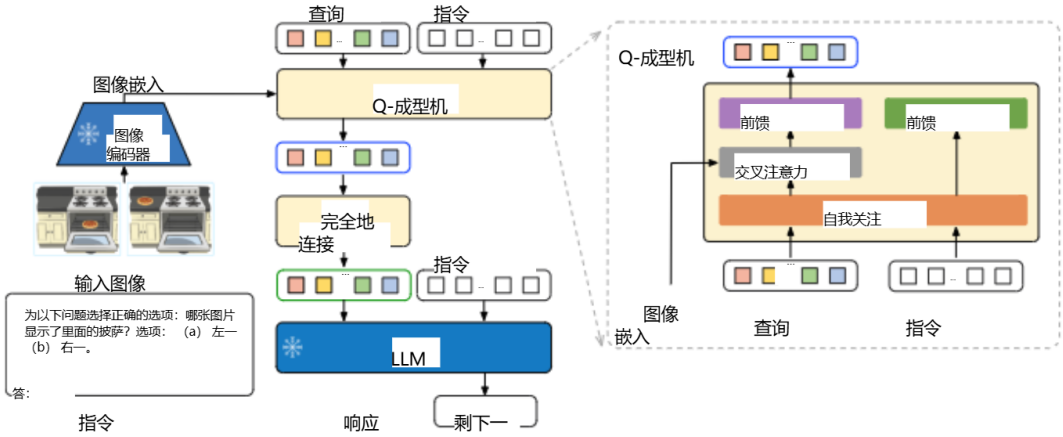
Extending BLIP-2, InstructBLIP proposes an instruction-aware Q-former module, which takes in the instruction text tokens as additional input. The instruction interacts with the query embeddings through self-attention layers of the Q-Former, and encourages the extraction of task-relevant image features. As a result, the LLM receives visual information conducive to instruction following. We demonstrate empirically (Table 2) that instruction-aware visual feature extraction provides substantial performance improvements for both held-in and held-out evaluations.

2.4 Balancing Training Datasets

Due to the large number of training datasets and the significant differences in the size of each dataset, mixing them uniformly could cause the model to overfit smaller datasets and underfit larger datasets. To mitigate the problem, we propose to sample datasets with probabilities proportional to the square root of their sizes, or the numbers of training samples. Generally, given D datasets with sizes $\{S_1, S_2, \ldots, S_D\}$, the probability of a data sample being selected from a dataset d during training is $p_d = \frac{\sqrt{S_d}}{\sum_{i=1}^{D} \sqrt{S_i}}$. On top of this formula, we make manual adjustments to the weights of certain datasets to improve optimization. This is warranted by inherent differences in the datasets and tasks that require varying levels of training intensity despite similar sizes. To be specific, we lower the weight of A-OKVQA, which features multiple-choice questions, and increase the weight of OKVQA, which requires open-ended text generation. In Table 2, we show that the balanced dataset sampling strategy improves overall performance for both held-in evaluation and held-out generalization.

2.5 Inference Methods

During inference time, we adopt two slightly different generation approaches for evaluation on different datasets. For the majority of datasets, such as image captioning and open-ended VQA, the instruction-tuned model is directly prompted to generate responses, which are subsequently compared to the ground truth to calculate metrics. On the other hand, for classification and multi-choice VQA tasks, we employ a vocabulary ranking method following previous works [46, 22, 21]. Specifically, we still prompt the model to generate answers, but restrict its vocabulary to a list of candidates. Then, we calculate log-likelihood for each candidate and select the one with the highest value as the final prediction. This ranking method is applied to ScienceQA, IconQA, A-OKVQA (multiple-choice), HatefulMemes, Visual Dialog, MSVD, and MSRVTT datasets. Furthermore, for binary classification,



每个查询嵌入一个图 3：InstructBLIP 的模型架构。Q-Former 从冻结图像编码器的输出嵌入中提取指令感知视觉特征，并将视觉特征作为软提示输入馈送到冻结LLM的 .我们使用语言建模损失对模型进行指令调整以生成响应。

在指令调整之前。第一阶段使用冻结图像编码器对 Q-Former 进行预训练，用于视觉语言表示学习。第二阶段将 Q-Former 的输出调整为使用冻结 LLM .预训练后，我们通过指令调整对 Q-Former 进行微调，其中LLM接收来自 Q-Former 和任务指令的视觉编码作为输入。

扩展 BLIP-2， InstructBLIP 提出了一个指令感知 Q-former模块，该模块将指令文本令牌作为附加输入。该指令通过 Q-Former 的自注意力层与查询嵌入进行交互，并鼓励提取与任务相关的图像特征。结果，接收LLM到有利于指令跟随的视觉信息。我们实证证明（表 2）指令感知视觉特征提取为保留和保留评估提供了实质性的性能改进。

2.4 平衡训练数据集

由于训练数据集数量众多，并且每个数据集的大小存在显著差异，因此将它们均匀混合可能会导致模型过度拟合较小的数据集和不足拟合较大的数据集。为了缓解这个问题，我们建议对概率与其大小或训练样本数量的平方根成正比的数据集进行采样。通常，给定大小为 {S， S， . . . ， S} 的 D 个数据集，在训练过程中从数据集 d 中选择数据样本的概率为 p=

.在此公式之上，我们对某些数据集的权重进行手动调整，以改进优化。这是由数据集和任务的固有差异保证的，尽管大小相似，但需要不同级别的训练强度。具体来说，我们降低了 A-OKVQA 的权重，它的特点是多项选择题，并增加了 OKVQA 的权重，这需要开放式文本生成。在表 2 中，我们表明平衡数据集采样策略提高了保留评估和保留泛化的整体性能。

2.5 推理方法

在推理期间，我们采用两种略有不同的生成方法对不同的数据集进行评估。对于大多数数据集，例如图像字幕和开放式 VQA，直接提示指令调整模型生成响应，然后将其与真实值进行比较以计算指标。另一方面，对于分类和多项选择 VQA 任务，我们采用遵循前人工作的词汇排序方法 [46， 22， 21]。具体来说，我们仍然提示模型生成答案，但将其词汇表限制为候选列表。然后，我们计算每个候选者的对数似然，并选择具有最高值的候选者作为最终预测。此排名方法适用于 ScienceQA、IconQA、A-OKVQA（多项选择）、HatefulMemes、Visual Dialog、MSVD 和 MSRVTT 数据集。此外，对于二元分类，

	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5$_{XL}$) [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5$_{XXL}$) [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5$_{XL}$)	119.9	**84.5**	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5$_{XXL}$)	120.0	83.5	47.9	**65.6**	**51.2**	46.6	**48.5**	54.1	30.9	**70.6**	**44.3**	**25.6**	**53.8**
InstructBLIP (Vicuna-7B)	**123.1**	82.4	49.2	54.3	43.1	50.1	45.2	**59.6**	**34.5**	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	**49.5**	52.1	44.8	**50.7**	45.4	57.5	33.4	63.1	41.2	24.8	51.0

Table 1: Zero-shot results on the held-out datasets. Here, Visdial, HM and SciQA denote the Visual Dialog, HatefulMemes and ScienceQA datasets, respectively. For ScienceQA, we only evaluate on the set with image context. Following previous works [4, 49, 32], we report the CIDEr score [42] for NoCaps and Flickr30K, iVQA accuracy for iVQA, AUC score for HatefulMemes, and Mean Reciprocal Rank (MRR) for Visual Dialog. For all other datasets, we report the top-1 accuracy (%).

we expand the positive and negative labels into a slightly broader set of verbalizers to exploit word frequencies in natural text (e.g., *yes* and *true* for the positive class; *no* and *false* for the negative class).

For the video question-answering task, we utilize four uniformly-sampled frames per video. Each frame is processed by the image encoder and Q-Former individually, and the extracted visual features are concatenated before being fed into the LLM.

2.6 Implementation Details

Architecture. Thanks to the flexibility enabled by the modular architectural design of BLIP-2, we can quickly adapt the model to a wide range of LLMs. In our experiments, we adopt four variations of BLIP-2 with the same image encoder (ViT-g/14 [10]) but different frozen LLMs, including FlanT5-XL (3B), FlanT5-XXL (11B), Vicuna-7B and Vicuna-13B. FlanT5 [7] is an instruction-tuned model based on the encoder-decoder Transformer T5 [34]. Vicuna [2], on the other hand, is a recently released decoder-only Transformer instruction-tuned from LLaMA [41]. During vision-language instruction tuning, we initialize the model from pre-trained BLIP-2 checkpoints, and only finetune the parameters of Q-Former while keeping both the image encoder and the LLM frozen. Since the original BLIP-2 models do not include checkpoints for Vicuna, we perform pre-training with Vicuna using the same procedure as BLIP-2.

Training and Hyper-parameters. We use the LAVIS library [19] for implementation, training, and evaluation. All models are instruction-tuned with a maximum of 60K steps and we validate model's performance every 3K steps. For each model, a single optimal checkpoint is selected and used for evaluations on all datasets. We employ a batch size of 192, 128, and 64 for the 3B, 7B, and 11/13B models, respectively. The AdamW [26] optimizer is used, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.05. Additionally, we apply a linear warmup of the learning rate during the initial 1,000 steps, increasing from 10^{-8} to 10^{-5}, followed by a cosine decay with a minimum learning rate of 0. All models are trained utilizing 16 Nvidia A100 (40G) GPUs and are completed within 1.5 days.

3 Experimental Results and Analysis

3.1 Zero-shot Evaluation

We first evaluate InstructBLIP models on the set of 13 held-out datasets with instructions provided in Appendix E. We compare InstructBLIP with the previous SOTA models BLIP-2 and Flamingo. As demonstrated in Table 1, we achieve new zero-shot SOTA results on all datasets. InstructBLIP consistently surpasses its original backbone, BLIP-2, by a significant margin across all LLMs,

	无帽	Flickr的30K	GQA	VSR	图标QA	文本VQA	Visdial	HM	VizWiz	科学技术和图像	MSVD QA	MSRVTT QA	iVQA
火烈鸟-3B [4]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
火烈鸟-9B [4]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
火烈鸟-80B [4]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 （法兰T5） [20]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 （法兰T5） [20]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 （骆马-7B）	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 （骆马-13B）	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP （FlanT5）	119.9	84.5	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP （FlanT5）	120.0	83.5	47.9	65.6	51.2	46.6	48.5	54.1	30.9	70.6	44.3	25.6	53.8
InstructBLIP （Vicuna-7B）	123.1	82.4	49.2	54.3	43.1	50.1	45.2	59.6	34.5	60.5	41.8	22.1	52.2
InstructBLIP （Vicuna-13B）	121.9	82.8	49.5	52.1	44.8	50.7	45.4	57.5	33.4	63.1	41.2	24.8	51.0

表 1：保留数据集的零镜头结果。在这里，Visdial、HM 和 SciQA 分别表示 Visual Dialog、HatefulMemes 和 ScienceQA 数据集。对于 ScienceQA，我们只在具有图像背景的集合上进行评估。根据以前的工作 [4 ， 49 ， 32]，我们报告了 NoCaps 和 Flickr30K 的 CIDEr 分数 [42]、iVQA 的 iVQA 准确性、HatefulMemes 的 AUC 分数和视觉对话的平均倒数排名 （MRR）。对于所有其他数据集，我们报告的准确率排名前 1 （%）。

我们将肯定和否定标签扩展为一组稍宽泛的动词器，以利用自然文本中的词频（例如，Yes 和 true 表示肯定类;no 和 false 表示负类）。

对于视频问答任务，我们为每个视频使用四个统一采样的帧。每一帧都由图像编码器和 Q-Former 单独处理，提取的视觉特征在馈送到 LLM.

2.6 实现细节

建筑。得益于 BLIP-2 的模块化架构设计所实现的灵活性，我们可以快速使模型适应各种 LLMs.在我们的实验中，我们采用了四种具有相同图像编码器 （ViT-g/14 [10]）但不同的冷冻LLMs的 BLIP-2 变体，包括 FlanT5XL （3B）、FlanT5-XXL （11B）、Vicuna-7B 和 Vicuna-13B。FlanT5 [7] 是一个基于编码器-解码器 Transformer T5 [34] 的指令调整模型。另一方面，Vicuna [2] 是最近发布的一款仅解码器的 Transformer 指令，由 LLaMA [41] 调优。在视觉语言指令调整期间，我们从预先训练的 BLIP-2 检查点初始化模型，并且只微调 Q-Former 的参数，同时保持图像编码器和LLM冻结。由于最初的 BLIP-2 模型不包括 Vicuna 的检查点，因此我们使用与 BLIP-2 相同的程序对 Vicuna 进行预训练。

训练和超参数。我们使用 LAVIS 库 [19] 进行实施、培训和评估。所有模型都经过指令调整，最大步长为 60K，我们每 3K 步验证一次模型的性能。对于每个模型，将选择一个最佳检查点，并将其用于对所有数据集进行评估。我们对 3B、7B 和 11/13B 模型分别采用 192、128 和 64 的批量大小。使用 AdamW [26] 优化器，β= 0.9，β= 0.999，权重衰减为 0.05。此外，我们在最初的 1,000 个步骤中应用学习率的线性预热，从 10 增加到 10，然后是最小学习率为 0 的余弦衰减。

所有模型都使用 16 个 Nvidia A100 （40G） GPU 进行训练，并在 1.5 天内完成。

3 实验结果与分析

3.1 零样本评估

我们首先根据附录 E 中提供的说明在 13 个保留数据集上评估 InstructBLIP 模型。我们将 InstructBLIP 与之前的 SOTA 型号 BLIP-2 和 Flamingo 进行了比较。如表 1 所示，我们在所有数据集上都获得了新的零样本 SOTA 结果。InstructBLIP 始终超过其原始主干 BLIP-2，在所有 LLMs、

Model	Held-in Avg.	GQA	ScienceQA (image-context)	IconQA	VizWiz	iVQA
InstructBLIP (FlanT5$_{XL}$)	94.1	48.4	70.4	50.0	32.7	53.1
w/o Instruction-aware Visual Features	89.8	45.9 (↓2.5)	63.4 (↓7.0)	45.8 (↓4.2)	25.1 (↓7.6)	47.5 (↓5.6)
w/o Data Balancing	92.6	46.8 (↓1.6)	66.0 (↓4.4)	49.9 (↓0.1)	31.8 (↓0.9)	51.1 (↓2.0)
InstructBLIP (Vicuna-7B)	100.8	49.2	60.5	43.1	34.5	52.2
w/o Instruction-aware Visual Features	98.9	48.2 (↓1.0)	55.2 (↓5.3)	41.2 (↓1.9)	32.4 (↓2.1)	36.8 (↓15.4)
w/o Data Balancing	98.8	47.8 (↓1.4)	59.4 (↓1.1)	43.5 (↑0.4)	32.3 (↓2.2)	50.3 (↓1.9)

Table 2: Results of ablation studies that remove the instruction-aware Visual Features (Section 2.3) and the balanced data sampling strategy (Section 2.4). For held-in evaluation, we compute the average score of four datasets, including COCO Caption, OKVQA, A-OKVQA, and TextCaps. For held-out evaluation, we show five datasets from different tasks.

demonstrating the effectiveness of vision-language instruction tuning. For instance, InstructBLIP FlanT5$_{XL}$ yields an average relative improvement of 15.0% when compared to BLIP-2 FlanT5$_{XL}$. Furthermore, instruction tuning boosts zero-shot generalization on unseen task categories such as video QA. InstructBLIP achieves up to 47.1% relative improvement on MSRVTT-QA over the previous SOTA despite having never been trained with temporal video data. Finally, our smallest InstructBLIP FlanT5$_{XL}$ with 4B parameters outperforms Flamingo-80B on all six shared evaluation datasets with an average relative improvement of 24.8%.

For the Visual Dialog dataset, we choose to report the Mean Reciprocal Rank (MRR) over the Normalized Discounted Cumulative Gain (NDCG) metric. This is because NDCG favors generic and uncertain answers while MRR prefers certain responses [32], making MRR better aligned with the zero-shot evaluation scenario.

3.2 Ablation Study on Instruction Tuning Techniques

To investigate the impact of the instruction-aware visual feature extraction (Section 2.3) and the balanced dataset sampling strategy (Section 2.4), we conduct ablation studies during the instruction tuning process. As illustrated in Table 2, the removal of instruction awareness in visual features downgrades performance significantly across all datasets. The performance drop is more severe in datasets that involve spatial visual reasoning (e.g., ScienceQA) or temporal visual reasoning (e.g., iVQA), where the instruction input to the Q-Former can guide visual features to attend to informative image regions. The removal of the data balancing strategy causes unstable and uneven training, as different datasets achieve peak performance at drastically different training steps. The lack of synchronized progress over multiple datasets harms the overall performance.

3.3 Qualitative Evaluation

Besides the systematic evaluation on public benchmarks, we further qualitatively examine InstructBLIP with more diverse images and instructions. As illustrated in Figure 1, InstructBLIP demonstrates its capacity for complex visual reasoning. For example, it can reasonably infer from the visual scene what could have happened and deduce the type of disaster from the location of the scene, which it extrapolates based on visual evidence like the palm trees. Moreover, InstructBLIP is capable of connecting visual input with embedded textual knowledge and generate informative responses, such as intruducing a famous painting. Furthermore, in descriptions of the overall atmosphere, InstructBLIP exhibits the ability to comprehend metaphorical implications of the visual imagery. Finally, we show that InstructBLIP can engage in multi-turn conversations, effectively considering the dialog history when making new responses.

In Appendix B, we qualitatively compare InstructBLIP with concurrent multimodal models (GPT-4 [33], LLaVA [25], MiniGPT-4 [52]). Although all models are capable of generating long-form responses, InstructBLIP's outputs generally contains more proper visual details and exhibits logically coherent reasoning steps. Importantly, we argue that long-form responses are not always preferable. For example, in Figure 2 of the Appendix, InstructBLIP directly addresses the user's intent by adaptively adjusting the response length, while LLaVA and MiniGPT-4 generate long and less

模型 Hold-in 平均 GQA			(图像上下文)	IconQA VizWiz iVQA
InstructBLIP （FlanT5） 94.1 48.4 70.4 50.0 32.7 53.1				
无指令感知视觉功能 89.8 45.9 （↓2.5） 63.4 （↓7.0） 45.8 （↓4.2） 25.1 （↓7.6） 47.5 （↓5.6）				
无数据平衡 92.6 46.8 （↓1.6） 66.0 （↓4.4） 49.9 （↓0.1） 31.8 （↓0.9） 51.1 （↓2.0）				
InstructBLIP （Vicuna-7B） 100.8 49.2 60.5 43.1 34.5 52.2				
无指令感知视觉功能 98.9 48.2 （↓1.0） 55.2 （↓5.3） 41.2 （↓1.9） 32.4 （↓2.1） 36.8 （↓15.4）				
无数据平衡 98.8 47.8 （↓1.4） 59.4 （↓1.1） 43.5 （↓0.4） 32.3 （↓2.2） 50.3 （↓1.9）				

表 2：去除指令感知视觉特征（第 2.3 节）和平衡数据采样策略（第 2.4 节）的消融研究结果。对于保留评估，我们计算了四个数据集的平均分数，包括 COCO Caption、OKVQA、A-OKVQA 和 TextCaps。对于保留评估，我们展示了来自不同任务的 5 个数据集。

证明视觉语言教学调整的有效性。例如，与 BLIP-2 FlanT5 相比，InstructBLIP FlanT5 的平均相对改进为 15.0%。此外，指令调优还促进了对视频 QA 等看不见的任务类别的零镜头泛化。尽管从未使用时间视频数据进行过训练，但与之前的 SOTA 相比，InstructBLIP 在 MSRVTT-QA 上实现了高达 47.1% 的相对改进。最后，我们最小的具有 4B 参数的 InstructBLIP FlanT5 在所有六个共享评估数据集上都优于 Flamingo-80B，平均相对改进 24.8%。

对于 Visual Dialog 数据集，我们选择报告标准化折扣累积增益 （NDCG） 指标的平均倒数秩（MRR）。这是因为 NDCG 偏爱通用和不确定的答案，而 MRR 偏爱某些答案 [32]，这使得 MRR 更符合零镜头评估情景。

3.2 指令调谐技术的消融研究

为了研究指令感知视觉特征提取（第 2.3 节）和平衡数据集采样策略（第 2.4 节）的影响，我们在指令调整过程中进行了消融研究。如表 2 所示，删除视觉特征中的指令感知会显著降低所有数据集的性能。在涉及空间视觉推理（例如 ScienceQA）或时间视觉推理（例如 iVQA）的数据集中，性能下降更为严重，其中 Q-Former 的指令输入可以指导视觉特征关注信息丰富的图像区域。删除数据均衡策略会导致训练不稳定和不均匀，因为不同的数据集在截然不同的训练步骤中实现了最佳性能。多个数据集上缺乏同步进度会损害整体性能。

3.3 定性评价

除了对公共基准测试的系统评估外，我们还进一步用更多样化的图像和指令对 InstructBLIP 进行了定性检查。如图 1 所示，InstructBLIP 演示了其进行复杂视觉推理的能力。例如，它可以从视觉场景中合理地推断可能发生的事情，并从场景的位置推断出灾难的类型，然后根据棕榈树等视觉证据进行推断。此外，InstructBLIP 能够将视觉输入与嵌入的文本知识联系起来，并生成信息响应，例如闯入一幅名画。此外，在对整体氛围的描述中，InstructBLIP 表现出理解视觉图像的隐喻含义的能力。最后，我们展示了 InstructBLIP 可以进行多轮次对话，在做出新响应时有效地考虑对话历史。

在附录 B 中，我们定性地将 InstructBLIP 与并发多模态模型（GPT4 [33]、LLaVA [25]、MiniGPT-4 [52]）进行比较。尽管所有模型都能够生成长格式响应，但 InstructBLIP 的输出通常包含更适当的视觉细节，并表现出逻辑连贯的推理步骤。重要的是，我们认为长篇回应并不总是可取的。例如，在附录的图 2 中，InstructBLIP 通过自适应调整响应长度直接解决用户的意图，而 LLaVA 和 MiniGPT-4 生成的 long 和更少

relevant sentences. These advantages of InstructBLIP are a result of the diverse instruction tuning data and an effective architectural design.

3.4 Instruction Tuning vs. Multitask Learning

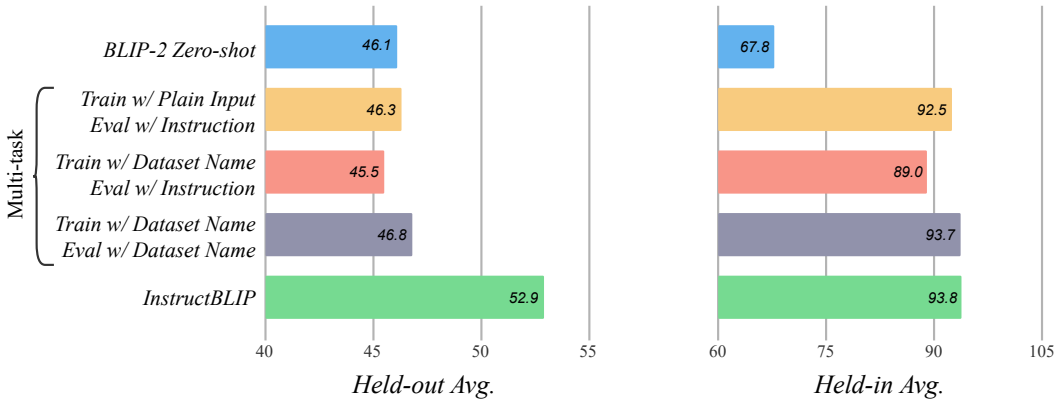


Figure 4: Comparison of instruction tuning and multitask training based on BLIP-2 FlanT5$_{XL}$ backbone. For held-in evaluation, we compute the average score across all held-in datasets. For held-out evaluation, we compute the average score across GQA, TextVQA, VSR, HatefulMemes, IconQA, ScienceQA, iVQA, VizWiz.

A direct analogue to instruction tuning is multitask learning, a widely used method that involves the simultaneous training of multiple datasets with the goal of improving the performance of each individual dataset. To investigate whether the improvement in zero-shot generalization observed in instruction tuning is mainly from the formatting of instructions or merely from multitasking, we conduct a comparative analysis between these two approaches under identical training settings.

Following [46], we consider two multitask training approaches. In the first approach, the model is trained using the vanilla input-output format of the training datasets without instructions. During evaluation, instructions are still provided to the model, indicating the specific task to be performed. However, an exception is made for image captioning, as the model achieves better scores when only receiving the image as input. For the second approach, we take a step towards instruction tuning by prepending a `[Task:Dataset]` identifier to the text input during training. For example, we prepend `[Visual question answering:VQAv2]` for the VQAv2 dataset. During evaluation, we explore both instructions and this identifier. Particularly, for the identifier of held-out datasets, we only use the task name since the model never sees the dataset name.

The results are shown in Figure 4, including BLIP-2 zero-shot, multitask training, and instruction tuning. All of these models are based on the BLIP-2 FlanT5$_{XL}$ backbone and adhere to the identical training configurations delineated in Section 2. Overall, we can conclude two insights from the results. Firstly, instruction tuning and multitask learning exhibit similar performance on the held-in datasets. This suggests that the model can fit these two different input patterns comparably well, as long as it has been trained with such data. On the other hand, instruction tuning yields a significant improvement over multitask learning on unseen held-out datasets, whereas multitask learning still performs on par with the original BLIP-2. This indicates that instruction tuning is the key to enhance the model's zero-shot generalization ability.

3.5 Finetuning InstructBLIP on Downstream Tasks

We further finetune the InstructBLIP models to investigate its performance on learning a specific dataset. Compared to most previous methods (e.g., Flamingo, BLIP-2) which increase the input image resolution and finetune the visual encoder on downstream tasks, InstructBLIP maintains the same image resolution (224×224) during instruction tuning and keeps the visual encoder frozen during finetuning. This significantly reduces the number of trainable parameters from 1.2B to 188M, thus greatly improves finetuning efficiency.

例如，相关句子。InstructBLIP 的这些优势是多样化的指令调优数据和有效的架构设计的结果。

3.4 指令调整与多任务学习

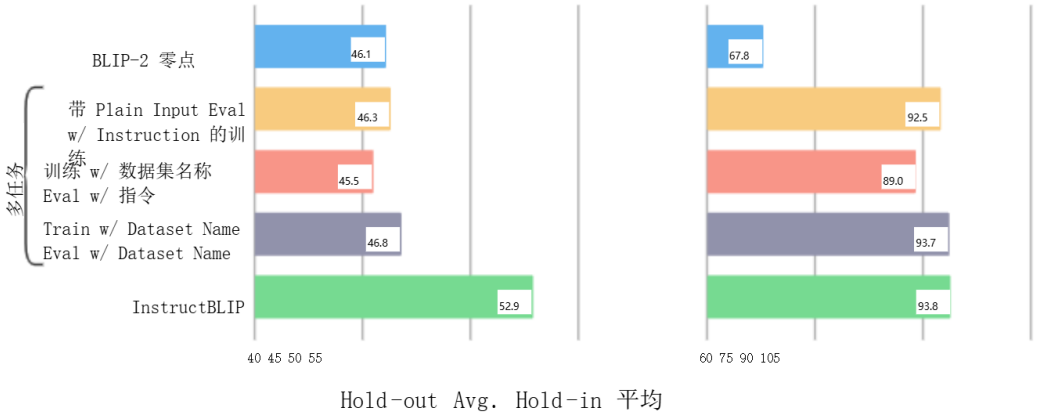


图 4：基于 BLIP-2 FlanT5 主干的指令调优和多任务训练的比较。对于暂留评估，我们计算所有暂留数据集的平均分数。对于保留评估，我们计算 GQA、TextVQA、VSR、HatefulMemes、IconQA、ScienceQA、iVQA、VizWiz 的平均分数。

与指令调优直接类似的是多任务学习，这是一种广泛使用的方法，涉及同时训练多个数据集，目的是提高每个数据集的性能。为了研究在指令调整中观察到的零镜头泛化的改进是主要来自指令的格式还是仅仅来自多任务处理，我们在相同的训练设置下对这两种方法进行了比较分析。

按照 [46]，我们考虑两种多任务训练方法。在第一种方法中，使用训练数据集的 vanilla input-output 格式对模型进行训练，无需说明。在评估期间，仍会向模型提供说明，指示要执行的特定任务。但是，图像描述是一个例外，因为当仅接收图像作为输入时，模型会获得更好的分数。对于第二种方法，我们通过在训练期间为文本输入预置 [Task：Dataset] 标识符来向指令优化迈出一步。例如，我们为 VQAv2 数据集预置了 [Visual question answering：VQAv2]。在评估过程中，我们会探索 instructions 和此标识符。特别是，对于保留数据集的标识符，我们只使用任务名称，因为模型永远不会看到数据集名称。

结果如图 4 所示，包括 BLIP-2 零样本、多任务训练和指令调优。所有这些模型都基于 BLIP-2 FlanT5backbone ，并遵循第 2 节中描述的相同训练配置。总的来说，我们可以从结果中得出两个见解。首先，指令调优和多任务学习在保留数据集上表现出相似的性能。这表明，只要该模型已经用这些数据进行了训练，就可以很好地拟合这两种不同的输入模式。另一方面，与在看不见的保留数据集上的多任务学习相比，指令调优产生了显着改进，而多任务学习的性能仍然与原始 BLIP-2 相当。这表明指令调优是增强模型零镜头泛化能力的关键。

3.5 对下游任务进行微调 InstructBLIP

我们进一步微调 InstructBLIP 模型，以研究其在学习特定数据集时的性能。与大多数以前的方法（例如，Flamingo、BLIP-2）相比，这些方法可以提高输入图像分辨率并在下游任务上微调视觉编码器，InstructBLIP 在指令调整期间保持相同的图像分辨率（224×224），并在微调期间保持视觉编码器冻结。这显着减少了可训练参数的数量，从 1.2B 减少到 188M，从而大大提高了微调效率。

	ScienceQA IMG	OCR-VQA	OKVQA	A-OKVQA			
				Direct Answer		Multi-choice	
				Val	Test	Val	Test
Previous SOTA	LLaVA [25] 89.0	GIT [43] 70.3	PaLM-E(562B) [9] **66.1**	[15] 56.3	[37] 61.6	[15] 73.2	[37] 73.6
BLIP-2 (FlanT5$_{XXL}$)	89.5	72.7	54.7	57.6	53.7	80.2	76.2
InstructBLIP (FlanT5$_{XXL}$)	**90.7**	**73.3**	55.5	57.1	54.8	**81.0**	**76.7**
BLIP-2 (Vicuna-7B)	77.3	69.1	59.3	60.0	58.7	72.1	69.0
InstructBLIP (Vicuna-7B)	79.5	72.8	62.1	**64.0**	**62.1**	75.7	73.4

Table 3: Results of finetuning BLIP-2 and InstructBLIP on downstream datasets. Compared to BLIP-2, InstructBLIP provides a better weight initialization model and achieves SOTA performance on three out of four datasets.

The results are shown in Table 3. Compared to BLIP-2, InstructBLIP leads to better finetuning performance on all datasets, which validates InstructBLIP as a better weight initialization model for task-specific finetuning. InstructBLIP sets new state-of-the-art finetuning performance on ScienceQA (IMG), OCR-VQA, A-OKVQA, and is outperformed on OKVQA by PaLM-E [9] with 562B parameters.

Additionally, we observe that the FlanT5-based InstructBLIP is superior at multi-choice tasks, whereas Vicuna-based InstructBLIP is generally better at open-ended generation tasks. This disparity can be primarily attributed to the capabilities of their frozen LLMs, as they both employ the same image encoder. Although FlanT5 and Vicuna are both instruction-tuned LLMs, their instruction data significantly differ. FlanT5 is mainly finetuned on NLP benchmarks containing many multi-choice QA and classification datasets, while Vicuna is finetuned on open-ended instruction-following data.

4 Related Work

Instruction tuning aims to teach language models to follow natural language instructions, which has been shown to improve their generalization performance to unseen tasks. Some methods collect instruction tuning data by converting existing NLP datasets into instruction format using templates [46, 7, 35, 45]. Others use LLMs (e.g., GPT-3 [5]) to generate instruction data [2, 13, 44, 40] with improved diversity.

Instruction-tuned LLMs have been adapted for vision-to-language generation tasks by injecting visual information to the LLMs. BLIP-2 [20] uses frozen FlanT5 models, and trains a Q-Former to extract visual features as input to the LLMs. MiniGPT-4 [52] uses the same pretrained visual encoder and Q-Former from BLIP-2, but uses Vicuna [2] as the LLM and performs training using ChatGPT [1]-generated image captions longer than the BLIP-2 training data. LLaVA [25] directly projects the output of a visual encoder as input to a LLaMA/Vinuca LLM, and finetunes the LLM on vision-language conversational data generated by GPT-4 [33]. mPLUG-owl [50] performs low-rank adaption [14] to a LLaMA [41] model using both text instruction data and vision-language instruction data from LLaVA. A separate work is MultiInstruct [48], which performs vision-language instruction tuning without a pretrained LLM, leading to less competitive performance.

Compared to existing methods, InstructBLIP uses a much wider range of vision-language instruction data, covering both template-based converted data and LLM-generated data. Architecture wise, InstructBLIP proposes an instruction-aware visual feature extraction mechanism. Furthermore, our paper provides a comprehensive analysis on various aspects of vision-language instruction tuning, validating its advantages on generalizing to unseen tasks.

5 Conclusion

In this paper, we present InstructBLIP, a simple yet novel instruction tuning framework towards generalized vision-language models. We perform a comprehensive study on vision-language instruction tuning and demonstrate the capability of InstructBLIP models to generalize to a wide range of unseen tasks with state-of-the-art performance. Qualitative examples also exhibit InstructBLIP's various

	科学质量保证 IMG	OCR-VQA	OKVQA		A-OKVQA Direct Answer		Multi-choice	
					Val	Test	Val	Test
上一页 SOTA	拉瓦 [25] 89.0	胃肠道 [43] 70.3	帕LM-E（562B） [9] 66.1	[15] 56.3	[37] 61.6	[15] 73.2	[37] 73.6	
BLIP-2 （FlanT5） 89.5 72.7 54.7 57.6 53.7 80.2 76.2								
InstructBLIP （FlanT5） 90.7 73.3 55.5 57.1 54.8 81.0 76.7								
BLIP-2 （骆马-7B） 77.3 69.1 59.3 60.0 58.7 72.1 69.0								
InstructBLIP （骆马-7B） 79.5 72.8 62.1 64.0 62.1 75.7 73.4								

表 3：在下游数据集上微调 BLIP-2 和 InstructBLIP 的结果。与 BLIP-2 相比，InstructBLIP 提供了更好的权重初始化模型，并在四个数据集中的三个数据集上实现了 SOTA 性能。

结果如表 3 所示。与 BLIP-2 相比，InstructBLIP 在所有数据集上都能获得更好的微调性能，这验证了 InstructBLIP 是用于特定任务微调的更好的权重初始化模型。InstructBLIP 在 ScienceQA （IMG）、OCR-VQA、A-OKVQA 上创造了新的最先进的微调性能，并且在 OKVQA 上以 562B 参数被 PaLM-E [9] 超越。

此外，我们观察到基于 FlanT5 的 InstructBLIP 在多项选择任务方面更胜一筹，而基于 Vicuna 的 InstructBLIP 通常在开放式生成任务方面表现更好。这种差异主要归因于他们的功能 冻结 LLMs，因为它们都使用相同的图像编码器。尽管 FlanT5 和 Vicuna 都是 instructiontuned LLMs，但它们的指令数据存在显著差异。FlanT5 主要根据包含许多多选 QA 和分类数据集的 NLP 基准测试进行微调，而 Vicuna 则根据开放式指令跟踪数据进行微调。

4 相关工作

指令调优旨在教语言模型遵循自然语言指令，这已被证明可以提高它们对看不见的任务的泛化性能。一些方法通过使用模板将现有的 NLP 数据集转换为指令格式来收集指令调优数据 [46， 7， 35， 45]。其他人使用 LLMs （例如，GPT-3 [5]）生成具有更高多样性的指令数据 [2， 13， 44， 40]。

Instruction-tuned LLMs 已通过将视觉信息LLMs注入 .BLIP-2 [20] 使用冻结的 FlanT5 模型，并训练 Q-Former 提取视觉特征作为输入。LLMsMiniGPT-4 [52] 使用与 BLIP-2 相同的预训练视觉编码器和 Q-Former，但使用 Vicuna [2] 作为，LLM并使用 ChatGPT [1] 生成的图像字幕进行训练，时间比 BLIP-2 训练数据更长。LLaVA [25] 直接将视觉编码器的输出投影为 LLaMA/Vinuca LLM的输入，并微调 GPT-4 [33] 生成的LLM视觉语言对话数据。mPLUG-owl [50] 使用来自 LLaVA 的文本指令数据和视觉语言指令数据对 LLaMA [41] 模型进行低秩适应 [14]。一个单独的工作是 MultiInstruct [48]，它在没有预训练LLM的情况下执行视觉语言指令调整，导致性能竞争力降低。

与现有方法相比，InstructBLIP 使用更广泛的视觉语言教学数据，涵盖基于模板的转换数据和LLM生成的数据。在架构方面，InstructBLIP 提出了一种指令感知的视觉特征提取机制。此外，本文对视觉语言指令调整的各个方面进行了全面分析，验证了其在泛化到看不见的任务方面的优势。

5 结论

在本文中，我们提出了 InstructBLIP，这是一个针对广义视觉语言模型的简单而新颖的指令调整框架。我们对视觉语言教学调整进行了全面研究，并展示了 InstructBLIP 模型以最先进的性能推广到各种看不见的任务的能力。定性示例还展示了 InstructBLIP 的各种

capabilities on instruction following, such as complex visual reasoning, knowledge-grounded image description, and multi-turn conversations. Furthermore, we show that InstructBLIP can serve as an enhanced model initialization for downstream task finetuning, achieving state-of-the-art results. We hope that InstructBLIP can spur new research in general-purpose multimodal AI and its applications.

References

[1] Chatgpt. https://openai.com/blog/chatgpt, 2023. 9

[2] Vicuna. https://github.com/lm-sys/FastChat, 2023. 3, 6, 9

[3] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, pages 8948–8957, 2019. 3, 16

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *NeurIPS*, 2022. 3, 6

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 9

[6] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*, 2021. 1

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 1, 3, 6, 9

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017. 3, 16

[9] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. 9

[10] Yuxin Fang, Wen Wang, Binhui Xie, Quan-Sen Sun, Ledell Yu Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *ArXiv*, abs/2211.07636, 2022. 6

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, July 2017. 3, 16

[12] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. 3, 16

[13] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. *ArXiv*, abs/2212.09689, 2022. 9

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 9

[15] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning, 2023. 9

[16] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 3, 16

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 16

[18] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020. 3, 16

我们提供了关于指令遵循的 InstructBLIP 功能，例如复杂的视觉推理、基于知识的图像描述和多轮次对话。此外，我们表明 InstructBLIP 可以作为下游任务微调的增强模型初始化，获得最先进的结果。我们希望 InstructBLIP 能够激发通用多模态 AI 及其应用的新研究。

引用

[1] 聊天网。https://openai.com/blog/chatgpt，2023 年。9 [2] 骆马。https://github.com/lm-sys/FastChat，2023 年。3, 6, 9

[3] Harsh Agrawal、Karan Desai、Yufei Wang、Xinlei Chen、Rishabh Jain、Mark Johnson、Dhruv Batra、Devi Parikh、Stefan Lee 和 Peter Anderson。nocaps：大规模的新奇对象标题。在 ICCV，第 8948-8957 页，2019 年。3， 16 [4] 让-巴蒂斯特·阿莱拉克， 杰夫·多纳休， 宝琳·吕克， 安托万·米奇， 伊恩·巴尔， 亚娜·哈森， 卡雷尔·伦克， 亚瑟·门施， 凯瑟琳·米利肯， 马尔科姆·雷诺兹， 罗曼·林， 伊丽莎·卢瑟福， 塞尔坎·卡比， 韩腾达， 龚志涛， 西娜·萨曼古伊， 玛丽安·蒙泰罗， 雅各布·梅尼克， 塞巴斯蒂安·博尔格奥， 安迪·布洛克， 艾达·内马扎德， 萨汉德·沙里夫扎德， 米科乌·阿伊·比恩科夫斯基， 里卡多·巴雷拉， 奥里奥尔·维亚尔斯，Andrew Zisserman 和 Karén Simonyan。Flamingo：用于小样本学习的视觉语言模型。在 S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh 编辑中，NeurIPS，2022 年。3， 6 [5] 汤姆·布朗，本杰明·曼， 尼克·莱德， 梅兰妮·苏比亚， 贾里德·卡普兰， 普拉弗拉·达里瓦尔， 阿尔温德·尼拉坎坦， 普拉纳夫·希亚姆， 吉里什·萨斯特里， 阿曼达·阿斯凯尔， 桑迪尼·阿加瓦尔， 阿里尔·赫伯特-沃斯， 格雷琴·克鲁格， 汤姆·亨尼汉， Rewon Child， 阿迪亚·拉梅什， 丹尼尔·齐格勒， 杰弗里·吴， 克莱门斯·温特， 克里斯托弗·黑塞， 马克·陈， 埃里克·西格勒， 马特乌斯·利特文， 斯科特·格雷， 本杰明·切斯， 杰克·克拉克， 克里斯托弗·伯纳、山姆·麦克坎德利什、亚历克·拉德福德、伊利亚·萨茨克弗和达里奥·阿莫迪。语言模型是少数机会的学习者。arXiv 预印本 arXiv：2005.14165,2020 年。9 [6] Jaemin Cho、Jie Lei、Hao Tan 和 Mohit Bansal。通过文本生成统一视觉和语言任务。

arXiv 预印本 arXiv：2102.02779,2021 年。1

[7] Hyung Won Chung， Le Hou， Shayne Longpre， Barret Zoph， Yi Tay， William Fedus， Eric Li，Xuezhi Wang， Mostafa Dehghani， Siddhartha Brahma， Albert Webson， Shixiang Shane Gu，Zhuyun Dai， Mirac Suzgun， Xinyun Chen， Aakanksha Chowdhery， Sharan Narang， Gaurav

赵、黄艳平、Andrew M. Dai、于洪坤、Slav Petrov、Ed H. Chi、Jeff Dean、Jacob Devlin、Adam Roberts、Denny 周、Quoc V. Le 和 Jason Wei。扩展指令微调的语言模型。arXiv 预印本 arXiv：2210.11416,2022 年。1, 3, 6, 9

[8] Abhishek Das、Satwik Kottur、Khushi Gupta、Avi Singh、Deshraj Yadav、Jose MF Moura、Devi Parikh 和 Dhruv Batra。可视对话框。在 CVPR，2017 年。3， 16 [9] 丹尼·德里斯、菲·夏、迈赫迪·萨贾迪、科里·林奇、阿坎莎·乔德里、布莱恩·伊赫特、艾赞·瓦希德、乔纳森·汤普森、全武旺、余天河、黄文龙、叶夫根·切博塔尔、皮埃尔·塞尔马内特、丹尼尔·达克沃斯、谢尔盖·莱文、文森特·范胡克、卡罗尔·豪斯曼、马克·杜桑、克劳斯·格雷夫、安迪·曾、伊戈尔·莫达奇和皮特·弗洛伦斯。Palm-e：一种具身的多模态语言模型，2023 年。9 [10] 方玉欣， 王温， 谢斌辉， 孙全森， 吴宇， 王兴刚， 黄铁军，

Xinlong Wang， 和 Yue Cao.Eva：探索大规模掩蔽视觉表示学习的局限性。ArXiv，abs/2211.07636,2022 年。6 [11] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。在 vqa 中生成 v

matter：提升图像理解在视觉问答中的作用。在 CVPR 中，2017 年 7 月。3， 16 [12] Danna Gurari， Qing Li， Abigale J. Stangl， Anhong Guo， Chi Lin， Kristen Grauman， Jiebo Luo， and Jeffrey P.

比格姆。Vizwiz 大挑战：回答盲人的视觉问题。在 CVPR，2018 年。3， 16 [13] 或霍诺维奇、托马斯·西亚洛姆、奥马尔·利维和蒂莫·希克。不自然的指令：调整语言

模型（几乎）没有人工。ArXiv，abs/2212.09689,2022 年。9 [14] Edward J. 胡， Shen Yelong Shen，Phillip Wallis， 朱泽远， 李元志， 王雪， 王璐， 和

陈伟珠.Lora：大型语言模型的低秩改编。在 ICLR，2022 年。9 [15] 胡宇石、华航、杨正元、史维佳、Noah A. Smith 和罗杰波。Promptcap：提示-

引导式任务感知图像字幕，2023 年。9 [16] 德鲁·哈德森和克里斯托弗·曼宁。Gqa：用于真实世界视觉推理的新数据集，以及

作文问答。在 CVPR，2019 年。3， 16 [17] 安德烈·卡尔帕西和李飞飞。用于生成图像描述的深度视觉语义对齐。在

IEEE 计算机视觉和模式识别会议 （CVPR） 会议记录，2015 年 6 月。16

[18] 杜韦·基拉、哈默德·菲鲁兹、阿拉文德·莫汉、维达努伊·戈斯瓦米、阿曼普雷特·辛格、普拉蒂克·林希亚、和 Davide Testuggine。仇恨模因挑战：检测多模态模因中的仇恨言论。在 NeurIPS 中，2020 年。3, 16

[19] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022. 6

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 3, 4, 6, 9, 16

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5, 16

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 5

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 16

[24] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 3, 7, 9, 16

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[27] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 1

[28] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 3, 16

[29] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 3, 16

[30] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 3, 16

[31] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 3, 16

[32] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 6, 7

[33] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 7, 9

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 2020. 3, 6

[35] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *ICLR*, 2022. 9

[36] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, 2022. 3, 16

[37] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 9

[38] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioningwith reading comprehension. 2020. 3, 16

[39] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 3, 16

[40] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 9

在 NeurIPS [19] Dongxu Li、Junnan Li、Hung Le、Guangsen Wang、Silvio Savarese 和 Steven C. H. Hoi 中。Lavis：一个库
语言视觉智能，2022 年。6

[20] Junnan Li， Dongxu Li， Silvio Savarese 和 Steven Hoi.Blip-2：引导语言图像预训练
使用冻结图像编码器和大型语言模型。在 ICML 中，2023 年。1, 3, 4, 6, 9, 16

[21] 李俊楠、李东旭、熊才明和史蒂文·海。Blip： Bootstrapping 语言图像预训练
用于统一视觉-语言的理解和生成。在 ICML 中，2022 年。5, 16

[22] Junnan Li， Ramprasaath Selvaraju， Akhilesh Gotmare， Shafiq Joty， Caim Xiong， 和 Steven Chu Hong
海。先对齐后融合：通过动量蒸馏进行视觉和语言表示学习。在 NeurIPS 中，2021 年。5

[23] 林宗义， 迈克尔·梅尔， 塞尔吉·贝隆吉， 詹姆斯·海斯， 彼得罗·佩罗纳， 德瓦·拉马南， 彼得·多拉尔，
和 C. Lawrence Zitnick。Microsoft coco：上下文中的常见对象。在 ECCV，2014 年。3, 16

[24] Fangyu Liu、Guy Edward Toh Emerson 和 Nigel Collier。视觉空间推理。的
计算语言学协会， 2023.3

[25] 刘浩天、李春元、吴庆阳和李勇宰。可视化指令调整。2023. 3, 7, 9, 16

[26] 伊利亚·洛什奇洛夫和弗兰克·胡特尔。解耦的权重衰减正则化。在 ICLR，2019 年。6

[27] Jiasen Lu、Vedanuj Goswami、Marcus Rohrbach、Devi Parikh 和 Stefan Lee。12 合 1：多任务视觉
和语言表征学习。在 CVPR，2020 年。1

[28] 潘璐， Swaroop Mishra， Tony Xia， 邱亮， 张开伟， 朱松春， Oyvind Tafjord， Peter
克拉克和阿什温·卡利安。学习解释：通过思维链进行多模态推理以进行科学问答。在 NeurIPS 中，2022
年。3, 16

[29] 潘璐、邱亮、陈佳琪、夏东、赵一洲、张伟、周宇、梁晓丹和
朱松春.Iconqa：抽象图理解和视觉语言的新基准
推理。在 NeurIPS 数据集和基准跟踪中，2021 年。3, 16

[30] 肯尼斯·马里诺、穆罕默德·拉斯特加里、阿里·法哈迪和鲁兹贝·莫塔吉。Ok-vqa：一个视觉问题
回答需要外部知识的基准。在 CVPR，2019 年。3, 16

[31] 阿南德·米什拉、沙尚克·谢哈尔、阿吉特·库马尔·辛格和阿尼班·查克拉博蒂。Ocr-vqa：视觉问题
通过阅读图像中的文本来回答。在 ICDAR，2019 年。3, 16

[32] 维什瓦克·穆拉哈里、德鲁夫·巴特拉、德维·帕里赫和阿比舍克·达斯。可视化对话的大规模预训练：
一个简单的最先进的基线。在 Andrea Vedaldi、Horst Bischof、Thomas Brox 和 Jan-Michael Frahm 编
辑中，ECCV，2020 年。6, 7

[33] OpenAI。GPT-4 技术报告。ArXiv，abs/2303.08774,2023 年。7, 9

[34] 科林·拉菲尔、诺姆·沙泽尔、亚当·罗伯茨、凯瑟琳·李、莎兰·纳朗、迈克尔·马特纳、周燕琪、
Wei Li 和 Peter J Liu。使用统一的文本转文本转换器探索迁移学习的极限。
机器学习研究杂志，2020 年。3, 6

[35] Victor Sanh， Albert Webson， Colin Raffel， Stephen H. Bach， Lintang Sutawika， Zaid Alyafeai， An-
托因·查芬、阿诺德·斯蒂格勒、阿伦·拉贾、马南·戴伊、M 赛义夫·巴里、徐灿文、乌尔米什·塔克尔、沙尼亚·夏
尔马·夏尔马、伊丽莎·什切克拉、泰云·金、贡詹·查布拉尼、尼哈尔·纳亚克、德巴约蒂·达塔、乔纳森·张、迈克·
江天剑、韩王、马泰奥·马尼卡、沈胜、郑欣勇、哈希特·潘迪、雷切尔·鲍登、托马斯·王、特丽莎拉·尼拉杰、
乔斯·罗森、阿比什特·夏尔马、安德烈·桑蒂利、 Thibault Févry、Jason Alan Fries、Ryan Teehan、Teven
Le Scao、Stella Biderman、Leo Gao、Thomas Wolf 和 Alexander M. Rush。多任务提示训练可实现零样
本任务泛化。在 ICLR，2022 年。9

[36] 达斯汀·施文克、阿普尔夫·坎德尔瓦尔、克里斯托弗·克拉克、肯尼斯·马里诺和鲁兹贝·莫塔吉。一个-
okvqa：使用世界知识进行视觉问答的基准测试。在 Shai Avidan、Gabriel Brostow、Moustapha Cissé、
Giovanni Maria Farinella 和 Tal Hassner 编辑中，ECCV，2022 年。3, 16

[37] 邵振伟， 周 Yu， 孟 旺， 和 Jun Yu.使用答案启发式方法提示大型语言模型
，用于基于知识的视觉问答。计算机视觉和模式识别 （CVPR），2023 年。9

[38] 奥列克西·西多罗夫、胡荣航、马库斯·罗尔巴赫和阿曼普利特·辛格。Textcaps：图像数据集
captioning与阅读理解。2020. 3, 16

[39] Amanpreet Singh、Vivek Natarjan、Meet Shah、Yu 江、Xinlei Chen、Devi Parikh 和 Marcus Rohrbach。
朝着可以读取的 vqa 模型。在 CVPR 中，第 8317-8326 页，2019 年。3, 16

[40] Rohan Taori， Ishaan Gulrajani， 张天一， Yann Dubois， 李雪辰， Carlos Guestrin， Percy Liang，
和 Tatsunori B. Hashimoto。斯坦福羊驼：一种遵循指令的美洲驼模型。https://github。
com/tatsu-lab/stanford_alpaca，2023 年。9

[41] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 6, 9

[42] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 6

[43] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022. 9

[44] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *ArXiv*, abs/2212.10560, 2022. 9

[45] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *EMNLP*, 2022. 9

[46] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022. 1, 5, 8, 9

[47] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1645–1653, 2017. 3, 16

[48] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *ArXiv*, abs/2212.10773, 2022. 9

[49] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021. 3, 6, 16

[50] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Chao Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. 2023. 9

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014. 3, 16

[52] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 7, 9

[41] 雨果·图夫隆、蒂博·拉夫里尔、戈蒂埃·伊扎卡德、泽维尔·马丁内、玛丽-安妮·拉肖、蒂莫西·拉克鲁瓦、巴蒂斯特·罗齐埃、纳曼·戈亚尔、埃里克·汉布罗、费萨尔·阿兹哈尔、奥雷利安·罗德里格斯、阿尔芒·朱兰、爱德华·格雷夫和纪尧姆·兰普尔。Llama：开放高效的基础语言模型。arXiv 预印本 arXiv：2302.13971,2023 年，2, 6, 9

[42] Ramakrishna Vedantam， C. Lawrence Zitnick， 和 Devi Parikh.Cider：基于共识的图像描述评估。2015 年 IEEE 计算机视觉和模式识别会议 （CVPR），第 4566–4575 页，2015 年。6 [43] 王剑锋、杨正元、胡晓伟、李林杰、林凯文、甘哲、刘子成、刘策和

王丽娟.Git：用于视觉和语言的生成式图像到文本转换器，2022 年。9

[44] Yizhong Wang， Yeganeh Kordi， Swaroop Mishra， Alisa Liu， Noah A. Smith， Daniel Khashabi， 和汉纳内·哈吉希尔齐。Self-instruct：将语言模型与自生成的指令保持一致。ArXiv，abs/2212.10560,2022 年。9

[45] Yizhong Wang， Swaroop Mishra， Pegah Alipoormolabashi， Yeganeh Kordi， Amirreza Mirzaei， Atharva 奈克、阿琼·阿肖克、阿鲁特·塞尔万·达纳塞卡兰、安贾娜·阿伦库马尔、大卫·斯塔普、伊桑·帕塔克、扬尼斯·卡拉马诺拉基斯、赖海志、伊山·普罗希特、伊沙尼·蒙达尔、雅各布·安德森、柯比·库兹尼亚、克里马·多西、昆塔尔·库马尔·帕尔、弥勒·帕特尔、梅赫拉德·莫拉沙希、米希尔·帕尔马、米拉利·普罗希特、尼拉杰·瓦什尼、法尼·罗希塔·卡扎、普尔基特·维尔马、拉夫塞哈吉·辛格·普里、鲁尚·卡里亚、沙万·多西、谢拉贾·凯乌尔·桑帕特、悉达多·米什拉、 Sujan Reddy A、Sumanta Patro、Tanay Dixit 和 Xudong Shen。SuperNaturalInstructions：通过对 1600+ NLP 任务的声明性指令进行泛化。在 EMNLP，2022 年。9

[46] Jason Wei， Maarten Bosma， Vincent Y. Zhao， Kelvin Guu， Adams Wei Yu， Brian Lester， Nan Du， Andrew M. Dai 和 Quoc V. Le.微调的语言模型是零样本学习器。在 ICLR，2022 年。1， 5， 8， 9 [47] 徐德静， 赵周， 肖军， 吴飞， 张汉旺， 何祥南， 庄月婷.视频通过逐渐细化对外表和动作的关注来回答问题。第 25 届 ACM 多媒体国际会议论文集，第 1645-1653 页，2017 年。3， 16 [48] 徐志阳、沈英和黄丽福。Multiinstruct：通过以下方式改进多模态零样本学习指令调整。ArXiv，abs/2212.10773,2022 年。9 [49] 安托万·杨、安托万·米奇、约瑟夫·西维奇、伊万·拉普捷夫和科迪莉亚·施密德。只需询问：学会回答来自数百万个旁白视频的问题。在 ICCV 中，第 1686-1697 页，2021 年。3， 6， 16 [50] 叶庆浩， 徐海洋， 徐国海， 叶家博， 闫明， 周毅， 王俊彦， 胡安文， 鹏程石、石亚亚、李晨良、徐元红、陈和宏、田俊峰、齐强、张吉超和黄飞燕。mplug-owl：模块化为大型语言模型提供了多模态。2023. 9 [51] 彼得·杨、爱丽丝·赖、米卡·霍多什和朱莉娅·霍肯迈尔。从图像描述到视觉nnotations：用于对事件描述进行语义推理的新相似性指标。计算语言学协会汇刊， 2， 2014.3， 16 [52] 朱德耀、陈军、沈小倩、李翔和穆罕默德·艾尔霍西尼。Minigpt-4：增强

使用高级大型语言模型进行视觉语言理解，2023 年。7, 9

A Broader Impact

InstructBLIP uses off-the-shelf frozen LLMs. Therefore it inherits some of the shortcomings from the original LLMs, such as hallucinating ungrounded text or generating outputs with bias. We mitigate such shortcomings by improving the model's grounding on the vision and instruction input, and performing vision-language instruction tuning on a diverse set of high-quality datasets. Nevertheless, we do not recommend applying InstructBLIP models to any downstream applications without a prior assessment on safety and fairness specific to that application.

B More Case Studies

[Input Image]	
[Input Instruction]	What is unusual about this image?
[Model Response]	
InstructBLIP	The unusual aspect of this image is that a man is ironing clothes on the back of a yellow taxi cab while it is driving down a bus city street. He is standing on top of the car's trunk, holding an iron and using it to press his clothing items. This situation is not only unconventional but also potentially hazardous, as the man could be at risk of falling off the vehicle or causing an accident due to his actions.
GPT-4	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
MiniGPT-4	The image shows a man standing on the back of a yellow taxi, with a yellow shirt and black pants, and a blue backpack on his back. The taxi is driving on a city street with othr cars and taxis in the background.

Figure 5: The response from InstructBLIP is more comprehensive than GPT-4, more visually-grounded than LLaVA, and more logical than MiniGPT-4. The responses of GPT-4 and LLaVA are obtained from their respective papers, while the official demo is used for MiniGPT-4. Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

更广泛的影响

InstructBLIP 使用现成的 frozen LLMs.因此，它继承了原版LLMs的一些缺点，例如产生幻觉，使文本不接地或生成带有偏见的输出。我们通过改进模型对视觉和指令输入的基础，并在各种高质量数据集上执行视觉-语言指令调整来减轻这些缺点。尽管如此，我们不建议将 InstructBLIP 模型应用于任何下游应用程序，而无需事先评估特定于该应用程序的安全性和公平性。

B 更多案例研究

[输入图像]

[输入指令] 这张图片有什么不寻常之处？

[模型响应]

InstructBLIP 这张图片的不寻常之处在于，一名男子正在一辆黄色出租车的后座上熨衣服，当时它正在公共汽车城市街道上行驶。他站在汽车后备箱的顶部，拿着熨斗，用它来压他的衣服。这种情况不仅非常规，而且具有潜在危险，因为该男子可能会因他的行为而面临从车辆上掉下来或造成事故的风险。

GPT-4 这张图片的不寻常之处在于，一名男子正在连接在行驶中的出租车车顶上的熨衣板上熨烫衣服。

拉瓦 这张图片的不寻常之处在于一名男子在小型货车或面包车的后座上熨烫衣服。这不是进行此活动的典型场所，因为人们通常会在更固定和安全的地方熨烫衣服，例如在家中，使用普通熨衣板。图片中描绘的场景很奇特，因为它涉及车辆上的临时熨烫装置，这可能既不安全又非常规。此外，目前尚不清楚该男子如何在如此不稳定的环境中熨烫衣服时保持平衡和稳定。

MiniGPT-4 该图像显示一名男子站在一辆黄色出租车的后座上，手里拿着他穿着黄色衬衫和黑色裤子，背着一个蓝色的背包。出租车在城市街道上行驶，背景是其他汽车和出租车。

图 5：InstructBLIP 的响应比 GPT-4 更全面，比 LLaVA 更直观，比 MiniGPT-4 更合乎逻辑。GPT-4 和 LLaVA 的回答来自各自的论文，而官方演示则用于 MiniGPT-4。源：

https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

[Input Image]	
[Input Instruction]	Do you know who drew this painting?
[Model Response]	
InstructBLIP	Leonardo da Vinci.
LLaVA	The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.
MiniGPT-4	The painting is the Mona Lisa, also known as the Mona Lisa, by Leonardo da Vinci. It is one of the most famous paintings in the world and is known for its enigmatic smile and the mysterious identity of the sitter. The painting is housed in the Louvre Museum in Paris, France.

Figure 6: In this example, we illustrate that long-form responses are not always preferable. InstructBLIP can directly address the user's intent by adaptively adjusting the response length, while other models tend to generate lengthy paragraphs with less-relevant sentences. The response from LLaVA is taken from the paper, and for MiniGPT-4, we utilize its official demo.

[输入图像]	
[输入指令]	你知道这幅画是谁画的吗？
[模型响应]	
InstructBLIP	列奥纳多·达·芬奇。
拉瓦	这幅画描绘了一个女人，通常被认为是达芬奇的著名作品蒙娜丽莎。这是一幅肖像画，展示了女人神秘的微笑，已成为世界上最著名和最具标志性的艺术作品之一。原作在巴黎卢浮宫博物馆展出，以其错综复杂的细节、油画颜料的使用以及艺术家的创新技术而闻名，这些技术促成了其持久的吸引力和神秘感。
迷你GPT-4	这幅画是达芬奇的《蒙娜丽莎》，也被称为蒙娜丽莎。它是世界上最著名的画作之一，以其神秘的微笑和坐着者的神秘身份而闻名。这幅画收藏在法国巴黎的卢浮宫博物馆。

图 6：在这个例子中，我们说明了长格式的响应并不总是更可取的。InstructBLIP 可以通过自适应调整响应长度来直接解决用户的意图，而其他模型往往会生成较长的段落，句子相关性较低。LLaVA 的回应取自论文，对于 MiniGPT-4，我们使用了它的官方演示。

[Input Image]

[Input Instruction] Introduce me this painting in detail.

[Model Response]

InstructBLIP The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.

LLaVA The painting features a woman with a headscarf, possibly of Dutch origin, dressed in blue and white clothing. She is wearing a large blue hat and has a red mouth, which adds a vibrant touch to the portrait. The woman is looking directly at the viewer, creating a sense of connection. The background of the painting is predominantly blue, further emphasizing the color theme of the piece. The overall composition and attire of the woman suggest that the painting might be inspired by the style of Rembrandt.

MiniGPT-4 The painting depicts a young woman with a pearl earring, wearing a blue dress and a white apron. She is shown with a serious expression, looking directly at the viewer. The background is a dark, neutral tone, with no visible objects or elements. The overall mood of the painting is somber and contemplative.
The painting is well-executed, with a high level of detail and realism in the depiction of the woman's features and clothing. The use of shading and lighting creates a sense of depth and dimension, making the woman appear three-dimensional. The artist has used a limited color palette, with the woman's dress and apron providing the only sources of color in the painting.
The painting is well-composed, with the woman's head and shoulders positioned in the center of the canvas, creating a sense of balance and symmetry. The use of negative space around the woman's figure adds to the overall somber mood of the painting.
Overall, this is a well-executed and thought-provoking painting that effectively conveys a sense of contemplation and introspection.

Figure 7: InstructBLIP shows the ability to connect textual knowledge with visual input, while the other models depict the picture plainly. MiniGPT-4 exhibits poorer results, which may be due to its training with only long captions. Responses of LLaVA and MiniGPT-4 are generated by their official demos.

[输入图像]

[输入指令]详细介绍一下这幅画。

[模型响应]

InstructBLIP 这幅画描绘了一个戴珍珠耳环的年轻女孩，也被称为“戴珍珠耳环的女孩”或“戴珍珠耳环的女孩”。她穿着一件蓝色的连衣裙，左耳上戴着珍珠耳环。她的头发向后挽成一个发髻，直视着观众。这幅画是荷兰画家约翰内斯·维米尔（Johannes Vermeer）于 1665 年创作的著名艺术作品。它被认为是他的杰作之一，并因其错综复杂的细节和对年轻女孩面部表情的逼真刻画而广受认可。

拉瓦 这幅画描绘了一位戴头巾的女人，可能是荷兰血统的，身着蓝白相间的衣服。她戴着一顶蓝色的大帽子，嘴巴红红的，为肖像增添了一丝生机勃勃的气息。女人直视着观众，营造出一种联系感。这

这幅画的背景主要是蓝色，进一步强调了这幅作品的色彩主题。这位女性的整体构图和着装表明，这幅画的灵感可能来自伦勃朗的风格。

迷你GPT-4 这幅画描绘了一位戴着珍珠耳环的年轻女子，穿着蓝色连衣裙和白色围裙。她表情严肃，直视着观众。背景是深色的中性色调，没有可见的物体或元素。这幅画的整体情绪是阴郁和沉思的。这幅画做得很好，在描绘女性的五官和服装方面具有高度的细节和真实感。阴影和照明的使用创造了一种深度和维度感，使女性看起来立体。艺术家使用了有限的调色板，女人的裙子和围裙是画中唯一的色彩来源。这幅画构图很好，女人的头和肩膀位于画布的中央，营造出一种平衡和对称的感觉。在女性形象周围使用负空间增加了这幅画的整体阴郁情绪。总的来说，这是一幅制作精良且发人深省的画作，有效地传达了一种沉思和内省的感觉。

图 7：InstructBLIP 显示了将文本知识与视觉输入联系起来的能力，而其他模型则清楚地描绘了图片。MiniGPT-4 表现出较差的结果，这可能是由于它的训练只有长字幕。LLaVA 和 MiniGPT-4 的响应由他们的官方演示生成。

C Instruction Tuning Datasets

Dataset Name	Held-out	Dataset Description
COCO Caption [23]	✗	We use the large-scale COCO dataset for the image captioning task. Specifically, Karpathy split [17] is used, which divides the data into 82K/5K/5K images for the train/val/test sets.
Web CapFilt	✗	14M image-text pairs collected from the web with additional BLIP-generated synthetic captions, used in BLIP [21] and BLIP-2 [20].
NoCaps [3]	✓ (val)	NoCaps contains 15,100 images with 166,100 human-written captions for novel object image captioning.
Flickr30K [51]	✓ (test)	The Flickr30k dataset consists of 31K images collected from Flickr, each image has five ground truth captions. We use the test split as the held-out which contains 1K images.
TextCaps [38]	✗	TextCaps is an image captioning dataset that requires the model to comprehend and reason the text in images. Its train/val/test sets contain 21K/3K/3K images, respectively.
VQAv2 [11]	✗	VQAv2 is dataset for open-ended image question answering. It is split into 82K/40K/81K for train/val/test.
VizWiz [12]	✓ (test-dev)	A dataset contains visual questions asked by people who are blind. 8K images are used for the held-out evaluation.
GQA [16]	✓ (test-dev)	GQA contains image questions for scene understanding and reasoning. We use the balanced test-dev set as held-out.
Visual Spatial Reasoning	✓ (test)	VSR is a collection of image-text pairs, in which the text describes the spatial relation of two objects in the image. Models are required to classify true/false for the description. We use the zero-shot data split given in its official github repository.
IconQA [29]	✓ (test)	IconQA measures the abstract diagram understanding and comprehensive cognitive reasoning abilities of models. We use the test set of its multi-text-choice task for held-out evaluation.
OKVQA [30]	✗	OKVQA contains visual questions that require outside knowledge to answer. It has been split into 9K/5K for train and test.
A-OKVQA [36]	✗	A-OKVQA is a successor of OKVQA with more challenging and diverse questions. It has 17K/1K/6K questions for train/val/test.
ScienceQA [28]	✓ (test)	ScienceQA covers diverse science topics with corresponding lectures and explanations. In out settings, we only use the part with image context (IMG).
Visual Dialog [8]	✓ (val)	Visual dialog is a conversational question answering dataset. We use the val split as the held-out, which contains 2,064 images and each has 10 rounds.
OCR-VQA [31]	✗	OCR-VQA contains visual questions that require models to read text in the image. It has 800K/100K/100K for train/val/test, respectively.
TextVQA [39]	✓ (val)	TextVQA requires models to comprehend visual text to answer questions.
HatefulMemes [18]	✓ (val)	A binary classification dataset to justify whether a meme contains hateful content.
LLaVA-Instruct-150K [25]	✗	An instruction tuning dataset which has three parts: detailed caption (23K), reasoning (77K), conversation (58K).
MSVD-QA [47]	✓ (test)	We use the test set (13K video QA pairs) of MSVD-QA for held-out testing.
MSRVTT-QA [47]	✓ (test)	MSRVTT-QA has more complex scenes than MSVD, with 72K video QA pairs as the test set.
iVQA [49]	✓ (test)	iVQA is a video QA dataset with mitigated language biases. It has 6K/2K/2K samples for train/val/test.

Table 4: Description of datasets in our held-in instruction tuning and held-out zero-shot evaluations.

C 指令调优数据集

数据集名称	保留	数据集描述
COCO 标题 [23]	✗	我们使用大规模 COCO 数据集进行图像描述任务。具体来说，使用了 Karpathy 分裂 [17]，它将数据划分为 train/val/test 集的 82K/5K/5K 图像。
Web CapFilt	✗	从网络收集的 14M 图像文本对，带有额外的 BLIP 生成的合成字幕，用于 BLIP [21] 和 BLIP-2 [20]。
无帽 [3]	✓ （val）	NoCaps 包含 15,100 张图像和 166,100 个人工编写的字幕，用于新颖的对象图像字幕。
Flickr30K的 [51]	✓ （测试）	Flickr30k 数据集由从 Flickr 收集的 31K 张图像组成，每张图像有 5 个真实字幕。我们使用测试拆分作为包含 1K 图像的保留。
文本大小写 [38]	✗	TextCaps 是一个图像描述数据集，需要模型理解和推理图像中的文本。它的 train/val/test 集分别包含 21K/3K/3K 图像。
VQAv2 [11]	✗	VQAv2 是用于开放式图像问答的数据集。它分为 82K/40K/81K，用于训练/评估/测试。
可视化 [12]	✓ （test-dev）	数据集包含盲人提出的视觉问题。8K 图像用于保留评估。
GQA [16]	✓ （test-dev）	GQA 包含用于场景理解和推理的图像问题。我们使用平衡的 test-dev 集作为 hold-out。
视觉空间推理	✓ （test）	VSR 是图像-文本对的集合，其中文本描述了 图像中的两个对象。模型需要对描述进行分类 true/false。我们使用其官方 github 存储库中给出的零样本数据拆分。
IconQA [29]	✓ （测试）	IconQA 衡量抽象图的理解和全面的认知研究 模型的测振能力。我们使用其多文本选择任务的测试集进行保留评估。
OKVQA [30]	✗	OKVQA 包含需要外部知识才能回答的视觉问题。它已分为 9K/5K 用于训练和测试。
A-OKVQA [36]	✗	A-OKVQA 是 OKVQA 的继任者，具有更具挑战性和多样化的问题。它有 17K/1K/6K 问题用于 train/val/test。
科学质量保证 [28]	✓ （测试）	ScienceQA 涵盖不同的科学主题以及相应的讲座和解释。在 out 设置中，我们只使用具有图像上下文 （IMG） 的零件。
可视对话框 [8]	✓ （val）	Visual dialog 是一个对话式问答数据集。我们使用 val split 作为保留，其中包含 2,064 张图像，每张图像有 10 轮。
OCR-VQA [31]	✗	OCR-VQA 包含需要模型读取图像中文本的视觉问题。它有 800K/100K/100K 分别用于训练/评估/测试。
文本VQA [39]	✓ （val）	TextVQA 要求模型理解视觉文本来回答问题。
HatefulMemes [18]	✓ （val）	一个二进制分类数据集，用于证明模因是否包含仇恨内容。
LLaVA-Instruct-150K [25]	✗	一个指令调优数据集，由三个部分组成：详细字幕 （23K）、推理 （77K）、对话 （58K）。
MSVD-质量保证 [47]	✓ （测试）	我们使用 MSVD-QA 的测试集（13K 视频 QA 对）进行保持测试。
MSRVTT-质量保证 [47]	✓ （测试）	MSRVTT-QA 的场景比 MSVD 更复杂，以 72K 视频 QA 对作为测试集。
iVQA [49]	✓ （测试）	iVQA 是一个减轻了语言偏见的视频 QA 数据集。它有 6K/2K/2K 样本用于训练/评估/测试。

表 4：我们的保留指令调整和保留零镜头评估中的数据集描述。

D Instruction Templates

Task	Instruction Template
Image Captioning	<Image>A short image caption: <Image>A short image description: <Image>A photo of <Image>An image that shows <Image>Write a short description for the image. <Image>Write a description for the photo. <Image>Provide a description of what is presented in the photo. <Image>Briefly describe the content of the image. <Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.
VQA	<Image>{Question} <Image>Question: {Question} <Image>{Question} A short answer to the question is <Image>Q: {Question} A: <Image>Question: {Question} Short answer: <Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Given the image, generate a question whose answer is: {Answer}. Question: <Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:

Table 5: Instruction templates used for transforming held-in datasets into instruction tuning data. For datasets with OCR tokens, we simply add "OCR tokens:" after the image query embeddings.

E Instructions for Zero-shot Inference

We provide instructions used for zero-shot inference. Note that for instructions with options, we separate options with the alphabetical order, e.g. (a) blue (b) yellow (c) pink (d) black.

GQA, VizWiz, iVQA, MSVD, MSRVTT <Image> Question: {} Short answer:

NoCaps, Flickr30k <Image> A short image description:

TextVQA <Image> OCR tokens: {}. Question: {} Short answer:

IconQA <Image> Question: {} Options: {}. Short answer:

ScienceQA <Image> Context: {} Question: {} Options: {}. Answer:

HatefulMemes <Image> This is an image with: "{}" written on it. Is it hateful? Answer:

VSR <Image> Based on the image, is this statement true or false? "{}" Answer:

Visual Dialog <Image> Dialog history: {}\n Question: {} Short answer:

D 指令模板

任务指令模板	
图像 字幕	<img>简短的图片说明： <img>简短的图片描述： <img><img>显示 <img>为图片撰写简短描述的 图片的照片。 <img>为照片编写描述。 <img>提供照片中呈现内容的描述。 <img>简要描述图片的内容。 <img>您能简要解释一下您在图像中看到的内容吗？ <img>您能用几个词来描述一下您在照片中看到的东西吗？<img>请提供图片的简短 描述。 <img>使用语言，提供图像的简短说明。 <img>用几个词来说明图片中发生的事情。
VQA	<img>{问题} <img>问题： {问题} <img>{问题} 问题的简短回答是 <img>Q： {问题} A： <img>问题： {问题} 简短回答： <img>给定图像，回答以下问题不超过三个单词。{问题}<img>根据图像，用简短的答案回答这个问题：{Question}。答案： <img>使用提供的图像回答问题： {问题} 提供尽可能简短的答案： <img>以下问题的答案是什么？ “{Question}” <img>问题 “{Question}” 可以使用图像进行回答。简短的回答是
VQG	<img>给定图像，生成一个答案为：{Answer} 的问题。问题：<img>根据图像，提供一个答案为：{Answer}的问题。问题：<img>给定视觉表示形式，创建一个答案为“{Answer}”的问题。<img>从提供的图像中，制作一个引出答案的问题：{Answer}。问题：<img>考虑图片，提出一个答案为：{Answer}的问题。<img>考虑到图像，生成一个答案为：{Answer} 的问题。问题：

表 5：用于将保留数据集转换为指令调整数据的指令模板。对于具有 OCR 标记的数据集，我们只需要在图像查询嵌入后添加 “OCR 标记：” 。

E 零样本推理说明

我们提供了用于零样本推理的说明。请注意，对于带有选项的说明，我们按字母顺序分隔选项，例如（a） 蓝色 （b） 黄色 （c） 粉红色 （d） 黑色。

GQA， VizWiz， iVQA， MSVD， MSRVTT <img> 问题： {} 简短的回答：

NoCaps， Flickr30k <img> 简短的图片描述：

TextVQA <img> OCR 令牌：{}。问题： {} 简短的回答：

IconQA <img> 问题：{} 选项：{}。简短的回答：

ScienceQA <img> 上下文： {} 问题： {} 选项： {}.答：

HatefulMemes <img> 这是一张写有： “{}” 的图片。这是可恨的吗？答：

VSR <img> 根据图片，这个说法是真是假？ “{}” 答案：

可视对话框 <img> 对话框历史记录： {}\n 问题： {} 简短的回答：