

# Visual Instruction Tuning

## 视觉指令调整

Haotian Liu<sup>1\*</sup>, Chunyuan Li<sup>2\*</sup>, Qingyang Wu<sup>3</sup>,  
Yong Jae Lee<sup>1</sup>

<sup>1</sup> University of Wisconsin-Madison    <sup>2</sup> Microsoft  
Research    <sup>3</sup> Columbia University

<sup>1</sup> 威斯康星大学麦迪逊分校    <sup>2</sup> 微软研究院    <sup>3</sup> 哥伦比亚大学

<https://llava-vl.github.io>

### Abstract 摘要

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has been shown to improve zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. We present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLaVA: Large Language and Vision Assistant, an end-to-end trained large multimodal model that connects a vision encoder and an LLM for generalpurpose visual and language understanding. To facilitate future research on visual instruction following, we construct two evaluation benchmarks with diverse and challenging application-oriented tasks. Our experiments show that LLaVA demonstrates impressive multimodal chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model, and code publicly available.

使用机器生成的指令跟随数据对大型语言模型（LLMs）进行指令调整，已被证明可以提高新任务的零拍能力，但这一想法在多模态领域的探索较少。我们首次尝试使用纯语言 GPT-4 生成多模态语言图像指令跟踪数据。通过对这些生成的数据进行指令调整，我们介绍了 LLaVA：大型语言和视觉助手，这是一个端到端训练有素的大型多模态模型，它连接了视觉编码器和用于通用视觉和语言理解的 LLM。为了促进未来的视觉指令跟踪研究，我们构建了两个评估基准，其中包含多种具有挑战性的面向应用的任务。我们的实验表明，LLaVA 展示了令人印象深刻的多模态聊天能力，有时在未见图像/指令上表现出多模态 GPT-4 的行为，并在合成多模态指令跟随数据集上与 GPT-4 相比获得了 85.1% 的相对分数。在对科学质量保证进行微调后，LLaVA 和 GPT-4 的协同作用达到了 92.53% 的最新准确度。我们公开了 GPT-4 生成的视觉指令调整数据、我们的模型和代码。

### 1 Introduction 1 引言

Humans interact with the world through many channels such as vision and language, as each individual channel has a unique advantage in representing and communicating certain concepts, and thus facilitates a better understanding of the world. One of the core aspirations in artificial intelligence is to develop a general-purpose assistant that can effectively follow multi-modal vision-and-language instructions, aligned with human intent to complete various real-world tasks in the wild [4, 27, 26].

人类通过视觉和语言等多种渠道与世界互动，因为每种渠道在表达和交流某些概念方面都具有独特的优势，从而有助于更好地理解世界。人工智能的核心愿望之一就是开发一种通用助手，它能有效地遵循多模态视觉和语言指令，与人类意图保持一致，在野外完成各种现实世界的任务[4, 27, 26]。

To this end, the community has witnessed an emergent interest in developing language-augmented foundation vision models [27, 16], with strong capabilities in open-world visual understanding such as classification [40, 21, 57, 54, 39], detection [29, 62, 33], segmentation [25, 63, 58] and captioning [50, 28], as well as visual generation and editing [42, 43, 56, 15, 44, 30]. We refer readers to the Computer Vision in the Wild reading list for a more up-to-date literature compilation [12]. In this line of work, each task is solved independently by one single large vision model, with the task instruction implicitly considered in the model design. Further, language is only utilized to describe the image content. While this allows language to play an important role in

mapping visual signals to language semantics-a common channel for human communication, it leads to models that usually have a fixed interface with limited interactivity and adaptability to the user's instructions.

为此，业界开始关注开发语言增强的基础视觉模型[27, 16]，这些模型在开放世界视觉理解（如分类[40, 21, 57, 54, 39]、检测[29, 62, 33]、分割[25, 63, 58]和字幕[50,28]）以及视觉生成和编辑[42, 43, 56, 15, 44, 30]方面具有很强的能力。读者可参阅“野外计算机视觉”（Computer Vision in the Wild）阅读清单，了解更多最新文献汇编[12]。在这一研究领域，每个任务都由一个大型视觉模型独立解决，而任务指令则隐含在模型设计中。此外，语言仅用于描述图像内容。虽然这可以让语言在将视觉信号映射到语言语义（人类交流的常用渠道）方面发挥重要作用，但却导致模型通常只有一个固定的界面，交互性和对用户指令的适应性有限。

Large language models (LLM), on the other hand, have shown that language can play a wider role: a universal interface for a general-purpose assistant, where various task instructions can be explicitly represented in language and guide the end-to-end trained neural assistant to switch to the task of interest to solve it. For example, the recent success of ChatGPT [35] and GPT-4 [36] have demonstrated the power of aligned LLMs in following human instructions, and have stimulated tremendous interest in developing open-source LLMs. Among them, LLaMA [49] is an opensource LLM that matches the performance of GPT-3. Alpaca [48], Vicuna [9], GPT-4-LLM [38]

另一方面，大型语言模型（LLM）表明，语言可以发挥更广泛的作用：作为通用助手的通用界面，各种任务指令可以明确地用语言表示，并引导端到端训练有素的神经助手切换到感兴趣的任务来解决它。例如，最近ChatGPT [35] 和 GPT-4 [36] 的成功证明了对齐 LLMs 在遵循人类指令方面的强大功能，并激发了人们对开发开源 LLMs 的极大兴趣。其中，LLaMA [49] 是一个开源 LLM，其性能与 GPT-3 不相上下。Alpaca [48], Vicuna [9], GPT-4-LLM[38]。

utilize various machine-generated high-quality instruction-following samples to improve the LLM's alignment ability, reporting impressive performance compared with proprietary LLMs. Importantly, this line of work is text-only.

利用各种机器生成的高质量指令跟随样本来提高LLM的对齐能力，与专有LLMs相比，其性能令人印象深刻。重要的是，这项工作只涉及文本。

In this paper, we present visual instruction-tuning, the first attempt to extend instruction-tuning to the language-image multimodal space, to pave the way towards building a general-purpose visual assistant. In particular, our paper makes the following contributions:

在本文中，我们提出了视觉指令调谐，首次尝试将指令调谐扩展到语言-图像多模态空间，为构建通用视觉助手铺平道路。我们的论文尤其有以下贡献：

Multimodal instruction-following data. One key challenge is the lack of vision-language instruction-following data. We present a data reformation perspective and pipeline to convert image-text pairs into an appropriate instruction-following format, using ChatGPT/GPT-4.

多模态教学跟踪数据。其中一个主要挑战是缺乏视觉语言教学数据。我们提出了一种数据重组观点和管道，利用 ChatGPT/GPT-4 将图像-文本对转换为适当的指令-跟读格式。

Large multimodal models. We develop a large multimodal model (LMM), by connecting the open-set visual encoder of CLIP [40] with the language decoder Vicuna [9], and fine-tuning end-to-end on our generated instructional vision-language data. Our empirical study validates the effectiveness of using generated data for LMM instruction-tuning, and suggests practical tips for building a general-purpose instruction-following visual agent. When ensembled with GPT-4, our approach achieves SoTA on the Science QA [34] multimodal reasoning dataset.

大型多模态模型。我们开发了一个大型多模态模型（LMM），将 CLIP [40] 的开放集视觉编码器与 Vicuna [9] 的语言解码器连接起来，并根据生成的视觉语言教学数据进行端到端的微调。我们的实证研究验证了使用生成数据进行 LMM 指令调整的有效性，并为构建通用指令跟随视觉代理提出了实用建议。当与 GPT-4 组合时，我们的方法在科学 QA [34] 多模态推理数据集上实现了 SoTA。

Multimodal instruction-following benchmark. We present LLaVA-Bench with two challenging benchmarks, with a diverse selection of paired images, instructions and detailed annotations.

多模态指令跟踪基准。我们为 LLaVA-Bench 提供了两个具有挑战性的基准，其中包含多种配对图像、指令和详细注释。

Open-source. We release the following assets to the public: the generated multimodal instruction data, the codebase, the model checkpoints, and a visual chat demo.

开源。我们向公众发布以下资产：生成的多模态指令数据、代码库、模型检查点和可视化聊天演示。

## 2 Related Work 2 相关工作

**Multimodal Instruction-following Agents.** In computer vision, existing works that build instruction-following agents can be broadly categorized into two classes: (i) End-to-end trained models, which are separately explored for each specific research topic. For example, the vision-language navigation task [3,19] and Habitat [47] require the embodied AI agent to follow natural language instructions and take a sequence of actions to complete goals in visual environments. In the image editing domain, given an input image and a written instruction that tells the agent what to do, InstructPix 2Pix [6] edits images by following the human instructions. (ii) A system that coordinates various models via LangChain [1] / LLMs [35], such as Visual ChatGPT [53], X-GPT [63], MM-REACT [55], VisProg [18], and ViperGPT [46]. While sharing the same goal in building instruction-following agents, we focus on developing an end-to-end trained language-vision multimodal model for multiple tasks.

多模态指令跟随代理。在计算机视觉领域，建立指令跟随代理的现有工作可大致分为两类：(i) 端到端训练模型，针对每个特定研究课题分别进行探索。例如，视觉语言导航任务 [3,19] 和 "栖息地" [47] 要求嵌入式人工智能代理遵循自然语言指令并采取一系列行动来完成视觉环境中的目标。在图像编辑领域，InstructPix 2Pix [6] 给定一幅输入图像和一条书面指令，告诉代理该做什么，代理按照人类指令编辑图像。(ii) 通过 LangChain [1] / LLMs [35] 协调各种模型的系统，如 Visual ChatGPT [53]、X-GPT [63]、MM-REACT [55]、VisProg [18] 和 ViperGPT [46]。虽然在构建指令跟随代理方面有着相同的目标，但我们的重点是为多种任务开发端到端训练有素的语言-视觉多模态模型。

**Instruction Tuning.** In the natural language processing (NLP) community, to enable LLMs such as GPT-3 [7], T5 [41], PaLM [10], and OPT [60] to follow natural language instructions and complete real-world tasks, researchers have explored methods for LLM instruction-tuning [37, 52, 51], leading to instruction-tuned counterparts such as InstructGPT [37]/ChatGPT [35], FLAN-T5 [11], FLAN-PaLM [11], and OPT-IML [22], respectively. It turns out that this simple approach can effectively improve the zero- and few-shot generalization abilities of LLMs. It is thus natural to borrow the idea from NLP to computer vision. More broadly, the teacher-student distillation ideas with foundation models have been studied in other topics such as image classification [14]. Flamingo [2] can be viewed as the GPT-3 moment in the multimodal domain, due to its strong performance on zero-shot task transfer and in-context-learning. Other LMMs trained on imagetext pairs include BLIP-2 [28], FROMAGe [24], and KOSMOS-1 [20]. PaLM-E [13] is an LMM for embodied AI. Based on the recent "best" open-source LLM LLaMA, OpenFlamingo [5] and LLaMA-Adapter [59] are open-source efforts that enable LLaMA to use image inputs, paving the way to build open-source multimodal LLMs. While these models present promising task transfer generalization performance, they are not explicitly tuned with vision-language instruction data, and their performance in multimodal tasks usually falls short compared to language-only tasks. In this paper, we aim to fill this gap and study its effectiveness. Finally, note that visual instruction tuning is different from visual prompt tuning [23]: the former aims to improve the model's instruction-following abilities, while the latter aims to improve the parameter-efficiency in model adaptation.

指令调整。在自然语言处理 (NLP) 领域，为了使 GPT-3 [7]、T5 [41]、PaLM [10] 和 OPT [60] 等 LLMs 遵循自然语言指令并完成实际任务，研究人员已经探索了 LLM 指令调整的方法 [37, 52, 51]，分别产生了 InstructGPT [37]/ChatGPT [35]、FLAN-T5 [11]、FLAN-PaLM [11] 和 OPT-IML [22] 等指令调整对应物。事实证明，这种简单的方法可以有效提高 LLMs 的零点泛化和少点泛化能力。因此，将这一想法从 NLP 借鉴到计算机视觉是很自然的。从更广泛的意义上讲，在图像分类 [14] 等其他课题中也研究了师生提炼基础模型的思路。Flamingo [2] 可被视为多模态领域的 GPT-3 moment，因为它在零镜头任务转移和上下文学习方面表现出色。其他在图像文本对上训练的 LMM 包括 BLIP-2 [28]、FROMAGe [24] 和 KOSMOS-1 [20]。PaLM-E [13] 是一种用于具身人工智能的 LMM。基于最近的 "最佳" 开源 LLM LLaMA，OpenFlamingo [5] 和 LLaMA-Adapter [59] 是使 LLaMA 能够使用图像输入的开源努力，为构建开源多模态 LLMs 铺平了道路。虽然这些模型具有良好的任务转移泛化性能，但它们并没有使用视觉语言指令数据进行明确调整，而且与纯语言任务相比，它们在多模态任务中的性能通常不尽如人意。本文旨在填补这一空白并研究其有效性。最后要注意的是，视觉指令调整不同于视觉提示调整 [23]：前者旨在提高模型的指令跟随能力，而后者旨在提高模型适应的参数效率。

## 3 GPT 辅助视觉指令数据生成

The community has witnessed a surge in the amount of public multimodal data such as image-text pairs, ranging from CC [8] to LAION [45]. However, when it comes to multimodal instruction-

从 CC [8] 到 LAION [45]，社会各界见证了图像-文本对等公共多模态数据数量的激增。然而，当涉及到多模态教学时

## Context type 1: Captions 语境类型 1：字幕

A group of people standing outside of a black vehicle with various luggage Luggage surrounds a vehicle in an underground parking area

一群人站在一辆装有各种行李的黑色汽车外 行李围绕着地下停车场的一辆汽车

People try to fit all of their luggage in an SUV.

人们试图把所有行李都塞进越野车。

The sport utility vehicle is parked in the public garage, being packed for a trip Some people with luggage near a van that is transporting it.

运动型多用途车停在公共车库里，正在收拾行李准备出行 一些人带着行李靠近一辆正在运送行李的面包车。

## Context type 2: Boxes 背景类型 2：盒子

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

人: [0.681, 0.242, 0.774, 0.694] , 背包: [0.384, 0.696, 0.485, 0.914] , 手提箱: .....<注释>

## Response type 1: conversation

## 回应类型 1：对话

Question: What type of vehicle is featured in the image?

问题图片中出现的是什么类型的车辆？

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

请回答：图片中是一辆黑色运动型多用途车 (SUV) ...<omitted>

## Response type 2: detailed description

## 答复类型 2：详细说明

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

画面是一个地下停车场，停着一辆黑色运动型多用途车 (SUV)。场景中有三个人，其中一个人站在离车左侧较近的地方，另一个人站在中间，第三个人站在右侧。他们正在一起将行李打包到 SUV 中，准备出行。...%3已提交>

## Response type 3: complex reasoning

答复类型 3：复杂推理

Question: What challenges do these people face?

问题这些人面临哪些挑战？

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

请回答：画面中，一群人站在停车场的一辆黑色 SUV 车外，周围摆满了各种行李，包括手提箱和背包。他们正面临着将所有行李装进黑色 SUV 的挑战。有多个行李箱和背包需要打包，这表明这群人有大量的物品...<omitted>

Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.

表 1：说明指令遵循数据的一个示例。上图显示了用于提示 GPT 的标题和方框等语境，下图显示了三种类型的回答。请注意，视觉图像并不是用来提示 GPT 的，我们在此只是将其作为参考。

following data, the available amount is limited, partially because the process for creating such data is time-consuming and less well-defined when human crowd-scouring is considered. Inspired by the success of recent GPT models in text-annotation tasks [17], we propose to leverage ChatGPT/GPT-4 for multimodal instruction-following data collection, based on the widely existing image-pair data.

由于创建此类数据的过程非常耗时，而且在考虑到人类人群搜索的情况下，此类数据的定义也不够明确，因此可用数据量非常有限。受近期 GPT 模型在文本标注任务中取得成功的启发 [17]，我们建议利用 ChatGPT/GPT-4 收集多模态指令跟随数据，并以广泛存在的图像对数据为基础。

For an image  $\mathbf{X}_v$  and its associated caption  $\mathbf{X}_c$ , it is natural to create a set of questions  $\mathbf{X}_q$  with the intent to instruct the assistant to describe the image content. We prompt GPT-4 to curate such a list of questions (see details in Appendix). Therefore, a simple way to expand an image-text pair to its instruction-following version is Human :  $\mathbf{X}_q \mathbf{X}_v < \text{STOP} >$  Assistant :  $\mathbf{X}_c < \text{STOP} >$ . Though cheap to construct, this simple expanded version lacks diversity and in-depth reasoning in both the instructions and responses.

对于图像  $\mathbf{X}_v$  及其相关标题  $\mathbf{X}_c$ ，自然可以创建一组问题  $\mathbf{X}_q$ ，以指示助手描述图像内容。我们提示 GPT-4 整理这样一个问题列表（详见附录）。因此，将图像-文本对扩展为其指令遵循版本的简单方法是 人类： $\mathbf{X}_q \mathbf{X}_v < \text{停止} >$  助手： $\mathbf{X}_c < \text{STOP} >$ 。这种简单的扩展版本虽然构建成本低，但在指令和回复中缺乏多样性和深入的推理。

To mitigate this issue, we leverage language-only GPT-4 or ChatGPT as the strong teacher (both accept only text as input), to create instruction-following data involving visual content. Specifically, in order to encode an image into its visual features to prompt a text-only GPT, we use two types of symbolic representations: (i) Captions typically describe the visual scene from various perspectives; (ii) Bounding boxes usually localize the objects in the scene, and each box encodes the object concept and its spatial location. One example is shown in the top block of Table 14.

为了缓解这一问题，我们利用纯语言 GPT-4 或 ChatGPT 作为强大的教师（两者都只接受文本作为输入），来创建涉及视觉内容的指令跟踪数据。具体来说，为了将图像编码为视觉特征以提示纯文本 GPT，我们使用了两类符号表示：(i) 标题通常从不同角度描述视觉场景；(ii) 边框通常定位场景中的物体，每个框编码物体概念及其空间位置。表 14 顶部的图块就是一个例子。

This symbolic representation allows us to encode the image as an LLM-recognizable sequence. We use COCO images [31] and generate three types of instruction-following data. One example per type is shown in the bottom block of Table 14. For each type, we first manually design a few examples. They are the only human annotations we have during data collection, and are used as seed examples in in-context-learning to query

通过这种符号表示法，我们可以将图像编码为LLM可识别序列。我们使用 COCO 图像 [31] 并生成三种类型的指令跟随数据。表 14 底部显示了每种类型的一个示例。对于每种类型，我们首先手动设计了几个示例。它们是我们在数据收集过程中唯一的人工注释，在上下文学习中用作种子示例，以查询 GPT-4。

Conversation. We design a conversation between the assistant and a person asking questions about this photo. The answers are in a tone as if the assistant is seeing the image and answering the question. A diverse set of questions are asked about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects. Only questions that have definite answers are considered. Please see Appendix for the detailed prompt.

对话。我们设计了一段对话，对话内容是助理和一个就这张照片提问的人之间的对话。回答者的语气就像助手看到图片并回答问题一样。我们会就图片的视觉内容提出一系列不同的问题，包括物体类型、物体数量、物体动作、物体位置、物体之间的相对位置等。只有有明确答案的问题才会被考虑。详细提示见附录。

Detailed description. To include a rich and comprehensive description for an image, we create a list of questions with such an intent. We prompt GPT-4 then curate the list (see detailed prompts

详细描述。为了对图像进行丰富而全面的描述，我们创建了一个具有这种意图的问题列表。我们会提示 GPT-4，然后对列表进行整理（请参阅详细提示

and curation process in Appendix). For each image, we randomly sample one question from the list to ask GPT-4 to generate the detailed description.

和整理过程见附录）。对于每张图片，我们从列表中随机抽取一个问题询问 GPT-4，以生成详细描述。

Complex reasoning. The above two types focus on the visual content itself, based on which we further create in-depth reasoning questions. The answers typically require a step-by-step reasoning process by following rigorous logic.

复杂推理。上述两类问题主要针对视觉内容本身，在此基础上，我们会进一步创建深入的推理问题。答案通常需要按照严密的逻辑进行逐步推理。

We collect 158 K unique language-image instruction-following samples in total, including 58 K in conversations, 23 K in detailed description, and 77 k in complex reasoning, respectively. We ablated the use of ChatGPT and GPT-4 in our early experiments, and found that GPT-4 consistently provides higher quality instruction-following data, such as spatial reasoning.

我们总共收集了 158 K 个独特的语言图像指令跟踪样本，包括 58 K 个对话样本、23 K 个详细描述样本和 77 K 个复杂推理样本。我们在早期实验中取消了 ChatGPT 和 GPT-4 的使用，结果发现 GPT-4 始终能提供更高质量的指令跟踪数据，如空间推理。

## 4 Visual Instruction Tuning

### 4 视觉指令调整

#### 4.1 Architecture 4.1 结构

The primary goal is to effectively leverage the capabilities of both the pre-trained LLM and visual model. The network architecture is illustrated in Figure 1. We choose Vicuna [9] as our LLM  $f_\phi(\cdot)$  parameterized by  $\phi$ , as it has the best instruction following capabilities in language tasks among publicly available checkpoints [48, 9, 38].

主要目标是有效利用预训练 LLM 和视觉模型的功能。网络架构如图 1 所示。我们选择 Vicuna [9] 作为由  $\phi$  参数化的 LLM  $f_\phi(\cdot)$ ，因为它在公开可用的检查点 [48, 9, 38] 中具有最佳的语言任务指令跟踪能力。

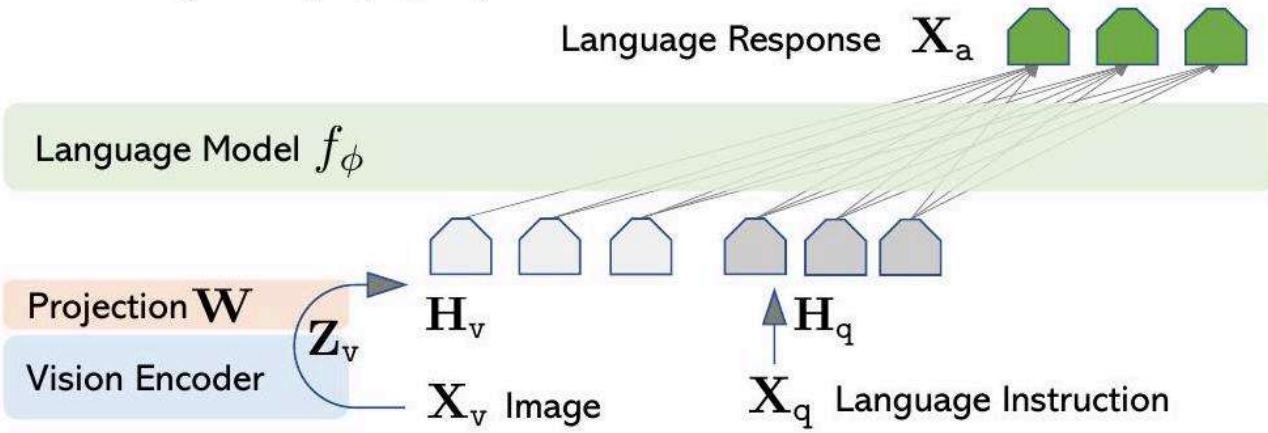


Figure 1: LLaVA network architecture.

图 1：LLaVA 网络架构。

For an input image  $\mathbf{X}_v$ , we consider the pre-trained CLIP visual encoder ViT-L/14 [40], which provides the visual feature  $\mathbf{Z}_v = g(\mathbf{X}_v)$ . The grid features before and after the last Transformer layer are considered in our experiments. We consider a simple linear layer to connect image features into the word embedding space. Specifically, we apply a trainable projection matrix  $\mathbf{W}$  to convert  $\mathbf{Z}_v$  into language embedding tokens  $\mathbf{H}_v$ , which have the same dimensionality as the word embedding space in the language model:

对于输入图像  $\mathbf{X}_v$ ，我们考虑使用预先训练好的 CLIP 视觉编码器 ViT-L/14 [40]，它提供了视觉特征  $\mathbf{Z}_v = g(\mathbf{X}_v)$ 。我们在实验中考虑了最后一个变换器层之前和之后的网格特征。我们考虑用简单的线性层将图像特征连接到词嵌入空间。具体来说，我们应用可训练的投影矩阵  $\mathbf{W}$  将  $\mathbf{Z}_v$  转换为语言嵌入标记  $\mathbf{H}_v$ ，其维度与语言模型中的单词嵌入空间相同：

$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

Thus, we have a sequence of visual tokens  $\mathbf{H}_v$ . Note that our simple projection scheme is lightweight, which allows us to iterate data centric experiments quickly. More sophisticated schemes to connect the image and language representations can also be considered, such as gated cross-attention in Flamingo [2] and Q-former in BLIP-2 [28]. We leave exploring possibly more effective and sophisticated architecture designs for LLaVA as future work.

这样，我们就得到了视觉标记  $\mathbf{H}_v$  的序列。请注意，我们的简单投影方案是轻量级的，这使我们能够快速迭代以数据为中心的实验。我们还可以考虑采用更复杂的方案来连接图像和语言表征，例如 Flamingo [2] 中的门控交叉注意和 BLIP-2 [28] 中的 Q-former。我们将在未来的工作中探索更有效、更复杂的 LLaVA 架构设计。

#### 4.2 Training 4.2 培训

For each image  $\mathbf{X}_v$ , we generate multi-turn conversation data  $(\mathbf{X}_q^1, \mathbf{X}_a^1, \dots, \mathbf{X}_q^T, \mathbf{X}_a^T)$ , where  $T$  is the total number of turns. We organize them as a sequence, by treating all answers as the assistant's response, and the instruction  $\mathbf{X}_{\text{instruct}}^t$  at the  $t$ -th turn as:

对于每个图像  $\mathbf{X}_v$ ，我们都会生成多回合对话数据  $(\mathbf{X}_q^1, \mathbf{X}_a^1, \dots, \mathbf{X}_q^T, \mathbf{X}_a^T)$ ，其中  $T$  是总回合数。我们将所有回答都视为助手的回应，并将第  $t$  个回合的指令  $\mathbf{X}_{\text{instruct}}^t$  视为一个序列：

$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], \text{ the first turn } t=1 \\ \mathbf{X}_q^t, \quad \text{the remaining turns } t > 1 \end{cases}$$

This leads to the unified format for the multimodal instruction-following sequence illustrated in Table 2. We perform instruction-tuning of the LLM on the prediction tokens, using its original auto-regressive training objective.

这就形成了表 2 中所示的多模态指令-跟读序列的统一格式。我们使用 LLM 的原始自动回归训练目标，对预测标记进行指令调整。

Specifically, for a sequence of length  $L$ , we compute the probability of the target answers  $\mathbf{X}_a$  by:

具体来说，对于长度为  $L$  的序列，我们通过以下方法计算目标答案  $\mathbf{X}_a$  的概率：

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^L p_{\theta}(x_i | \mathbf{X}_v, \mathbf{X}_{\text{instruct}, < i}, \mathbf{X}_{a, < i})$$

$\mathbf{X}_{\text{system-message}} < \text{STOP} >$

Human :  $\mathbf{X}_{\text{instruct}}^1 < \text{STOP} >$  Assistant:  $\mathbf{X}_a^1 < \text{STOP} >$

人类:  $\mathbf{X}_{\text{instruct}}^1$

Human :  $\mathbf{X}_{\text{instruct}}^2 < \text{STOP} >$  Assistant:  $\mathbf{X}_a^2 < \text{STOP} > \dots$

人类:  $\mathbf{X}_{\text{instruct}}^2$

Table 2: The input sequence used to train the model. Only two conversation turns are illustrated here; in practice, the number of turns varies based on the instruction-following data. In our current implementation, we follow Vicuna-vo [9] to set the system message  $\mathbf{X}_{\text{system-message}}$  and we set  $<\text{STOP}> = \#\#\#$ . The model is trained to predict the assistant answers and where to stop, and thus only green sequence/tokens are used to compute the loss in the auto-regressive model.

表 2：用于训练模型的输入序列。这里只说明了两个对话回合；实际上，回合数会根据指令跟随数据而变化。在我们目前的实现中，我们按照 Vicuna-vo [9] 来设置系统信息  $\mathbf{X}_{\text{system-message}}$ ，并设置

where  $\theta$  is the trainable parameters,  $\mathbf{X}_{\text{instruct}, < i}$  and  $\mathbf{X}_{a, < i}$  are the instruction and answer tokens in all turns before the current prediction token  $x_i$ , respectively. Please see Table 2 for an illustration of the prediction tokens. For the conditionals in (3), we explicitly add  $\mathbf{X}_v$  to emphasize the fact that the image is grounded for all answers, and we omit  $\mathbf{X}_{\text{system-message}}$  and all previous  $<\text{STOP}>$  for better readability. For LLaVA model training, we consider a two-stage instruction-tuning procedure.

其中  $\theta$  是可训练参数， $\mathbf{X}_{\text{instruct}, < i}$  和  $\mathbf{X}_{a, < i}$  分别是当前预测标记  $x_i$  之前所有回合中的指令和答案标记。有关预测标记的说明，请参见表 2。对于(3)中的条件句，我们明确添加了  $\mathbf{X}_v$ ，以强调图像对所有答案都是有根据的这一事实，并且我们省略了  $\mathbf{X}_{\text{system-message}}$  和之前所有的

Stage 1: Pre-training for Feature Alignment. To strike a balance between concept coverage and training efficiency, we filter CC3M to 595K image-text pairs. Please see Appendix for details of the filtering process. These pairs are converted to the instruction-following data using the naive expansion method described in Section 3. Each sample can be treated as a single-turn conversation. To construct the input  $\mathbf{X}_{\text{instruct}}$  in (2), for an image  $\mathbf{X}_v$ , a question  $\mathbf{X}_q$  is randomly sampled, which is a language instruction to request the assistant to describe the image briefly. The ground-truth prediction answer  $\mathbf{X}_a$  is the original caption. In training, we keep both the visual encoder and LLM weights frozen, and maximize the likelihood of (3) with trainable parameters  $\theta = \mathbf{W}$  (the projection matrix) only. In this way, the image features  $\mathbf{H}_v$  can be aligned with the pre-trained LLM word embedding. This stage can be understood as training a compatible visual tokenizer for the frozen LLM.

第 1 阶段：特征对齐的预训练。为了在概念覆盖率和训练效率之间取得平衡，我们将 CC3M 筛选为 595K 个图像-文本对。有关过滤过程的详情，请参阅附录。使用第 3 节中描述的天真扩展方法将这些图像对转换为指令跟随数据。每个样本都可视为一次单轮对话。为了构建(2)中的输入  $\mathbf{X}_{\text{instruct}}$ ，对于图像  $\mathbf{X}_v$ ，会随机抽取一个问题  $\mathbf{X}_q$ ，这是一个要求助手简要描述图像的语言指令。地面真实预测答案  $\mathbf{X}_a$  是原始标题。在训练过程中，我们冻结视觉编码器和LLM权重，仅使用可训练参数  $\theta = \mathbf{W}$ （投影矩阵）最大化(3)的可能性。这样，图像特征  $\mathbf{H}_v$  就可以与预先训练的LLM字嵌入对齐。这一阶段可以理解为为冻结的LLM训练一个兼容的视觉标记器。

Stage 2: Fine-tuning End-to-End. We always keep the visual encoder weights frozen, and continue to update both the pre-trained weights of the projection layer and LLM in LLaVA; i.e., the trainable parameters are  $\theta = \{\mathbf{W}, \phi\}$  in (3). We consider two specific use case scenarios:

第二阶段：端到端的微调。我们始终保持视觉编码器权重的冻结，并继续更新 LLaVA 中投影层和 LLM 的预训练权重；即 (3) 中的可训练参数为  $\theta = \{\mathbf{W}, \phi\}$ 。我们考虑了两个特定的使用场景：

Multimodal Chatbot. We develop a Chatbot by fine-tuning on the 158 K language-image instruction-following data in Section 3. Among the three types of responses, conversation is multi-turn while the other two are single-turn. They are uniformly sampled in training.

多模态聊天机器人。我们通过对第 3 节中的 158 K 语言-图像指令-跟读数据进行微调，开发了一个聊天机器人。在这三种回答类型中，对话是多转向的，而其他两种是单转向的。它们在训练中统一采样。

Science QA. We study our method on the ScienceQA benchmark [34], the first large-scale multimodal science question dataset that annotates the answers with detailed lectures and explanations. Each question is provided a context in the form of natural language or an image. The assistant provides the reasoning process in natural language and selects the answer among multiple choices. For training in (2), we organize the data as a single turn conversation, the question & context as  $\mathbf{X}_{\text{instruct}}$ , and reasoning & answer as  $\mathbf{X}_a$ .

科学 QA。我们在 ScienceQA 基准[34]上研究了我们的方法，该基准是首个大规模多模态科学问题数据集，为答案注释了详细的讲座和解释。每个问题都有一个自然语言或图像形式的上下文。助手用自然语言提供推理过程，并从多个选项中选出答案。对于 (2) 中的训练，我们将数据组织为单轮对话，问题和上下文为  $\mathbf{X}_{\text{instruct}}$ ，推理和答案为  $\mathbf{X}_a$ 。

## 5 Experiments 5 项实验

We assess the performance of LLaVA in instruction-following and visual reasoning capabilities with two primary experimental settings: multimodal chatbot and the ScienceQA dataset, respectively. We train all models with  $8 \times$  A100s, following Vicuna’s hyperparameters [9]. We pre-train our model on the filtered CC-595K subset for 1 epoch with a learning rate of  $2e - 3$  and a batch size of 128, and fine-tune on the proposed LLaVA-Instruct-158K dataset for 3 epochs, with a learning rate of  $2e - 5$  and a batch size of 32. See Appendix for more training details.

我们通过两个主要实验设置来评估 LLaVA 在指令跟随和视觉推理能力方面的表现：多模态聊天机器人和 ScienceQA 数据集。我们按照 Vicuna 的超参数 [9]，使用  $8 \times$  A100s 训练所有模型。我们在经过过滤的 CC-595K 子集上对模型进行了 1 个历元的预训练，学习率为  $2e - 3$ ，批量大小为 128，并在提议的 LLaVA-Instruct-158K 数据集上进行了 3 个历元的微调，学习率为  $2e - 5$ ，批量大小为 32。更多训练详情，请参见附录。

### 5.1 Multimodal Chatbot 5.1 多模态聊天机器人

We developed a chatbot demo to show the image understanding and conversation abilities of LLaVA, and to study how well LLaVA is able to digest visual inputs and exhibit instruction-following capabilities. We first use the examples in the original GPT-4 paper [36], shown in Table 3 (more examples in Appendix), that require in-depth image understanding. For comparisons, we quote the prompt and response of the multimodal GPT-4 from their paper, and query BLIP-2 and OpenFlamingo model checkpoints to get their response.

我们开发了一个聊天机器人演示，以展示 LLaVA 的图像理解和对话能力，并研究 LLaVA 在消化视觉输入和展示指令跟随能力方面的表现。我们首先使用 GPT-4 原文[36]中需要深入理解图像的示例，如表 3 所示（更多示例见附录）。为了便于比较，我们引用了其论文中多模态 GPT-4 的提示和响应，并查询 BLIP-2 和 OpenFlamingo 模型检查点以获得其响应。

Visual input example, Extreme Ironing:

视觉输入示例：极限熨烫：



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

What is unusual about this image?

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.

来源：<https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg> 这幅图片有什么不同寻常之处？这幅图片的不同寻常之处在于，一名男子正在一辆小型货车或面包车的后座上熨烫衣物。这不是进行这项活动的典型场所，因为人们通常会在更固定、更安全的场所（如家中）使用普通熨衣板熨烫衣服。图片中描绘的场景很奇特，因为它涉及到在车辆上临时熨烫衣服，这可能既不安全又不合常规。此外，尚不清楚该男子在如此不稳定的环境中熨烫衣服时是如何保持平衡和稳定的。

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around

the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are

two other people in the vicinity, one close to the man on the car and another further away.

The man on the ladder appears to be attracting attention due to his unconventional choice

of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User  
LLaVA  
用户名

	<p>画面描绘了一名男子站在繁忙城市街道上的一辆黄色 SUV 车顶上。他手持一把便携式梯子，似乎是站在车上熨烫衣服。场景周围有其他汽车、一个交通灯和一个消防栓。此外，附近还有两个人，一个离车上的男子很近，另一个离得较远。梯子上的男子似乎正在吸引人们的注意，因为他选择了在行驶的汽车顶上熨烫衣服这一非常规的方式。他周围的城市街道熙熙攘攘，更增添了这一场景的独特性。</p>
User GPT-4 [36] 用户GPT-4 [36]	<p>What is unusual about this image? 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.</p> <p>这幅图片有什么不同寻常之处？这幅图片的不同寻常之处在于，一名男子正在移动的出租车车顶上熨烫衣服。</p>
User BLIP-2 用户BLIP-2	<p>What is unusual about this image? a man is sitting on the back of a yellow cab</p> <p>这幅图片有何不同寻常之处？</p>
User OpenFlamingo 用户打开火烈鸟	<p>What is unusual about this image? The man is drying his clothes on the hood of his car.</p> <p>这幅图有什么不寻常之处？</p>
	<p>The man is drying his clothes on the hood of his car.</p> <p>这名男子正在汽车引擎盖上晾衣服。</p>

Table 3: Example prompt from GPT-4 paper [36] to compare visual reasoning and chat capabilities. Compared to BLIP-2 [28] and OpenFlamingo [5], LLaVA accurately follows the user’s instructions, instead of simply describing the scene. LLaVA offers a more comprehensive response than GPT-4. Even when merely asked to describe the image, LLaVA identifies atypical aspects of the image.

表 3：GPT-4 论文[36]中的提示示例，用于比较视觉推理和聊天功能。与 BLIP-2 [28] 和 OpenFlamingo [5] 相比，LLaVA 能准确地按照用户的指令进行操作，而不是简单地描述场景。与 GPT-4 相比，LLaVA 提供了更全面的响应。即使只要求描述图像，LLaVA 也能识别图像的非典型方面。

Surprisingly, although LLaVA is trained with a small multimodal instruction-following dataset ( $\sim 80$  K unique images), it demonstrates quite similar reasoning results with multimodal GPT-4 on these examples. Note that while these images are out-of-domain for LLaVA, LLaVA is still able to understand the scenes and follow the question instruction to provide a reasonable response. In contrast, BLIP-2 and OpenFlamingo focus on describing the image, instead of following the user instruction to answer in an appropriate manner.

令人惊讶的是，虽然 LLaVA 是通过一个小型的多模态指令跟踪数据集 ( $\sim 80$  K 唯一图像) 进行训练的，但它在这些示例上的推理结果与多模态 GPT-4 非常相似。请注意，虽然这些图像对 LLaVA 来说是域外图像，但 LLaVA 仍然能够理解这些场景，并根据问题指令做出合理的回答。相比之下，BLIP-2 和 OpenFlamingo 则侧重于描述图像，而不是按照用户指令以适当的方式进行回答。

Quantitative Evaluation. To gain a systematic understanding of the performance of LLaVA, we propose a quantitative metric to measure the model’s instruction-following capability on multimodal data. Inspired by [9], we leverage GPT-4 to measure the quality of generated responses. Specifically, we create triplets consisting of image, ground-truth textual descriptions, and question. The candidate models (e.g., LLaVA) predict the

answers based on the question and the image. To provide an approximate theoretical upper bound, we create a reference prediction based on the question and the ground-truth textual descriptions, using the text-only GPT-4. After obtaining the responses from both models, we feed the question, visual information (in the format of textual descriptions), and the generated responses from both assistants, to the judge (i.e., text-only GPT-4). It evaluates the helpfulness, relevance, accuracy, and level of detail of the responses from the assistants, and gives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. It is also asked to provide a comprehensive explanation for the evaluation, for us to better understand the

定量评估。为了系统地了解 LLaVA 的性能，我们提出了一个量化指标来衡量模型在多模态数据上的指令遵循能力。受 [9] 的启发，我们利用 GPT-4 来衡量生成响应的质量。具体来说，我们创建了由图像、真实文本描述和问题组成的三元组。候选模型（如 LLaVA）根据问题和图像预测答案。为了提供一个近似的理论上限，我们使用纯文本 GPT-4 创建了一个基于问题和真实文本描述的参考预测。在获得两个模型的回复后，我们将问题、视觉信息（以文本描述的形式）和两个助手生成的回复反馈给法官（即纯文本 GPT-4）。它将对助手回答的有用性、相关性、准确性和详细程度进行评估，并按 1-10 分制给出总分，分值越高表示整体表现越好。此外，还要求其对评价进行全面解释，以便我们更好地了解

	Conversation 对话	Detail description 详细说明	Complex reasoning 复杂推理	All 全部
Full data 完整数据	83.1	75.3	96.5	85.1
Detail + Complex 细节 + 复杂	81.5(-1.6)	73.3(-2.0)	90.8(-5.7)	81.9(-3.2)
Conv +5% Detail +10% Complex				
Conv +5% Detail +10% Complex	81.0(-2.1)	68.4(-7.1)	91.5(-5.0)	80.5(-4.4)
Conversation 对话	76.5(-6.6)	59.8(-16.2)	84.9(-12.4)	73.8(-11.3)
No Instruction Tuning 无 指令调谐	22.0(-61.1)	24.0(-51.3)	18.5(-78.0)	21.5(-63.0)

Table 4: Ablation on LLaVA-Bench (COCO) with different training data. We report relative scores w.r.t. a text-only GPT-4 model that uses ground truth image captions and bounding boxes as visual input. We prompt GPT-4 with the answers from our model outputs and the answers by GPT-4 (text-only), and let it compare between both responses and give a rating with an explanation.

表 4：不同训练数据在 LLaVA-Bench (COCO) 上的消融情况。我们报告的是与纯文本 GPT-4 模型相比的相对分数，该模型使用地面实况图像标题和边界框作为视觉输入。我们用模型输出的答案和 GPT-4 的答案（纯文本）提示 GPT-4，让它比较两种答案并给出评分和解释。

	Conversation 对话	Detail description 详细说明	Complex reasoning 复杂推理	All 全部
OpenFlamingo [5]	$19.3 \pm 0.5$	$19.0 \pm 0.5$	$19.1 \pm 0.7$	$19.1 \pm 0.4$
BLIP-2 [28]	$54.6 \pm 1.4$	$29.1 \pm 1.2$	$32.9 \pm 0.7$	$38.1 \pm 1.0$
LLaVA	$57.3 \pm 1.9$	$52.5 \pm 6.3$	$81.7 \pm 1.8$	$67.3 \pm 2.0$
LLaVA <sup>†</sup>	$58.8 \pm 0.6$	$49.2 \pm 0.8$	$81.4 \pm 0.3$	$66.7 \pm 0.3$

Table 5: Instruction-following capability comparison using relative scores on LLaVA-Bench (In-theWild). The results are reported in the format of mean  $\pm$  std. For the first three rows, we report three inference runs. LLaVA performs significantly better than others. <sup>†</sup> For a given set of LLaVA decoding sequences, we evaluate by querying GPT-4 three times; GPT-4 gives a consistent evaluation.

表 5：使用 LLaVA-Bench (In-the-Wild) 上的相对分数进行指令遵循能力比较。结果以均值  $\pm$  std. 在前三行中，我们报告了三次推理运行。LLaVA 的表现明显优于其他推理。<sup>†</sup> 对于一组给定的 LLaVA 解码序列，我们通过查询 GPT-4 进行了三次评估；GPT-4 给出了一致的评估结果。

models. We report relative scores w.r.t. the text-only GPT-4 model that uses the textual ground truth description as visual input. We create two benchmarks to evaluate the model’s performance.

模型的相对分数。我们报告了与纯文本 GPT-4 模型相比的相对分数，后者使用文本地面实况描述作为视觉输入。我们创建了两个基准来评估模型的性能。

LLaVA-Bench (COCO). We randomly select 30 images from COCO-Val-2014, and for each image, we generate three types of questions (conversation, detailed description, complex reasoning) using the proposed data generation pipeline in Sec. 3, totaling 90 questions. This benchmark studies the model’s alignment behavior and capabilities with consistent visual inputs. We vary the training datasets to study the effectiveness of different types of instruction-following data, and show the results in Table 4. First, with instruction tuning, the model’s ability of following user instructions improves significantly by over 50 points. Second, adding a small amount of detailed description and complex reasoning questions contributes to a considerable improvement of the model’s overall capability by 7 points. Furthermore, it also improves the model’s performance on conversational questions, suggesting that improvements in reasoning capabilities complement conversational abilities. Finally, we show that having all three types of data yields the best performance at 85.1%.

LLaVA-Bench (COCO)。我们从 COCO-Val-2014 中随机选取 30 幅图像，并使用第 3 章中提出的数据生成管道为每幅图像生成三种类型的问题（对话、详细描述、复杂推理），共计 90 个问题。该基准测试研究了模型在一致的视觉输入下的对齐行为和能力。我们改变了训练数据集，以研究不同类型的指令跟随数据的有效性，结果如表 4 所示。首先，通过指令调整，模型遵循用户指令的能力显著提高了 50 多点。其次，增加少量的详细描述和复杂推理论题，使模型的整体能力大幅提高了 7 分。此外，它还提高了模型在对话问题上的表现，这表明推理论题的提高与对话能力的提高相辅相成。最后，我们表明，在 85.1% 时，三种类型的数据都能产生最佳性能。

LLaVA-Bench (In-the-Wild). To evaluate the model’s capability in more challenging tasks and generalizability to novel domains, we collect a diverse set of 24 images with 60 questions in total, including indoor and outdoor scenes, memes, paintings, sketches, etc., and associate each image with a highly-detailed and manually-curated description and a proper selection of questions. We compare LLaVA, BLIP, and OpenFlamingo in Table 5. Thanks to visual instruction tuning, LLaVA achieves significantly better performance compared with BLIP-2 (+29%) and OpenFlamingo (+48%). Compared to the text-only GPT-4 that has access to ground-truth labels, LLaVA achieves an impressive 81.7% performance on complex reasoning questions, with an overall score of 67.3%.

LLaVA-Bench (野外)。为了评估该模型在更具挑战性的任务中的能力以及在新领域中的通用性，我们收集了一组共包含 60 个问题的 24 幅图像，其中包括室内和室外场景、备忘录、绘画、素描等，并将每幅图像与人工编辑的高度详细描述和适当选择的问题关联起来。我们在表 5 中对 LLaVA、BLIP 和 OpenFlamingo 进行了比较。与 BLIP-2 (+29%) 和 OpenFlamingo (+48%) 相比，得益于可视化指令调整，LLaVA 取得了明显更好的性能。与可以访问地面实况标签的纯文本 GPT-4 相比，LLaVA 在复杂推理论题上取得了令人印象深刻的 81.7% 性能，总得分为 67.3%。

Limitations. This LLaVA-Bench (In-the-Wild) is designed to be challenging and to reveal a model’s weaknesses. We provide two examples with associated captions and questions in Table 6. For the ramen example (left), to correctly answer the name of the restaurant, it requires the model to have a large knowledge coverage and multilingual understanding capability; to correctly describe the side dishes, the model may need to retrieve relevant multimodal information from Internet. For the fridge example (right), perceiving the correct brand of the yogurt requires the model to process high resolution images and possess extensive knowledge coverage. We also observed an interesting failure of LLaVA, as it responds with yes when asked if strawberry-flavored yogurt is present, even though the fridge contains only yogurt and strawberries. This indicates that, at times, LLaVA perceives the image as a “bag of patches”, failing to grasp the complex semantics within the image. We hope LLaVA serves as a solid baseline on the benchmarks, on which our

findings can inspire future work in developing more capable LMMs.

局限性。本 LLaVA-Bench (野外) 旨在提高挑战性并揭示模型的弱点。我们在表 6 中提供了两个例子以及相关的标题和问题。在拉面例子 (左) 中, 要正确回答餐厅名称, 需要模型具有较大的知识覆盖面和多语言理解能力; 要正确描述配菜, 模型可能需要从互联网上检索相关的多模态信息。在冰箱的例子中 (右图), 感知酸奶的正确品牌需要模型处理高分辨率图像并拥有广泛的知识覆盖面。我们还观察到 LLaVA 的一个有趣失误, 当被问及是否存在草莓味酸奶时, 它的回答是 "是", 尽管冰箱里只有酸奶和草莓。这表明, LLaVA 有时会将图像视为 "一袋补丁", 而无法把握图像中的复杂语义。我们希望 LLaVA 能在基准测试中发挥坚实的基础作用, 我们的研究结果能为今后开发更强大的 LMM 提供启发。

Challenging examples from LLaVA-Bench (In-the-Wild):

来自 LLaVA-Bench (野外) 的挑战性实例:

ICHIRAN Ramen [source] 一兰拉面 [来源]

Table 6: Challenging examples from LLaVA-Bench (In-the-Wild), we provide extremely-detailed annotation for each image for an accurate evaluation. Some questions require the model to extract details from high resolution image and to have a broad knowledge coverage.

表 6: 来自 LLaVA-Bench 的挑战性示例 (野外), 我们为每张图像提供了极为详细的注释, 以便进行准确评估。有些问题要求模型从高分辨率图像中提取细节, 并具有广泛的知识覆盖面。

## 5.2 ScienceQA 5.2 科学质量保证

ScienceQA [34] contains 21 k multimodal multiple choice questions with rich domain diversity across 3 subjects, 26 topics, 127 categories, and 379 skills. The benchmark dataset is split into training, validation, and test splits with 12726, 4241, and 4241 examples, respectively. We consider two representative methods, including GPT-3.5 model (text-davinci-002) with and without chainof-thought (CoT), LLaMA-Adapter [59], as well as multimodal chain-of-thought (MM-CoT) [61], which is the current SoTA method on this dataset. For more baseline numbers, please see [34].

ScienceQA [34] 包含 21 k 道多模态选择题, 涉及 3 个学科、26 个主题、127 个类别和 379 种技能, 具有丰富的领域多样性。基准数据集分为训练、验证和测试三个部分, 分别有 12726、4241 和 4241 个示例。我们考虑了两种具有代表性的方法, 包括带有和不带思维链 (CoT) 的 GPT-3.5 模型 (text-davinci-002)、LLaMA-Adapter [59] 以及多模态思维链 (MM-CoT) [61], 后者是目前该数据集上的 SoTA 方法。更多基线数据, 请参阅 [34]。

The results are reported in Table 7. For LLaVA, we use the visual features before the last layer, ask the model to first predict reasons and then the answer, and train it for 12 epochs. It yields 90.92% accuracy, which is quite close to the SoTA 91.68%. To explore the limit of LLMs, we also prompt GPT-4 using 2-shot in-context-learning and achieve 82.69% accuracy, which is a 7.52% absolute gain compared with 75.17% from GPT-3.5. For a substantial number of questions, we note that GPT-4 fails simply because it reports that there is insufficient context such as images or plots. We consider two schemes to combine the outcomes from our model and GPT-4. (i) A GPT-4 complement. Whenever GPT-4 fails to provide answers, we use the prediction from our method. This scheme yields 90.97% accuracy, which is almost the same as applying our method alone. (ii) GPT-4 as the judge. Whenever GPT-4 and LLaVA produce different answers, we prompt GPT-4 again, asking it to provide its own final answer based on the question and two outcomes. The spirit is similar with CoT, but with the external knowledge from the other model. Surprisingly, this scheme is able to provide consistent improvement over all question classes, and achieves a new SoTA accuracy of 92.53%. Interestingly, the text-only GPT-4, which cannot process images, improves the overall performance of the model on questions that have an image as context. This is because some of these questions do not actually require the image context for a correct answer. The GPT-4 judge can identify such cases and correct some of the errors

that LLaVA makes. See the example in Appendix. To the best of our knowledge,

结果见表 7。对于 LLaVA，我们在最后一层之前使用视觉特征，要求模型首先预测原因，然后预测答案，并对其进行 12 次历时训练。它的准确率为 90.92%，与 SoTA 91.68% 相当接近。为了探索LLMs的极限，我们还使用 2 次情境内学习提示 GPT-4，并获得了 82.69% 的准确率，与 GPT-3.5 中的 75.17% 相比，绝对增益为 7.52%。我们注意到，对于相当多的问题，GPT-4 的失败仅仅是因为它报告说没有足够的上下文，如图片或图表。我们考虑了两种将我们的模型和 GPT-4 的结果结合起来的方案。(i) GPT-4 补充。每当 GPT-4 无法提供答案时，我们就使用我们方法的预测结果。该方案可获得 90.97% 的准确度，与单独使用我们的方法几乎相同。(ii) GPT-4 作为法官。每当 GPT-4 和 LLaVA 得出不同答案时，我们会再次提示 GPT-4，要求它根据问题和两个结果给出自己的最终答案。其精神与 CoT 类似，但使用了来自另一个模型的外部知识。出乎意料的是，该方案能够在所有问题类别中提供一致的改进，并实现了 92.53% 的新 SoTA 准确率。有趣的是，不能处理图像的纯文本 GPT-4 在有图像作为上下文的问题上提高了模型的整体性能。这是因为其中有些问题实际上并不需要图像上下文就能得出正确答案。GPT-4 法官可以识别这种情况，并纠正 LLaVA 所犯的一些错误。参见附录中的示例。据我们所知

Method 方法	Subject 主题			Context Modality 语境模式			Grade 等级		Average 平均	
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12		
Representative & SoTA methods with numbers reported in the literature										
具有代表性的方法和 SoTA 方法，以及文献中报告的数字										
Human [34] 人类 [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40	
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97	
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17	
LLaMA-Adapter [59] LLaMA 适配器 [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19	
MM-CoT Base [61] MM- CoT Base [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91	
MM-CoT Large [61]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68	
Results with our own experiment runs										
我们自己的实验结果										
GPT-4 <sup>†</sup>	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69	
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92	
LLaVA+GPT-4 <sup>†</sup> (complement)										
LLaVA+GPT-4 <sup>†</sup> (补 体)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97	
LLaVA+GPT-4 <sup>†</sup> (judge)										
LLaVA+GPT-4 <sup>†</sup> (法 官)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	92.53	

Table 7: Accuracy (%) on Science QA dataset. Question categories: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. <sup>†</sup> Text-only GPT-4, our eval. Our novel model ensembling with the text-only GPT-4 consistently improves the model's performance under all categories, setting the new SoTA performance.

表 7：科学 QA 数据集的准确率 (%)。问题类别：NAT = 自然科学，SOC = 社会科学，LAN = 语言科学，TXT = 文本上下文，IMG = 图像上下文，NO = 无上下文，G1-6 = 1-6 年级，G7-12 = 7-12 年级。<sup>†</sup>  
纯文本 GPT-4，我们的评估。我们使用纯文本 GPT-4 进行的新颖模型组合持续提高了模型在所有类别下的性能，并创造了新的 SoTA 性能。

this is the first time that GPT-4 is used for model ensembling. We hope this finding can encourage future research to explore more effective methods to leverage LLMs for model ensembling.

这是 GPT-4 首次用于模型集合。我们希望这一发现能鼓励未来的研究，探索更有效的方法来利用LLMs进行模型集合。

Ablations. We ablate several design choices on ScienceQA in Table 8. (i) Visual features. We tried using the last layer feature from CLIP vision encoder, which yields 89.96% and is 0.96% lower than the feature before the last layer. We hypothesize that this is because CLIP’s last layer features may focus more on global and abstract image properties compared to the layer before it, which can focus more on

消融。表 8 列出了我们在 ScienceQA 上消融的几种设计选择。(i) 视觉特征。我们尝试使用 CLIP 视觉编码器的最后一层特征，其结果为 89.96%，0.96% 低于最后一层之前的特征。我们推测，这是因为 CLIP 的最后一层特征可能更侧重于全局和抽象的图像属性，而前一层特征则可以更侧重于

Visual features 视觉特征	Before 之前	Last 最后
Best variant 最佳变体	90.92	89.96
Predict answer first 先预测答案	-	89.77 <sub>(-1.15)</sub>
Training from scratch 从零开始的培训	85.81 <sub>(-5.11)</sub>	-
7B model size 7B 型号尺寸	89.84 <sub>(-1.08)</sub>	-

Table 8: Design choice ablations (%). The difference with the best variant is reported in red text. localized properties that are useful for under-

表 8：设计选择消融率 (%)。与最佳变体的差异用红色文字表示。

standing specific image details. (ii) Chain-of-thought. To decide the order between the answer and reasoning process in the model prediction, we run both variants and observe that answer-first reports the best number 89.77% accuracy in 12 epochs, while reasoning-first can quickly reach 89.77% accuracy in 6 epochs, but no further improvement with more training. Training the model for 24 epochs does not improve the performance. We conclude that CoT-like reasoning-first strategy can largely improve convergence, but contributes relatively little to the final performance. (iii) Pre-training. We skip pre-training and directly train on Science QA from scratch - performance drops to 85.81% accuracy. The 5.11% absolute degradation indicates the importance of our pre-training stage, in aligning multimodal features while preserving the vast pre-trained knowledge. (iv) Model size. We keep all configurations the same as our best 13B model, and train a 7B model. This yields 89.84% accuracy, which is 1.08% lower than 90.92%, demonstrating the importance of model scale.

立具体的图像细节。(ii) 思维链。为了决定模型预测中答案和推理过程的先后顺序，我们运行了这两个变体，观察到答案优先在 12 个历时周期内报告了最佳的 89.77% 准确率，而推理优先在 6 个历时周期内可以快速达到 89.77% 准确率，但随着训练次数的增加没有进一步提高。对模型进行 24 个历时的训练也不会提高性能。我们的结论是，类似于 CoT 的 "推理优先" 策略可以在很大程度上提高收敛性，但对最终性能的贡献相对较小。(iii) 预训练。我们跳过预训练，直接对科学 QA 进行从头开始的训练--性能下降到 85.81% 准确率。5.11% 的绝对下降表明了我们预训练阶段的重要性，即在对齐多模态特征的同时保留大量预训练知识。(iv) 模型大小。我们保持所有配置与最佳 13B 模型相同，并训练一个 7B 模型。这将产生 89.84% 的准确率，1.08% 低于 90.92%，这表明了模型规模的重要性。

## 6 Conclusion 6 结论

This paper demonstrated the effectiveness of visual instruction tuning. We presented an automatic pipeline to create language-image instruction-following data, based on which we train LLaVA, a multimodal model to follow human intent to complete visual tasks. It achieves the new SoTA accuracy when fine-tuned on ScienceQA, and excellent visual chat capabilities when fine-tuned on multimodal chat data. Besides, we present the first benchmark to study multimodal instructionfollowing capability. This paper is an initial step in visual instruction tuning, and mainly focuses on real-life tasks. For more quantitative results of LLaVA on academic benchmarks, please refer to the improved baselines with visual instruction tuning [32]. We hope our work can inspire future research on building more capable multimodal models.

本文展示了视觉指令调整的有效性。我们介绍了一种创建语言-图像指令跟踪数据的自动流水线，在此基础上，我们训练了一个多模态模型 LLaVA，以跟踪人类完成视觉任务的意图。在对 ScienceQA 进行微调时，它达到了新的 SoTA 准确率；在对多模态聊天数据进行微调时，它具有出色的视觉聊天能力。此外，我们还首次提出了研究多模态指令跟随能力的基准。本文是视觉指令调整的第一步，主要关注现实生活中的任务。关于 LLaVA 在学术基准上的更多量化结果，请参阅视觉指令调整的改进基准[32]。我们希望我们的工作能对未来建立更强大的多模态模型的研究有所启发。

Acknowledgements. We thank Baolin Peng and Pan Lu for valuable discussions on instruction-tuning language models and Science QA, respectively. We thank the LLaMA team for giving us access

致谢。我们感谢彭宝林和潘璐分别就指令调谐语言模型和科学质量保证进行的宝贵讨论。感谢 LLaMA 团队为我们提供了访问