
Visual Instruction Tuning

Haotian Liu^{1*}, Chunyuan Li^{2*}, Qingyang Wu³, Yong Jae Lee¹

¹University of Wisconsin–Madison ²Microsoft Research ³Columbia University

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

摘要

利用机器生成的指令跟随数据对大语言模型（Large Language Models, LLMs）进行指令微调，已被证明能够显著提升模型在新任务上的零样本（zero-shot）能力，但这一思想在多模态领域中仍然缺乏深入研究。本文首次尝试使用仅具备语言能力的 GPT-4 来生成多模态的语言 - 图像指令跟随数据。基于上述生成数据进行指令微调，本文提出了 LLaVA（Large Language and Vision Assistant），一种端到端训练的大规模多模态模型，该模型通过连接视觉编码器与大语言模型，实现通用的视觉与语言理解能力。为促进视觉指令跟随方向的后续研究，本文构建了两个评测基准，涵盖多样且具有挑战性的应用导向任务。实验结果表明，LLaVA 展现出令人印象深刻的多模态对话能力，在面对未见过的图像和指令时，有时能够表现出类似多模态 GPT-4 的行为；在合成的多模态指令跟随数据集上，其性能相较于 GPT-4 达到了 85.1% 的相对得分。此外，在 Science QA 数据集上进行微调后，LLaVA 与 GPT-4 的协同作用取得了 92.53% 的准确率，达到了该任务上的最新最优性能。本文将由 GPT-4 生成的视觉指令微调数据、模型以及相关代码全部公开。

1 引言

人类通过多种信息通道（如视觉和语言）与世界进行交互，不同的通道在表示和传达特定概念方面各具优势，从而共同促进了对世界的更全面理解。人工智能领域的一个核心目标，是构建一种通用型智能助手，使其能够有效地理解并执行多模态的视觉 - 语言指令，与人类意图保持一致，从而在真实开放环境中完成多样化的实际任务 [4, 27, 26]。为实现这一目标，研究社区近年来逐渐兴起了对语言增强的基础视觉模型的研究热潮 [27, 16]。这些模型在开放世界视觉理解任务中展现出了强大的能力，包括图像分类 [40, 21, 57, 54, 39]、目标检测 [29, 62, 33]、图像分割 [25, 63, 58]、图像描述生成 [50, 28]，以及视觉生成与编辑任务 [42, 43, 56, 15, 44, 30]。关于该方向更为最新和系统的研究工作，读者可参考 Computer Vision in the Wild 阅读列表 [12]。

在上述研究范式中，通常由单一的大规模视觉模型分别解决各类任务，任务指令往往被隐式地融入模型设计之中。此外，语言主要用于对图像内容进行描述。尽管这种方式使语言在将视觉信号映射到语言语义（这一人类交流的通用通道）方面发挥了重要作用，但也导致模型接口较为固定，交互性有限，且难以灵活适应用户多样化、动态变化的指令需求。

相比之下，大语言模型（Large Language Models, LLMs）展示了语言更为广泛的潜力：语言可以作为通用型智能助手的统一交互接口，通过显式地以语言形式表达不同任务指令，引导端到端训练的神经网络助手在不同任务之间进行切换并加以解决。例如，ChatGPT [35] 和 GPT-4 [36] 的成功，充分展示了经过对齐的大语言模型在理解和执行人类指令方面的强大能力，同时也极大推动了开源大语言模型的发展。其中，LLaMA [49] 是一种性能可与 GPT-3 相媲美的开源大语言模型；Alpaca [48]、Vicuna [9] 以及 GPT-4-LLM [38] 等模型，则利用多种机器生成的高质量指令跟随样本来提升模型的指令对齐能力，其性能已可

与部分专有大语言模型相竞争。需要强调的是，上述研究均局限于纯文本场景。

在本文中，我们提出了视觉指令微调（visual instruction-tuning），首次将指令微调方法扩展至语言-图像多模态空间，为构建通用型视觉智能助手奠定基础。具体而言，本文的主要贡献包括：

- **模态指令跟随数据。**视觉-语言指令跟随数据的匮乏是当前面临的关键挑战之一。本文从数据重构的视角出发，提出了一套数据处理流程，利用 ChatGPT / GPT-4 将图像-文本对转换为适用于指令跟随任务的数据格式。
- **大规模多模态模型。**我们构建了一种大规模多模态模型（Large Multimodal Model, LMM），通过将 CLIP [40] 的开放集视觉编码器与语言解码器 Vicuna [9] 相连接，并在生成的多模态指令数据上进行端到端微调。实验结果验证了使用生成数据进行多模态指令微调的有效性，并总结了构建通用指令跟随视觉智能体的实践经验。当与 GPT-4 进行集成时，该方法在 Science QA [34] 多模态推理数据集上取得了最新最优（SoTA）性能。
- **多模态指令跟随评测基准。**我们提出了 LLaVA-Bench，包含两个具有挑战性的评测基准，涵盖多样化的图像、指令配对以及细粒度标注。
- **开源资源。**我们向社区公开了由模型生成的多模态指令数据、完整代码库、模型权重以及可交互的视觉对话演示系统。

2 相关工作

多模态指令跟随智能体（Multimodal Instruction-following Agents）在计算机视觉领域，现有构建指令跟随智能体的研究工作大致可以分为两类：（i）端到端训练的模型，通常针对每一个具体研究任务分别进行探索。例如，在视觉-语言导航任务中 [3, 19] 以及 Habitat 平台 [47] 中，具身人工智能（Embodied AI）智能体需要理解自然语言指令，并在视觉环境中执行一系列动作以完成既定目标。在图像编辑领域，给定一张输入图像及一条描述操作意图的文本指令，InstructPix2Pix [6] 能够通过遵循人类指令对图像进行编辑。（ii）通过 LangChain [1] 或大语言模型（LLMs）[35] 协调多个模型的系统，例如 Visual ChatGPT [53]、X-GPT [63]、MM-REACT [55]、VisProg [18] 以及 ViperGPT [46] 等。尽管上述方法均以构建指令跟随智能体为目标，本文关注的是开发一种端到端训练的视觉-语言多模态模型，以统一支持多种不同任务。

指令微调（Instruction Tuning）在自然语言处理（NLP）领域，为了使 GPT-3 [7]、T5 [41]、PaLM [10] 以及 OPT [60] 等大语言模型能够更好地理解自然语言指令并完成真实世界任务，研究者提出并系统探索了多种**指令微调（instruction tuning）方法** [37, 52, 51]，从而催生了相应的指令微调模型，如 InstructGPT [37] / ChatGPT [35]、FLAN-T5 [11]、FLAN-PaLM [11] 以及 OPT-IML [22]。研究表明，这种相对简单的训练策略能够显著提升大语言模型在零样本（zero-shot）和小样本（few-shot）场景下的泛化能力。因此，将该思想从 NLP 领域引入计算机视觉领域具有天然的合理性。

更广泛而言，基于基础模型的教师-学生蒸馏思想也已在其他研究方向中得到探索，例如图像分类任务 [14]。Flamingo [2] 可被视为多模态领域中的“GPT-3 时刻”，其在零样本任务迁移和上下文学习（in-context learning）方面展现了强大的性能。其他基于图像-文本对训练的大规模多模态模型还包括 BLIP-2 [28]、FROMAGE [24] 以及 KOSMOS-1 [20]。PaLM-E [13] 则是一种面向具身人工智能的多模态模型。

基于当前性能表现最优的开源大语言模型 LLaMA，OpenFlamingo [5] 和 LLaMA-Adapter [59] 等工作进一步赋予 LLaMA 处理图像输入的能力，为构建开源多模态大语言模型奠定了基础。尽管上述模型在任务迁移泛化方面展现出良好潜力，但它们**并未显式地利用视觉-语言指令数据进行微调**，因此在多模态任务上的性能通常不及其在纯语言任务中的表现。本文旨在弥补这一不足，并系统研究视觉指令微调方法的有效性。

最后需要指出的是，**视觉指令微调（visual instruction tuning）**不同于**视觉提示微调（visual prompt tuning）** [23]：前者的目标是提升模型的指令理解与执行能力，而后者则主要关注在模型适配过程中提高参数利用效率。

3 GPT辅助的视觉指令数据生成

近年来，研究社区见证了大量公开多模态数据的涌现，例如由图像 - 文本对构成的数据集，从 Conceptual Captions (CC) [8] 到 LAION [45]。然而，在**多模态指令跟随数据**方面，现有可用数据的规模仍然十分有限，其原因在于此类数据的构建过程往往耗时且缺乏明确规范，尤其是在依赖人工众包标注的情况下。

上下文类型1：描述

一群人站在一辆黑色车辆外，身边放着各种行李。
行李将一辆地下停车场的车团团围住。
人们正视图讲他们的所有行李塞进一辆SUV里。
这辆SUV停在公共车库中，人们正在为出行打包行李。
一些人带着行李站在一辆正在运输行李的厢式货车旁。



上下文类型2：箱子

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

回应类型1：对话

文体：图像中展示的是哪种类型的车辆？

回答：图像中展示的事一辆黑色SUV。...<省略>

回应类型2：详细描述

该图像展示了一个地下停车场场景，其中停着一辆黑色运动型多用途车（SUV）。画面中共有三人：一人站在车辆左侧附近，另一人位于车辆中部，第三人则在右侧。他们正协作将行李装入SUV，为即将出行做准备。...<省略>

回应类型3：复杂推理

问题：这些人面临哪些挑战？

回答：图像中，一群人站在地下停车场的一辆黑色SUV旁，周围散放着多个行李箱和背包。他们当前面临的主要挑战是如何将所有行李装入这辆SUV。现场存在大量行李物品，包括多个行李箱和背包，表明该群体携带的个人物品数量较多，可能超出车辆后备箱的常规装载容量，从而需要协调空间安排与装载策略。...<省略>

表 1：一个示例，用于说明指令遵循数据的构造方式。上半部分展示了用于提示 GPT 的上下文信息（如图像标题和边界框），下半部分展示了三种类型的回应。需要注意的是，在提示 GPT 时并未使用实际的视觉图像，此处仅作为参考展示。

现有的高质量多模态指令遵循数据数量有限，部分原因在于当依赖人工众包（human crowd-sourcing）时，此类数据的构建过程既耗时又缺乏明确标准。受近期 GPT 模型在文本标注任务中取得成功的启发 [17]，我们提出利用 ChatGPT/GPT-4 基于广泛存在的图像-文本对数据，自动生成多模态指令遵循数据。

对于一张图像 X_v 及其对应的标题 X_c ，很自然地可以构造一组问题 X_q ，用以指示助手描述图像内容。我们通过提示 GPT-4 来生成这类问题列表（详见附录）。因此，将一个图像 - 文本对扩展为指令遵循格式的一种简单方式是：**Human:** X_q X_v **Assistant:** X_c 。尽管这种扩展方式构建成本低廉，但其在指令和回应两方面均缺乏多样性与深度推理能力。

为缓解这一问题，我们利用纯文本的 GPT-4 或 ChatGPT 作为强大的“教师模型”（二者仅接受文本输入），来生成涉及视觉内容的指令遵循数据。具体而言，为了将图像编码为可用于提示纯文本 GPT 的形式，我们采用两类符号化表示：(i) **标题** (Captions) 通常从多个角度描述视觉场景；(ii) **边界框** (Bounding boxes) 通常定位场景中的物体，每个框包含物体类别及其空间位置信息。表14的上半部分展示了一个此类符号化表示的示例。

这种符号化表示使我们能够将图像编码为大语言模型（LLM）可识别的文本序列。我们基于 COCO 图像 [31] 生成了三种类型的指令遵循数据，表14下半部分分别为每种类型展示了一个示例。对于每种类型，我们首先人工设计少量示例——这些是整个数据收集过程中**唯一的人工标注内容**，并作为上下文学习（in-context learning）中的种子示例用于查询 GPT-4。

- **对话 (Conversation)：**我们设计一段助手与用户围绕该图像的多轮对话。回答的语气仿佛助手正在“看到”图像并作出回应。所提问题涵盖图像的多种视觉内容，包括

物体类别、物体数量、物体动作、物体位置以及物体之间的相对空间关系等。仅保留具有明确答案的问题（详细提示见附录）。

- 详细描述（Detailed description）：为生成对图像丰富而全面的描述，我们构造了一组旨在引导详细描述的问题，并提示 GPT-4 对其进行整理（详见附录中的提示模板与整理流程）。对于每张图像，我们从该问题列表中随机采样一个问题，要求 GPT-4 生成对应的详细描述。
- 复杂推理（Complex reasoning）：上述两类聚焦于视觉内容本身，而此类则在此基础上进一步构造需要深入推理的问题。其答案通常需遵循严谨逻辑，进行逐步推理。

我们共收集了 **158K 条** 唯一的语言-图像指令遵循样本，其中包含 **58K 条对话**、**23K 条详细描述** 和 **77K 条复杂推理** 样本。在早期实验中，我们对比了 ChatGPT 与 GPT-4 的使用效果，发现 GPT-4 在生成高质量指令遵循数据方面表现更优，尤其在空间推理等任务上更为可靠。

4 视觉指令微调

4.1 模型架构

本研究的主要目标是**有效结合预训练大语言模型（LLM）与视觉模型的能力**。整体网络架构如图 1 所示。我们选择 Vicuna [9] 作为语言模型 $f_\phi(\cdot)$ ，其参数为 ϕ 。在当前公开可用的模型权重中，Vicuna 在语言任务上的指令跟随能力表现最佳 [48, 9, 38]。

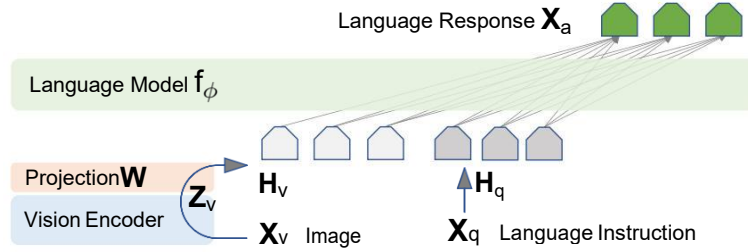


Figure 1: LLaVA network architecture.

对于输入图像 X_v ，我们采用预训练的 CLIP 视觉编码器 ViT-L/14 [40]，以获得视觉特征 $Z_v = g(X_v)$ 。在我们的实验中，考虑了最后一层 Transformer 之前和之后的网格特征。我们使用一个简单的线性层将图像特征映射到词嵌入空间。具体而言，我们引入一个可训练的投影矩阵 W ，将 Z_v 转换为语言嵌入 token H_v ，其维度与语言模型中的词嵌入空间维度一致：

$$H_v = W \cdot Z_v, \text{ with } Z_v = g(X_v) \quad (1)$$

因此，我们得到一个视觉 token 序列 H_v 。需要注意的是，我们采用的这种简单投影方案非常轻量，使我们能够快速迭代以数据为中心的实验。当然，也可以考虑更复杂的图像与语言表征融合方案，例如 Flamingo [2] 中的门控交叉注意力机制（gated cross-attention）和 BLIP-2 [28] 中的 Q-former。我们将探索可能更有效、更复杂的 LLaVA 架构设计留作未来的工作。

4.2 训练策略

对于每张图像 X_v ，我们生成多轮对话数据 $(X_{q_1}, X_{a_1}, \dots, X_{q_T}, X_{a_T})$ ，其中 T 表示总对话轮数。我们将这些数据组织成一个统一序列，并将所有答案视为助手（Assistant）的回复，第 t 轮的指令 X_{instruct}^t 定义如下：

$$X_{\text{instruct}}^t = \begin{cases} \text{随机选择 } [X_{q_1}, X_v] \text{ 或 } [X_v, X_{q_1}], & t = 1 \\ X_{q_t}, & t > 1 \end{cases} \quad (2)$$

这形成了表 2 所示的统一多模态指令遵循序列格式。我们使用大语言模型（LLM）原有的自回归训练目标，在预测标记上对其进行指令微调。具体而言，对于长度为 L 的序列，目标答案 X_a 的概率计算如下：

$$p(X_a | X_v, X_{\text{instruct}}) = \prod_{i=1}^L p_{\theta}(x_i | X_v, X_{\text{instruct}, <i}, X_{a, <i}) \quad (3)$$

$X_{\text{system-message}} <\text{STOP}>$
 Human: $X_{\text{instruct}}^1 <\text{STOP}>$ Assistant: $X_a^1 <\text{STOP}>$
 Human: $X_{\text{instruct}}^2 <\text{STOP}>$ Assistant: $X_a^2 <\text{STOP}> \dots$

表 2: 用于训练模型的输入序列。此处仅示意展示了两轮对话；在实际训练过程中，对话轮数会根据具体的指令跟随数据而有所不同。在当前实现中，我们按照 Vicuna-v0 [9] 的设置方式构造系统提示语 $X_{\text{system-message}}$ ，并将特殊终止标记设为 $<\text{STOP}> = \text{###}$ 。模型的训练目标是预测助手（Assistant）的回复内容以及生成终止标记的位置，因此在自回归训练过程中，仅对助手回复部分（图中绿色序列/token）计算损失。

其中， θ 表示可训练参数； $X_{\text{instruct}, <i}$ 和 $X_{a, <i}$ 分别表示在当前预测 token x_i 之前所有轮次中的指令 token 和答案 token。预测 token 的具体示例可参见表 2。在公式 (3) 的条件概率表示中，我们显式加入图像输入 X_v ，以强调所有答案均以图像为基础进行生成。为便于阅读，在表达式中省略了 $X_{\text{system-message}}$ 以及此前出现的所有 $<\text{STOP}>$ 标记。在 LLaVA 模型的训练过程中，我们采用两阶段的指令微调策略。

第一阶段：用于特征对齐的预训练。 为了在概念覆盖范围与训练效率之间取得平衡，我们对 CC3M 数据集进行了筛选，保留了 595K 个图像-文本对。具体的筛选过程详见附录。这些数据对通过第 3 节所述的朴素扩展方法（naive expansion method）转换为指令遵循格式，每个样本可视为一轮单轮对话。为构建公式 (2) 中的输入 X_{instruct} ，对于给定图像 X_v ，我们随机采样一个问题 X_q ，该问题是一条语言指令，要求助手简要描述该图像；其对应的预测答案 X_a （即真实标签）即为原始图像标题（caption）。在训练过程中，我们冻结视觉编码器和大语言模型（LLM）的权重，仅使用可训练参数 $\theta = W$ （即投影矩阵）来最大化公式 (3) 的似然。通过这种方式，图像特征 H_v 可以与预训练 LLM 的词嵌入空间对齐。该阶段可理解为为冻结的 LLM 训练一个兼容的视觉分词器（visual tokenizer）。

第二阶段：端到端微调。 我们始终冻结视觉编码器的权重，同时继续更新 LLaVA 中投影层和大语言模型（LLM）的预训练权重；即公式 (3) 中的可训练参数为 $\theta = \{W, \phi\}$ 。我们考虑以下两种具体应用场景：

- 多模态聊天机器人（Multimodal Chatbot）：我们基于第 3 节中所述的 158K 条语言-图像指令遵循数据进行微调，构建一个聊天机器人。在三种回应类型中，“对话”为多轮交互，其余两种为单轮。在训练过程中，这三类样本被均匀采样。
- 科学问答（Science QA）：我们在 ScienceQA 基准 [34] 上验证所提方法。该基准是首个大规模多模态科学问题数据集，其答案配有详细的讲解与推理说明。每个问题的上下文以自然语言或图像形式提供，助手需用自然语言给出推理过程，并从多个选项中选择正确答案。针对公式 (2) 的训练设置，我们将数据组织为单轮对话：将问题及其上下文作为 X_{instruct} ，将推理过程与答案作为 X_a 。

5 实验

我们从两个主要实验设置出发，评估 LLaVA 在指令跟随能力与视觉推理能力方面的表现：多模态聊天机器人任务以及 ScienceQA 数据集。所有模型均使用 8 张 NVIDIA A100 GPU 进行训练，并采用与 Vicuna 相同的超参数设置 [9]。在训练流程上，我们首先在筛选后的 CC-595K 子集上进行预训练，共 1 个 epoch，学习率设为 $2e-3$ ，batch size 为 128。随后，在本文提出的 LLaVA-Instruct-158K 数据集上进行 3 个 epoch 的微调，学习率设为 $2e-5$ ，batch size 为 32。更多训练细节请参见附录。

5.1 多模态机器人

我们开发了一个聊天机器人演示系统，用以展示 LLaVA 在图像理解与对话生成方面的能力，并研究其对视觉输入信息的整合能力以及指令跟随能力的表现。首先，我们采用原始 GPT-4 论文 [36] 中的示例（见表 3，更多示例见附录），这些示例通常需要较为深入的图像理解能力。为进行对比，我们引用其论文中多模态 GPT-4 的提示语（prompt）及对应回复结果，同时调用 BLIP-2 和 OpenFlamingo 的公开模型权重进行推理，并获取其生成结果进行比较分析。

Visual input example, Extreme Ironing:



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

User	这张图片有什么不寻常之处？
LLaVA	这张图片中不寻常之处在于，一名男子正在一辆小型货车（或面包车）的车顶上熨烫衣服。通常情况下，人们会在更为固定且安全的环境中熨衣服，例如在家中使用标准熨衣板进行操作。因此，在车辆上搭建一个临时熨衣装置本身就显得非常不寻常。该场景之所以怪异，是因为男子选择在一辆车辆上进行熨衣，这种做法既不安全，也不符合常规。此外，在如此不稳定的环境中保持平衡并进行熨烫操作，本身也令人感到疑惑。
	[Start a new conversation, and clear the history]
User	场景中正在发生什么？
LLaVA	图像展示了一名男子站在一辆黄色 SUV 的车顶上，地点似乎是在一条繁忙的城市街道上。他手持一个便携式梯子，正在车顶上熨烫衣服。周围环境包括其他车辆、交通信号灯以及消防栓。此外，现场还有另外两人，一人靠近车辆附近，另一人则在稍远处。由于在行驶车辆上熨衣服这一行为十分反常，该男子的举动显得格外引人注目。繁忙的城市街景进一步增强了这一场景的独特性与荒诞感。
User	这张图片有什么不寻常之处？
GPT-4 [36]	这张图片的不寻常之处在于，一名男子正在一辆正在行驶的出租车车顶上，使用固定在车顶上熨衣板熨烫衣物。
User	这张图片有什么不寻常之处？
BLIP-2	一名男子坐在一辆黄色出租车的后部。
User	这张图片有什么不寻常之处？
OpenFlamingo	这名男子正在汽车引擎盖上晾晒衣服。

表 3：来自 GPT-4 论文 [36] 的示例提示，用于比较视觉推理与对话能力。与 BLIP-2 [28] 和 OpenFlamingo [5] 相比，LLaVA 能够准确遵循用户指令，而不仅仅是对场景进行简单描述。LLaVA 给出的回答比 GPT-4 更加全面。即便在仅被要求描述图像时，LLaVA 也能够识别图像中的异常之处。

令人惊讶的是，尽管 LLaVA 仅使用了规模较小的多模态指令跟随数据集（约 8 万张独立图像）进行训练，但在这些示例中，其推理结果与多模态 GPT-4 表现出高度相似性。需要注意的是，这些图像对于 LLaVA 来说属于域外数据（out-of-domain），然而 LLaVA 依然能够理解图像内容，并根据问题指令给出合理的回答。相比之下，BLIP-2 和 OpenFlamingo 更倾向于对图像进行描述，而未能有效遵循用户指令进行针对性回答。

定量评估为了系统性地评估 LLaVA 的性能，我们提出了一种定量指标，用于衡量模型在多模态数据上的指令跟随能力。受文献 [9] 的启发，我们采用 GPT-4 作为自动评估工具来衡量生成结果的质量。具体而言，我们构建了由“图像—真实文本描述—问题”组成

的三元组数据。候选模型（如 LLaVA）基于图像和问题生成对应答案。为了提供一个近似的理论性能上界，我们利用仅基于问题和真实文本描述的文本版 GPT-4 生成参考预测结果。随后，我们将问题、视觉信息（以文本描述形式表示）以及两个模型生成的回答一同输入至评判模型（即文本版 GPT-4）。评判模型从以下几个维度对回答进行综合评估：有用性（Helpfulness）相关性（Relevance）准确性（Accuracy）细节程度（Level of Detail）最终给出 1 - 10 分的综合评分，其中分数越高表示整体性能越优。同时，评判模型还需给出详细的文字解释，以便我们更深入理解模型性能差异的来源。

	Conversation	Detail description	Complex reasoning	All
Full data	83.1	75.3	96.5	85.1
Detail + Complex	81.5 (-1.6)	73.3 (-2.0)	90.8 (-5.7)	81.9 (-3.2)
Conv + 5% Detail + 10% Complex	81.0 (-2.1)	68.4 (-7.1)	91.5 (-5.0)	80.5 (-4.4)
Conversation	76.5 (-6.6)	59.8 (-16.2)	84.9 (-12.4)	73.8 (-11.3)
No Instruction Tuning	22.0 (-61.1)	24.0 (-51.3)	18.5 (-78.0)	21.5 (-63.6)

表 4: 在 LLaVA-Bench (COCO) 上对不同训练数据的消融实验。我们报告相对于一个纯文本 GPT-4 模型的相对得分，该 GPT-4 模型以真实图像标题（ground truth captions）和边界框（bounding boxes）作为视觉输入。我们使用本模型的输出答案和纯文本 GPT-4 生成的答案共同提示 GPT-4，让其对两个回答进行比较，并给出评分及解释。

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

表 5: 在 LLaVA-Bench (In-the-Wild) 上的指令跟随能力对比结果（采用相对评分）。结果以“均值 ± 标准差”的形式报告。对于前三行结果，我们报告了三次推理运行的平均结果。LLaVA 的性能显著优于其他模型。† 对于同一组 LLaVA 解码序列，我们向 GPT-4 进行三次评估查询，GPT-4 给出了高度一致的评估结果。我们报告了相对于文本版 GPT-4 模型的相对得分，其中该 GPT-4 使用真实文本描述作为视觉输入。我们构建了两个基准测试，以评估模型性能。

LLaVA-Bench (COCO)。我们从 COCO-Val-2014 中随机选取 30 张图像，并针对每张图像使用第 3 节提出的数据生成流程构建三种类型的问题（对话、详细描述、复杂推理），共计 90 个问题。该基准主要用于研究模型在视觉输入一致条件下的对齐行为与能力表现。我们通过改变训练数据集的组成，研究不同类型指令跟随数据的有效性，结果如表 4 所示。实验结果表明：1、经过指令微调后，模型在遵循用户指令方面的能力显著提升，得分提升超过 50 分。2、在训练数据中加入少量详细描述和复杂推理问题，使模型整体能力进一步提升 7 分。3、推理能力的增强同时也提升了模型在对话类问题上的表现，说明推理能力的提升能够促进对话能力的提高。4、当同时使用三种类型的数据进行训练时，模型性能达到最佳，相对得分为 85.1%。

LLaVA-Bench (In-the-Wild)。为了评估模型在更具挑战性的任务中的能力以及在新领域上的泛化能力，我们收集了 24 张多样化图像，共 60 个问题。这些图像涵盖室内外场景、网络表情包、绘画、素描等多种类型。每张图像均配有人工精细整理的高质量描述以及精心设计的问题集合。我们在表 5 中比较了 LLaVA、BLIP-2 和 OpenFlamingo 的性能。得益于视觉指令微调，LLaVA 的表现显著优于 BLIP-2 (+29%) 和 OpenFlamingo (+48%)。与能够访问真实文本标签的文本版 GPT-4 相比，LLaVA 在复杂推理问题上取得了 81.7% 的相对性能，整体得分达到 67.3%，表现令人印象深刻。

局限性。LLaVA-Bench (In-the-Wild) 旨在具有挑战性，并用于揭示模型的潜在弱点。我们在表 6 中展示了两个示例及其对应的描述与问题。拉面示例（左图）：若要正确回答餐厅名称，模型需要具备广泛的知识覆盖以及多语言理解能力；而若要准确描述配菜内容，则可能需要从互联网检索相关多模态信息。冰箱示例（右图）：若要正确识别酸奶品牌，模型需要处理高分辨率图像并具备较强的知识储备。我们还观察到一个有趣的失败案例：当被问及冰箱中是否存在草莓味酸奶时，LLaVA 回答“是”，然而图中实际上仅包含普通酸奶和草莓。这表明在某些情况下，LLaVA 将图像视为“局部图块的集合”（a bag of

patches)，未能正确理解图像中的复杂语义关系。我们希望 LLaVA 能够作为这些基准测试上的一个坚实基线模型，并以此为基础，推动未来更强大多模态大模型（LMM）的研究与发展。

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

	
ICHIRAN Ramen [source]	Filled fridge [source]
注释	<p>这是一张在拉面连锁店 ICHI-RAN 拍摄的餐食特写照片，构图精致，细节清晰。画面中央摆放着一碗热气腾腾的叉烧拉面，旁边配有一把金属勺子，整体摆盘极具日式料理的简约美感。拉面上淋有辣椒酱、撒有切碎的葱花，并配有两片叉烧。碗的右侧放着一双筷子，仍包裹在纸套中，尚未拆开。拉面左侧配有一片海苔。照片上方从左至右依次摆放着以下配菜：一碗橙色调味料（可能是蒜蓉辣酱）、一碟烟熏风味的炖叉烧配葱花，以及一杯抹茶绿茶。</p> <p>这张照片展示了一台打开的家用冰箱内部，冷藏室隔间被各类食材填得满满当当。整个空间被自然划分为左、中、右三个区域，每个区域从前到后都精心（或随意）地堆放着不同的食品。在隔间的左侧靠前位置，有一个装草莓的塑料盒，上面放着一小袋迷你胡萝卜。靠后位置则叠放着几盒酱料容器。隔间中间靠前处有一个绿色塑料盒，其上放着一个无法辨识内容的塑料袋；靠后位置放着一盒牛奶。隔间右侧靠前处有一盒蓝莓，上面叠放着三杯酸奶；最下面是一大瓶 Fage 无脂酸奶，中间是一小杯 Fage 蓝莓味酸奶，另一小杯的品牌和口味则无法确定。右侧靠后位置还有一个容器，内容物未知。</p>
问题 1	餐馆的名字是什么？
问题 2	详细地描述这张照片。
	蓝莓风味的酸奶是什么品牌？
	这里面有草莓风味的酸奶吗？

表 6：来自 LLaVA-Bench (In-the-Wild) 的具有挑战性的示例。我们为每张图像提供了极为详细的标注，以确保评估的准确性。其中一些问题要求模型能够从高分辨率图像中提取细节，并具备广泛的知识覆盖能力。

5.2 ScienceQA

ScienceQA [34] 包含 21,000 道多模态选择题，涵盖丰富的领域多样性，涉及 3 个学科 (Subjects)、26 个主题 (Topics)、127 个类别 (Categories) 以及 379 项技能 (Skills)。该基准数据集划分为训练集、验证集和测试集，分别包含 12,726、4,241 和 4,241 个样本。我们选取了具有代表性的对比方法，包括：GPT-3.5 (text-davinci-002)，分别测试是否使用链式思维 (Chain-of-Thought, CoT)；LLaMA-Adapter [59]；多模态链式思维方法 MM-CoT [61]（当前该数据集上的最新最优方法，SoTA）。更多基线结果请参见文献 [34]。

结果如表7所示。对于 LLaVA，我们采用视觉编码器最后一层之前的视觉特征，在推理时要求模型先生成推理过程，再给出最终答案，并在训练中迭代了12个epoch。该设置取得了90.92%的准确率，与当前最优 (SoTA) 的91.68%非常接近。为了探索大语言模型 (LLM) 的能力边界，我们还尝试使用两样本上下文学习 (2-shot in-context learning) 的方式直接提示 GPT-4，获得了82.69%的准确率，相比 GPT-3.5 的 75.17% 提升了 7.52 个百分点 (绝对值)。然而，我们注意到，在大量问题上，GPT-4 的失败并非因为推理能力不足，而是因为它明确表示“缺乏足够的上下文（如图像或图表）”，因而拒绝作答。

为此，我们设计了两种融合 LLaVA 与 GPT-4 输出结果的策略：(i) GPT-4 补充方案 (GPT-4 complement)：当 GPT-4 无法提供答案时（例如因缺少图像而拒绝回答），我们转而采用 LLaVA 的预测结果。该方案最终达到 90.97% 的准确率，几乎与单独使用

LLaVA 相当，说明 GPT-4 在此设定下并未带来显著增益。（ii）GPT-4 作为裁判（GPT-4 as the judge）：当 GPT-4 与 LLaVA 给出的答案不一致时，我们再次提示 GPT-4，要求它基于原始问题以及两个模型的输出，重新判断并给出最终答案。这一思路类似于思维链（Chain-of-Thought, CoT），但引入了来自另一模型的外部知识。令人惊讶的是，该方案在所有问题类别上均带来稳定提升，最终实现了 92.53% 的新 SoTA 准确率。有趣的是，尽管纯文本版的 GPT-4 无法处理图像，但它却能提升那些以图像为上下文的问题的整体性能。原因在于：部分问题实际上并不真正依赖图像信息即可正确回答（例如常识性问题或仅需题干文字即可推理的问题）。在此类情况下，GPT-4 裁判能够识别出 LLaVA 因过度依赖视觉信号而产生的错误，并加以纠正（详见附录中的示例）。

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
<i>Representative & SoTA methods with numbers reported in the literature</i>									
Human [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]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
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
<i>Results with our own experiment runs</i>									
GPT-4 [†]	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 [†] (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4 [†] (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	92.53

表 7: Science QA 数据集上的准确率（%）。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。† 纯文本版 GPT-4，由我们评估。我们提出的与纯文本 GPT-4 的新型模型集成方法在所有类别上均持续提升模型性能，取得了新的最先进（SoTA）成绩。

Visual features	Before	Last
Best variant	90.92	89.96 (-0.96)
Predict answer first	-	89.77 (-1.15)
Training from scratch	85.81 (-5.11)	-
7B model size	89.84 (-1.08)	-

表 8: 设计选择的消融实验（%）。与最佳变体的差异以红色文字标出。

这是首次将 GPT-4 用于模型集成。我们希望这一发现能够激励未来的研究探索更有效的方法，以利用大语言模型（LLMs）进行模型集成。消融实验。我们在 ScienceQA 上对若干设计选择进行了消融研究，结果见表8。（i）视觉特征。我们尝试使用 CLIP 视觉编码器最后一层的特征，得到 89.96% 的准确率，比倒数第二层之前的特征低了 0.96%。我们推测这是因为 CLIP 最后一层的特征可能更关注全局和抽象的图像属性，而其前一层的特征则更侧重于局部特性，这对于理解图像中的具体细节更为有用。（ii）思维链（Chain-of-thought）。为确定模型预测中答案与推理过程的顺序，我们测试了两种变体：采用“先答案”策略时，在 12 个 epoch 内达到最佳准确率 89.77%；而“先推理”策略在 6 个 epoch 内即可快速达到 89.77% 的准确率，但继续训练后未见进一步提升。即使将训练延长至 24 个 epoch，性能也未改善。我们得出结论：类似思维链的“先推理”策略可显著加快收敛速度，但对最终性能的贡献相对有限。（iii）预训练。我们跳过预训练阶段，直接从零开始在 Science QA 上训练，性能降至 85.81% 的准确率。这一 5.11% 的绝对下降表明我们的预训练阶段至关重要——它在对齐多模态特征的同时，保留了大语言模型中丰富的预训练知识。（iv）模型规模。我们在保持所有配置与最佳 13B 模型一致的前提下，训练了一个 7B 模型，其准确率为 89.84%，比 90.92% 低了 1.08%，这说明了模型规模的重要性。

6 结论

本文验证了视觉指令微调 (visual instruction tuning) 的有效性。我们提出了一条自动化流程, 用于构建语言-图像指令跟随数据集, 并基于该数据训练了 LLaVA——一种能够理解人类意图并完成视觉任务的多模态模型。在 ScienceQA 上进行微调后, LLaVA 取得了新的最先进 (SoTA) 准确率; 在多模态对话数据上微调后, 模型展现出优异的视觉对话能力。此外, 我们提出了首个用于研究多模态指令跟随能力的基准测试。本文是视觉指令微调方向的初步探索, 主要聚焦于现实生活场景任务。关于 LLaVA 在学术基准测试上的更多定量结果, 请参阅基于视觉指令微调改进的基线模型工作 [32]。我们希望本研究能够为构建更强大的多模态模型提供启发。

References

- [1] Langchain. <https://github.com/hwchase17/langchain>, 2022. 2
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2, 4
- [3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 2
- [4] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021. 1
- [5] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. 2, 6, 7
- [6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instruct pix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 2
- [9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 1, 2, 4, 5, 6
- [10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2
- [11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2
- [12] CVinW. Computer vision in the wild. https://github.com/Computer-Vision-in-the-Wild/CVinW_Readings, 2022. 1
- [13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2
- [14] Fartash Faghri, Hadi Pouransari, Sachin Mehta, Mehrdad Farajtabar, Ali Farhadi, Mohammad Rastegari, and Oncel Tuzel. Reinforce data, multiply impact: Improved model accuracy and

- robustness with dataset reinforcement. *arXiv preprint arXiv:2303.08983*, 2023. 2
- [15] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *ArXiv*, abs/2203.13131, 2022. 1
- [16] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 2022. 1
- [17] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023. 3
- [18] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *arXiv preprint arXiv:2211.11559*, 2022. 2
- [19] Weituo Hao, Chunyuan Li, Xiujuan Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, 2020. 2
- [20] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 2
- [21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. July 2021. If you use this software, please cite it as below. 1
- [22] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-1ml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 2
- [23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2
- [24] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023. 2
- [25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ICLR*, 2022. 1
- [26] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023. 1
- [27] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS Track on Datasets and Benchmarks*, 2022. 1
- [28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 4, 6, 7
- [29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 1
- [30] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023. 1
- [31] 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

- [32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 9, 14
- [33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1
- [34] Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 2022. 2, 5, 8, 9
- [35] OpenAI. ChatGPT. <https://openai.com/blog/chatgpt/>, 2023. 1, 2
- [36] OpenAI. Gpt-4 technical report, 2023. 1, 5, 6, 15
- [37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 2
- [38] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*, 2023. 1, 4
- [39] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv: 2111.10050*, 2021. 1
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 4
- [41] 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. 2
- [42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1
- [43] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, pages 10674–10685, 2022. 1
- [44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 1
- [45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2
- [46] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 2
- [47] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

- [48] 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. 1, 4
- [49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [50] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1
- [51] 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 preprint arXiv:2212.10560*, 2022. 2
- [52] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022. 2
- [53] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2
- [54] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Lu Yuan, Ce Liu, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *CVPR*, 2022. 1
- [55] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2
- [56] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv*, abs/2206.10789, 2022. 1
- [57] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1
- [58] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 1
- [59] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2, 8, 9
- [60] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2
- [61] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multi-modal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 8, 9
- [62] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 1
- [63] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. *arXiv preprint arXiv:2212.11270*, 2022. 1, 2

A 更广泛影响

作为一种通用视觉助手，LLaVA 的部署和发布可能带来潜在收益与风险。其中部分问题源于其视觉特性，另一些则与现有的指令跟随型大语言模型（如 Alpaca、Vicuna 等）类似。由于 LLaVA 构建于 LLaMA、Vicuna 和 CLIP 之上，因此也继承了部分来自大语言模型与视觉编码器的相关问题。以下我们概述模型发布时面临的风险及相应的缓解策略。

恶意输入 (Malicious Input)。为尽量减少潜在滥用和有害后果，我们为 LLaVA 采取了两项预防措施：1、使用 OpenAI 的过滤 API 对用户输入文本进行筛查，防止有害或不当文本指令被模型处理；2、对用户上传的图像使用 NSFW 过滤器，以检测并阻止不适宜工作场所 (NSFW) 或其他潜在有害的图像内容。

幻觉问题 (Hallucination)。与大语言模型类似，LLaVA 可能生成与事实或输入数据不一致的内容。这在关键应用场景（例如医疗领域）中尤为令人担忧，因为模型可能会做出不可靠推断。

偏见 (Biases)。偏见可能从基础模型迁移至 LLaVA，包括来自视觉编码器 (CLIP) 和语言解码器 (LLaMA/Vicuna) 的偏见。这可能导致输出结果存在偏见或对多样化内容产生不公平表征。

能源消耗 (Energy Consumption)。由于预训练数据规模相对较小（详见附录 C），能源消耗目前并非 LLaVA 的主要问题。但在扩大预训练数据规模或增加模型规模（例如扩展到更大的 LLaMA 版本，如 65B 模型）时，能源消耗可能成为值得关注的问题。

评估复杂性 (Evaluation Complexities)。评估 LLaVA 的性能具有挑战性，因为其同时涉及语言与视觉任务。我们构建的评测基准涵盖多个方面，包括准确率、概念覆盖度、推理能力与创造力。然而，还需进一步考虑其他指标，例如视觉内容幻觉程度以及对视觉内容的细粒度理解能力。虽然在我们的研究中，基于文本的 GPT-4 多模态评估表现出一致性和准确性，但其在不同场景下的鲁棒性以及尚未探索方面的评估能力仍有待未来研究。

尽管存在上述风险，我们认为将 LLaVA 向研究社区开放的收益大于潜在风险。这有助于持续研究和改进模型，并促进社区共同开发更完善的缓解策略。此外，LLaVA 的发布也将推动新应用与研究方向的发展，促进视觉-语言基础模型的进步与负责任部署。

B 更多结果

我们提供了更多 LLaVA 的定性结果，以分析其涌现行为与已观察到的弱点。有关 LLaVA 在学术基准测试上的更多定量结果，请参阅基于视觉指令微调改进的基线模型工作 [32]。在表 9 中，LLaVA 在论文中的另一个示例里展现出与 GPT-4 类似的行为。类似于 OpenAI 的 GPT-4 现场演示，LLaVA 能够根据图 2 中用户提供的简化草图生成交互式笑话网站的 HTML/JS/CSS 代码（尽管存在一个小错误）。如图 3 所示，LLaVA 能够以对话风格遵循用户指令，并提供详细回答或创造性写作。此外，LLaVA 还能将视觉内容与预训练语言模型中的文本知识联系起来（见图 4 和图 5）。

一个有趣的涌现行为是，LLaVA 能够理解训练数据中未出现过的视觉内容。例如，在图 6 中，它可以识别 Elon Musk 的头像，以及一张他打扮成 Doge（柴犬梗图风格）的幽默表情图，尽管在视觉特征对齐阶段和视觉指令微调阶段的训练数据中都未包含 Elon Musk。LLaVA 还在表 9 和图 2 中展现出令人印象深刻的 OCR（光学字符识别）能力，而这在我们的训练数据中覆盖极少。我们希望这些补充结果与观察能够展示 LLaVA 在多种应用场景中的潜力。未来研究需要更深入地探究这些涌现行为，并理解其背后的机制，从而构建更强大的多模态模型 (LMM)，包括提升鲁棒性、减少偏见，以及改进视觉-语言表示的对齐程度与覆盖范围。

Visual input example, Chicken Nugget Map:

<div><div>Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.</div><div></div></div>	
User	你能详细解释这个梗图吗？
LLaVA	这张图中的梗图是一种富有创意且幽默的表达方式，将鸡块作为“宇宙的中心”。梗图开头写道：“有时候我会看着从太空拍摄的地球照片，惊叹于它是多么美丽……”主图是一盘炸的金黄的熟鸡块，这些鸡块被精心地摆放成类似世界地图的形状，看起来像大陆和岛屿。梗图巧妙地暗示，这些鸡块代表地球，而照片中展示的不同位置其实是“鸡块版本”的世界各地。梗图的笑点出现在底部的小字中：“我是说，这不是真正的地球，但它真的很美。”这张轻松有趣的梗图将食物与幽默结合，通过鸡块的形象创造出一个滑稽而富有想象力的“世界”。
User	你能解释这个梗图吗？
GPT-4 [36]	这个梗图把两种毫不相关的事物结合在一起：从太空拍摄的地球照片和鸡块。文字暗示下方是一张美丽的地球照片，但实际上图像是鸡块被摆放成类似世界地图的形状。幽默来自文字与图像之间出人意料的对比。文字营造出宏伟壮观的期待，但图像却是平凡又滑稽的鸡块。
User	你能详细解释这个梗图吗？
BLIP-2	有时候我只是看着从太空拍摄的地球照片，惊叹它是多么美丽。（仅重复图片文字，未完成解释。）
User	你能详细解释这个梗图吗？
OpenFlamingo	这是一张在国际空间站上的鸡块照片。（理解错误。）

表 9：该示例对比了 LLaVA、GPT-4、BLIP-2 和 OpenFlamingo 在理解幽默方面的视觉推理能力。BLIP-2 和 OpenFlamingo 未能遵循用户指令。LLaVA 和 GPT-4 都成功解释了梗图及其幽默之处，其中 GPT-4 的回答更为简洁。



```
<!DOCTYPE html>
<html>
<head>
<title>My Joke Website</title>
<style>
body { background-color: white; font-family: Arial, sans-serif; }
.container { display: flex; flex-direction: column; align-items: center; justify-content: center; height: 100vh; }
h1 { font-size: 36px; margin-bottom: 20px; }
p { font-size: 20px; }
button { background-color: blue; color: white; border: none; padding: 10px 20px; font-size: 18px; cursor: pointer; }
</style>
</head>
<body>
<div class="container">
<h1>My Joke Website</h1>
<p id="punchline">Welcome to my joke website. Push the button to reveal the punchline!</p>
<button onclick="revealPunchline()">Push me!</button>
</div>
<script>
function revealPunchline() {
document.getElementById("punchline").innerHTML = "The punchline is: 'Why was the math book sad? Because it had too many
problems.'";
}
</script>
</body>
</html>
```

Place the generated HTML code into an HTML renderer to render the web page.

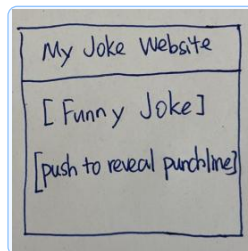


Rendered website



Rendered website
(clicked the button)

图 2: LLaVA 根据用户的手绘草图生成用于交互式网站的 HTML/JS 代码。在对生成结果中的一处小错误（标红）进行修正后，该交互界面即可正常运行。LLaVA 的输出仍有改进空间，例如将笑话和笑点分别放在两行，并在用户点击按钮后才显示笑点，以更好地体现用户的意图。



User

把这份模拟图用 HTML/JS 做成一个可交互的网站，并把里面的笑话换成一个真实的笑话。



User



我能用这些食材做什么菜？



LLaVA

冰箱里有多种食材可供选择，你可以制作多种餐点。例如，你可以用草莓、蓝莓和胡萝卜制作一份水果沙拉。此外，还可以用这些水果和蔬菜制作美味的奶昔，作为健康又营养的零食或早餐选择。

牛奶和鸡蛋可以用来制作各种菜肴，例如牛奶奶昔、烘焙砂锅菜或炒鸡蛋。果汁也可以用来制作新鲜的水果果汁混合饮品，或者做成清爽的自制冰棒。

选择非常丰富，具体可以做什么取决于你的口味偏好和饮食需求。

我想做水果沙拉，请展示食谱。



User



LLaVA

当然！以下是一份使用你冰箱中食材制作的简单水果沙拉食谱：

材料：

- 2杯新鲜草莓（切片）
- 1杯新鲜蓝莓
- 1根大胡萝卜
- 半个柠檬的汁水
- 一小把切碎的新鲜欧芹或薄荷
- 几块冰块
- 可选：蜂蜜或枫糖浆（根据口味调整）

做法：

1. 在一个大碗中，将切片草莓、蓝莓和胡萝卜丝混合在一起，轻轻翻拌均匀。
2. 将半个柠檬的汁均匀淋在水果和蔬菜上，以防止氧化，并增添清新的柑橘风味。
3. 撒上切碎的新鲜欧芹或薄荷，再轻轻翻拌一次。
4. 如需增加甜味，可以淋上少量蜂蜜或枫糖浆，根据个人口味调整用量。
5. 盖上保鲜膜，放入冰箱冷藏至少 30 分钟或数小时，使味道充分融合，口感更清爽。
6. 将水果沙拉盛入杯子或碗中，如需装饰，可再撒少许新鲜欧芹或薄荷

这道水果沙拉非常适合作为健康零食或配菜。利用冰箱中丰富的食材，可以轻松快速地完成这道清爽美味的料理。



User



我来这里参观需要注意什么



LLaVA

撰写一篇引人入胜的旅行博客文章，讲述我这次旅行到这个地方的经历，突出美丽的风景、愉快的体验，并包括可能的趣闻轶事。



User



LLaVA

前往这个风景如画的地方，背景是宁静的湖泊和群山，是一次令人耳目一新、身心焕发的体验。伸向平静水面的码头提供了一个绝佳的观景台，让人能够欣赏到令人叹为观止的景色。当我站在码头上时，不禁感受到与大自然的深刻联系，停下来片刻，沉浸在宁静的氛围中。

此次旅行的亮点之一无疑是钓鱼体验。这个地区生机勃勃，尽管湖面不大，我还是很高兴钓到了一条不小的鱼。在如此美丽的环境中钓鱼的乐趣真是令人满足，也让我得以从繁忙的生活中抽身，真正沉浸在大自然的宁静之中。

在探索周围的区域时，我遇到了一群在码头附近筑巢的鸟，为这个地方增添了独特的魅力。与野生动物互动并观察它们的行为是一种充实的体验，我将珍藏多年。总之，这个田园般的目的地给了我一个放松和充电的机会，四周环绕着迷人的自然美景和水声、野生动植物的轻柔声音。如果你想寻找一个远离日常喧嚣的宁静之地，这个俯瞰湖泊的码头是寻找安慰并重新与自然连接的完美地方。

图 3：LLaVA能够根据用户的意图识别视觉内容，而无需直接提示进行视觉识别。当收到后续请求时，它还能够提供详细的响应，并且生成的响应与提供的视觉内容密切相关。

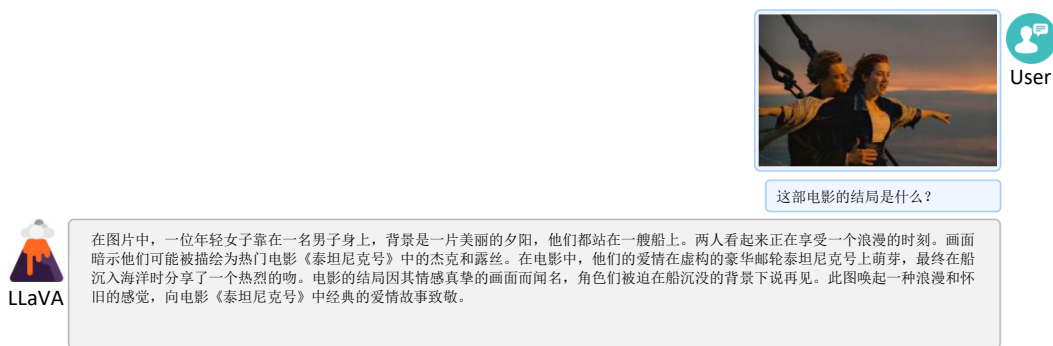


图 4：LLaVA 将电影场景与预训练大语言模型的文本知识关联起来。



图 5：LLaVA 能识别出列奥纳多·达·芬奇的著名艺术作品《蒙娜丽莎》。当我们开始新对话时，它也会解释网络上创作的模仿《蒙娜丽莎》的幽默作品。



图 6：LLaVA的一个有趣的涌现行为是它能够识别埃隆·马斯克，无论是在头像中还是在他打扮成Doge的幽默表情包中。这意味着预训练的CLIP视觉编码器可能见过埃隆·马斯克的图片。然而，这仍然令人惊讶，因为埃隆·马斯克从未出现在LLaVA的视觉特征对齐或视觉指令微调阶段的训练数据中，这表明基础语言模型能够推广到未见过的视觉概念。

Question:

这把摇椅是用什么材料做的？

选项：(A) 木头 (B) 丝绸

LLaVA answer:

讲义：材料是一种物质。

木材、玻璃、金属和塑料都是常见的材料。

有些物体仅由一种材料制成。

例如大多数钉子是用金属制成的。

而其他物体则由多种材料组合而成。

比如这把锤子就是由金属和木材制成的。

解答：观察摇椅的图片。这把摇椅由两种不同的材料制成：椅腿是木制的，而椅背和座面则是丝绸材质。

因此，正确答案是 B。

GPT-4 (text-only) answer:

答案是A。因为：摇椅通常由坚固的材料制成，能够承受人的重量并承受摇动。木材是制作摇椅的常用材料，因为它坚固、耐用且外观自然。



GPT-4 (text-only) judge:

由于没有上下文或图片可供参考，我无法独立分析所讨论的摇椅。

但是，我可以尝试分析助教们给出的解释。

助教1声称看过图片，并提到摇椅的椅腿是木制的，椅背和椅面是丝绸的，暗示它是两种材料的组合。

然而，这种说法似乎不太合理，因为丝绸由于稳定性和耐用性问题，并非摇椅的常用材料。通常情况下，摇椅的框架、椅腿和椅面或椅背会由木材制成，而坐垫和填充物则使用棉或涤纶等织物。

助教2认为，摇椅通常由坚固的材料制成，而木材因其强度高、耐用且外观自然，是制作摇椅的常用材料。这种说法是正确的，并且基于常识。

根据现有信息，最合理的答案是摇椅是由木头制成的，正如助手 1 所建议的那样。答案是 A。

表 10：一个示例说明纯文本 GPT-4 如何作为裁判来整合 LLaVA 和纯文本 GPT-4 的预测，并给出正确的最终答案。

C 训练细节

我们使用经过过滤的 CC-595K 子集对模型进行 1 个 epoch 的预训练，学习率为 $2e-3$ ，批大小为 128。然后，我们使用提出的 LLaVA-Instruct-158K 数据集进行 3 个 epoch 的微调，学习率为 $2e-5$ ，批大小为 32。与 Vicuna 类似，我们使用 Adam 优化器，不启用权重衰减，并采用余弦学习率和 3% 的预热比。在微调过程中，我们使用 FSDP（全分片数据并行）和梯度检查点来节省 GPU 内存，并且未使用卸载。我们启用了 BF16 和 TF32 以平衡速度和精度。

所有模型均使用 8 个 A100 显卡进行训练。在 CC-595K 上的预训练耗时 4 小时。在 Instruct-158K 上的微调耗时 10 小时。在 ScienceQA 上的微调耗时 4 小时。

D 资源

我们的源代码、生成的指令调优数据和提出的基准测试都已上传到匿名化的 GitHub 存储库中：[LLaVA-Annonymous/LLaVA](#)。

1. Source Code: [link](#)
2. README: [link](#)
3. Instructions to launch the demo: [link](#)
4. All prompts and few shot examples for querying GPT-4: [link](#)
5. LLaVA-Instruct-158K: [link](#)
6. LLaVA-Bench: [COCO](#), [In-The-Wild](#)
7. 模型检查点。压缩后的模型检查点大小为 25GB，超过了 GitHub LFS（大型文件存储）5GB 的限制。我们将向公众发布该检查点，或应本次提交审阅者的要求发布。

E 数据

图像简要描述说明。用于简要描述图像内容的说明列表如表 11 所示。这些说明含义相同，但语言表达方式略有不同。

- “准确描述这张图片。”
- “简洁地介绍下所给的图片。”
- “请对所展示的图片作简要说明。”
- “总结这张图片的视觉内容。”
- “请对下图进行简明扼要的解释。”
- “请对所提供的图片进行简明扼要的解读。”
- “简明扼要地描述照片的主要特征。”
- “简明扼要地描述图中所示内容。”
- “对照片进行清晰简洁的描述。”
- “请用简洁但信息丰富的语言描述这张图片。”
- “用简洁的语言描述所呈现的图像。”

表 11：简要图像描述说明列表。

图像详细描述说明。用于详细描述图像内容的说明列表如表 12 所示。这些说明含义相同，但语言表达方式略有不同。

CC3M。我们使用 Spacy 对整个 CC3M 数据集集中的每个标题提取名词短语，并统计每个唯一名词短语的出现频率。我们跳过出现频率小于 3 的名词短语，因为它们通常是已经涵盖过的罕见概念和属性组合。

- “请详细描述以下图片。”
- “请提供给定图像的详细描述。”
- “请详细描述你所看到的图像。”
- “分享对所展示图像的全面概述。”
- “对图像进行全面分析。”
- “在你开始之前，请解释图像的各个方面。”
- “详细阐明所显示图像的内容。”
- “用详尽的描述来刻画图像的特征。”
- “详细分解图像的各个元素。”
- “逐一分析图像的重要细节。”
- “用丰富、生动的叙述来描绘这幅画面。”
- “准确描述图像内容。”
- “以全面、详细的方式分析图像。”
- “通过描述性解释来阐释图片。”
- “仔细观察图片并分享其细节。”
- “请对给定图像进行详尽的描述。”

表 12：详细图像描述的说明列表。

我们首先筛选出其他标题。然后，我们从剩余频率最低的名词短语开始，将包含该名词短语的标题添加到候选库中。如果名词短语的频率大于 100，则从其所有标题中随机抽取一个大小为 100 的子集。这样最终得到大约 59.5 万个图像-文本对。

图 7 显示了 CC3M 过滤前后名词短语统计数据的比较。过滤后的数据集对频率高于 3 的概念覆盖良好，但图像-文本对的数量较少。

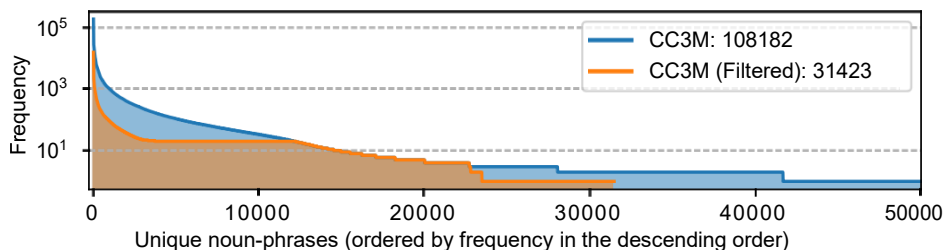


图 7：CC3M 过滤前后名词短语统计数据的比较。图例中报告了唯一名词短语的总数。

F 提示词

表 13 显示了 ChatGPT/GPT-4 生成基于图像的对话所使用的提示。

```
messages = [ {"role": "system", "content": f"" 你是人工智能视觉助手，你正在看到
一张图片。你看到的图片会附带五个句子，描述这张图片。请在观看图片的同时回答所有
问题。"" }
```

设计一段你和询问这张照片的人之间的对话。回答的语气应该像是视觉AI助手在看图回答问题。提出各种各样的问题，并给出相应的回答。

问题应包含有关图像视觉内容的描述，例如物体类型、物体数量、物体动作、物体位置、物体间的相对位置等。问题必须有明确的答案：

- (1) 能够从图像中清晰地看到问题所询问的内容并能给出确切答案；
- (2) 能够从图像中明确判断问题所询问的内容不在图像中。请勿提出任何无法给出确切答案的问题。

此外，还要提出与图片内容相关的复杂问题，例如，询问图片中物体的背景知识，引导讨论图片中发生的事件等等。再次强调，不要询问不确定的细节。回答复杂问题时，要提供详细的答案。例如，给出详细的例子或推理步骤，使内容更具说服力且条理清晰。如有必要，可以包含多个段落。

```
for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']} )
messages.append({"role": "user", "content": '\n'.join(query)})
```

表 13：对于每个查询，我们展示了 ChatGPT/GPT-4 如何构建提示，以使用少样本上下文学习从查询['context']中收集查询['response']。其中，示例来自 `fewshot_samples`，每个示例包含输入样本 `sample['context']` 和输出样本 `sample['response']`。请注意，`messages` 是最终提示。在本例中，我们提供了用于生成对话回复的提示，详情请参见表 15 和表 16 中的上下文学习示例。我们建议读者查看代码库，了解生成其他两种类型回复（包括详细描述和复杂推理）的提示。

Context type 1: Captions

一群人站在一辆黑色汽车外，车里堆满了行李。
在地下停车场，行李包围着一辆车。人们试图把所有行李

都塞进一辆SUV里。
这辆运动型多用途车停在公共停车场，人们正在打包准备出行。一些人带着行李，靠近一辆运送行李的货车。



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Response type 1: conversation

问：图片中是什么类型的车辆？
答：图片中是一辆黑色运动型多用途车（SUV）。
问：这辆车停在哪里？
答：这辆车停在地下停车场，很可能是公共停车场。
问：图片中的人在做什么？
答：他们正在把所有行李塞进SUV里，很可能是在准备旅行。

Response type 2: detailed description

图片显示的是一个地下停车场，一辆黑色运动型多用途车（SUV）停在那里。画面中有三人，一人靠近车辆左侧，一人位于中间，另一人位于右侧。他们正在一起将行李装进SUV，准备出行。

停车场内，各种行李散落在车辆周围。有两个背包，一个靠近左后轮，另一个靠近车辆右侧。此外，还有两个行李箱，一个在车辆右侧，另一个则位于停车场中央附近。车辆左侧还停着一辆自行车。

其他车辆也停放在SUV周围，一辆车位于其后方偏左的位置，另一辆车位于其后方偏右的位置，第三辆车则位于其后方更靠右的位置。

Response type 3: complex reasoning

问：这些人面临哪些挑战？
答：图中，一群人站在停车场一辆黑色SUV旁，周围堆满了各种行李，包括行李箱和背包。他们面临的挑战是如何将所有行李塞进这辆黑色SUV。由于需要装下的行李箱和背包数量众多，这表明他们携带的物品相当可观。他们可能需要制定策略，合理安排行李，确保所有物品都能妥善装入车内。此外，他们还需要考虑乘客的舒适度和驾驶视野，因此行李的摆放位置不能阻挡驾驶员的视线，也不能让乘客在旅途中感到不适。

表 14：一个用于说明指令跟随数据的示例。上部方框显示了用于提示 GPT 的上下文信息，例如标题和对话框，下部方框显示了三种类型的响应。请注意，此处展示的图像并非用于提示 GPT，仅作参考。

Captions

电影院门口上方张贴着放映时间表。一个红色的消防栓深深地埋在雪里。
消防栓位于一条刚被清理过的路边人行道附近，被雪埋住了。
这座城市今年冬天下下了很大的雪。
冬天，一家狗狗旅馆被雪覆盖。

Conversation

Question:

图中的消防栓是什么颜色的？

====

Answer:

图中消防栓是红色的。

====

Question:

消防栓是完全被雪覆盖还是部分可见？

====

Answer:

消防栓部分可见，大部分被厚厚的积雪覆盖。

====

Question:

消防栓附近的人行道状况如何？

====

Answer:

消防栓附近的人行道最近刚被铲过雪，所以与周围地区相比，积雪相对较少。

====

Question:

这张图片展现了这座城市冬季的哪些景象吗？

====

Answer:

图片显示，这座城市的冬季气候十分严寒，充满挑战。图中可见的大量积雪表明，该市经历了强降雪，这可能会给居民带来各种困难和不便。

例如，消防栓上方厚厚的积雪可能会在紧急情况下阻碍其使用，从而造成安全隐患。最近清理过的行人道表明，市政维护人员一直在积极努力保持人行道的畅通和安全，但如此大的降雪量可能会使持续的除雪工作难以跟上。

此外，极端冬季天气会影响交通运输，导致交通拥堵、公共交通延误，以及因路面湿滑而增加事故风险。它还会给企业和学校带来问题，因为人们通勤上班或上课可能会遇到困难。此外，大雪还会给屋顶和电力线路等基础设施带来额外压力，增加结构损坏或停电的可能性。

总之，照片中深埋雪中的红色消防栓和刚刚清理过的路边人行道表明，这座城市经历了异常严寒的冬季，大量的降雪可能给居民和基础设施带来了各种挑战和干扰。

表 15：用于情境学习构建视觉对话数据的一个示例。

Captions

一名男子在开阔的雪山上滑雪。一名滑雪者正滑入雪中。
一名滑雪者在巨大山脉的山麓。
一名滑雪者站在雪道指示牌附近眺望群山。
一个人滑雪经过一片灌木丛生的区域。

Conversation

Question:

这个滑雪者在干什么？

===

Answer:

滑雪者正在白雪皑皑的山坡上滑行，穿梭在雪地中，欣赏着美丽的群山景色。他们还站在一块指示牌附近，指示牌标明了他们在山上的位置。

===

Question:

这张照片中能看到其他的人或者滑雪者吗？

===

Answer:

不，图片中只有一个人在滑雪，而且他周围有很多灌木丛。

===

Question:

这个人是在进行哪种滑雪？越野滑雪还是高山滑雪？

===

Answer:

根据照片内容，无法确定滑雪者是在进行越野滑雪还是高山滑雪。但是，考虑到有雪道标志，且滑雪者是在山麓地带滑雪，因此他更有可能是在进行越野滑雪。高山滑雪通常在更陡峭的山坡上进行，并且需要借助缆车；而越野滑雪则通常在相对平坦的地形上进行，并设有标记清晰的雪道。

===

Question:

你能再详细解释一下越野滑雪和高山滑雪吗？

===

Answer:

越野滑雪和高山滑雪是两种截然不同的滑雪方式，各有其独特的技术、装备和地形偏好。

1. 越野滑雪：

越野滑雪，又称北欧式滑雪，是指使用滑雪板和雪杖在相对平坦的地形或缓坡上滑行。它是一项广受欢迎的休闲和竞技运动，注重耐力、平衡和有氧运动能力。越野滑雪者依靠自身动力在雪地上滑行，而不是依赖缆车或其他辅助设备。越野滑雪主要有两种技术：

- 传统式：这种技术要求滑雪者以平行轨迹滑行，通过前后滑行来保持双板平行。滑雪者使用雪杖推动身体前进。
- 自由式：这种技术类似于滑冰，滑雪者用一只滑雪板斜向蹬地，另一只滑雪板滑行。它需要平整的雪道，通常比传统式速度更快。

越野滑雪装备包括轻便的滑雪板、雪鞋、固定器和雪杖。这种靴子比高山滑雪靴更灵活，可以更好地活动脚踝并控制脚部。

2. 高山滑雪：

高山滑雪，又称速降滑雪，是指滑雪者使用滑雪板和雪杖以高速滑下山坡，保持平衡和控制。这项运动更注重速度、技巧以及应对各种复杂地形，包括陡坡、雪包，甚至跳台。速降滑雪可以进一步细分为多个项目，例如回转、大回转、超级大回转和速降赛。每个项目都有其自身的规则、赛道和滑雪装备。

速降滑雪的装备包括比越野滑雪更重、更硬的滑雪板、雪鞋、固定器和雪杖。雪鞋更加坚固，以便在高速滑行和急转弯时提供更好的支撑和控制。

总而言之，越野滑雪是一项耐力型运动，主要在平坦或缓坡地形上滑行；而速降滑雪则更注重速度和技巧，滑雪者需要应对更陡峭的坡度和复杂的地形。这两种运动都需要专门的器材和技巧，但它们为参与者提供了不同的体验和挑战。

表 16：用于情境学习构建视觉对话数据的一个示例。