Improved Baselines with Visual Instruction Tuning

Haotian Liu[1] Chunyuan Li[2] Yuheng Li[1] Yong Jae Lee[1]
[1]University of Wisconsin–Madison [2]Microsoft Research, Redmond
https://llava-vl.github.io

Abstract

Large multimodal models (LMM) have recently shown encouraging progress with visual instruction tuning. In this paper, we present the first systematic study to investigate the design choices of LMMs in a controlled setting under the LLaVA framework. We show that the fully-connected vision-language connector in LLaVA is surprisingly powerful and data-efficient. With simple modifications to LLaVA, namely, using CLIP-ViT-L-336px with an MLP projection and adding academic-task-oriented VQA data with response formatting prompts, we establish stronger baselines that achieve state-of-the-art across 11 benchmarks. Our final 13B checkpoint uses merely 1.2M publicly available data, and finishes full training in ∼1 day on a single 8-A100 node. Furthermore, we present some early exploration of open problems in LMMs, including scaling to higher resolution inputs, compositional capabilities, and model hallucination, etc. We hope this makes state-of-the-art LMM research more accessible. Code and model will be publicly available.

1. Introduction

Large multimodal models (LMMs) have become increasingly popular in the research community, as they are the key building blocks towards general-purpose assistants [1, 28, 41]. Recent studies on LMMs are converging on a central concept known as visual instruction tuning [34]. The results are promising, *e.g*. LLaVA [34] and MiniGPT-4 [59] demonstrate impressive results on natural instruction-following and visual reasoning capabilities. To better understand the capability of LMMs, multiple benchmarks [16, 26, 32, 35, 52] have been proposed. Recent works further demonstrate improved performance by scaling up the pretraining data [2, 13, 51], instruction-following data [13, 17, 27, 55], visual encoders [2], or language models [37], respectively. The LLaVA architecture is also leveraged in different downstream tasks and domains, including region-level [7, 53] and pixel-level [25, 48] understanding, biomedical assistants [29], image generation [4], adversarial studies [5, 56].

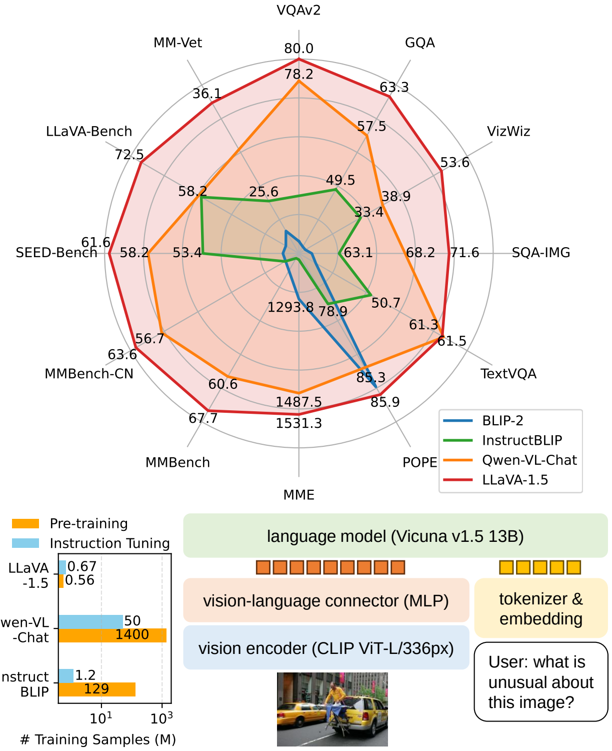


Figure 1. **LLaVA-1.5** achieves SoTA on a broad range of 11 tasks (Top), with high training sample efficiency (Left) and simple modifications to LLaVA (Right): an MLP connector and including academic-task-oriented data with response formatting prompts.

However, despite many benchmarks and developments, it still remains unclear what the best recipe is to train LMMs towards the goal of general-purpose assistants. For example, LLaVA [34] excels in conversational-style visual reasoning and even outperforms later approaches like InstructBLIP [13] on such benchmarks [52], while InstructBLIP excels in traditional VQA benchmarks that demands single-word or short answers. Given the significant differences in the model architecture and training data between them, the root cause of the disparity in their capabilities remains elusive, despite conjectures [35, 52]: the amount of training data, the usage of resamplers like Qformer [30], *etc*. To this end, we present the first systematic study to investigate the design choices of LMMs in a controlled setting. Our study originates from LLaVA and builds a road map by carefully making effective contributions from the perspectives of the input, model, and data.

First, we unveil that the fully-connected vision-language connector in LLaVA is surprisingly powerful and data-efficient, and we establish stronger and more feasible baselines built upon the LLaVA framework. We report that two simple improvements, namely, an MLP cross-modal connector and incorporating academic task related data such as VQA, are orthogonal to the framework of LLaVA, and when used with LLaVA, lead to better multimodal understanding capabilities. In contrast to InstructBLIP [13] or Qwen-VL [2], which trains specially designed visual resamplers on hundreds of millions or even billions of image-text paired data, LLaVA uses one of the simplest architecture design for LMMs and requires only training a simple fully-connected projection layer on merely 600K image-text pairs. Our final model can finish training in $\sim$1 day on a single 8-A100 machine and achieves state-of-the-art results on a wide range of benchmarks. Moreover, unlike Qwen-VL [2] that includes in-house data in training, LLaVA utilizes only publicly available data.

Next, we delve into an early exploration of other open problems of large multimodal models. Our findings include: (1) Scaling to high-resolution image inputs. We show that LLaVA's architecture is versatile in scaling to higher resolutions by simply dividing images into grids and maintains its data efficiency; with the increased resolution, it improves the model's detailed perception capabilities and reduces hallucination. (2) Compositional capabilities. We find that large multimodal models are capable of generalizing to compositional capabilities. For example, training on long-form language reasoning together with shorter visual reasoning can improve the model's writing capability for multimodal questions. (3) Data efficiency. We show that randomly downsampling LLaVA's training data mixture by up to 75% does not significantly decrease the model's performance, suggesting that the possibility of a more sophisticated dataset compression strategy can further improve LLaVA's already efficient training pipeline. (4) Data scaling. We provide empirical evidence for the scaling of data granularity in conjunction with the model's capability is crucial for an improved capability without introducing artifacts like hallucination.

In sum, we perform a systematic study on the training of large multimodal models, and introduce a simple yet effective approach to balance the multitask learning and effective scaling for large multimodal models. Our improved baselines, LLaVA-1.5, uses only *public* data, achieves the state-of-the-art on a broad range of 11 tasks, and is significantly more data-efficient than previous approaches. By rethinking the conventional approaches and exploring the open problems in visual instruction tuning, we pave the way for more robust and capable systems for LMMs. We hope these improved and easily-reproducible baselines will provide a reference for future research in open-source LMMs.

2. Related Work

Instruction-following large multimodal models (LMMs). Common architectures include a pre-trained visual backbone to encode visual features, a pre-trained large language model (LLM) to comprehend the user instructions and produce responses, and a vision-language cross-modal connector to align the vision encoder outputs to the language models. As shown in Fig. 1, LLaVA [34] is perhaps the simplest architecture for LMMs. Optionally, visual resamplers (*e.g*. Qformer [30]) are used to reduce the number of visual patches [2, 13, 59]. Training an instruction-following LMM usually follows a two-stage protocol. First, the vision-language alignment pretraining stage leverages image-text pairs to align the visual features with the language model's word embedding space. Earlier works utilize relatively few image-text pairs (*e.g*. $\sim$600K [34] or $\sim$6M [59]), while some recent works pretrain the vision-language connector for a specific language model on a large amount of image-text pairs (*e.g*. 129M [13] and 1.4B [2]), to maximize the LMM's performance. Second, the visual instruction tuning stage tunes the model on visual instructions [34], to enable the model to follow users' diverse requests on instructions that involve the visual contents.

Multimodal instruction-following data. In NLP, studies show that the quality of instruction-following data largely affects the capability of the resulting instruction-following models [58]. For visual instruction tuning, LLaVA [34] is the pioneer to leverage text-only GPT-4 to expand the existing COCO [33] bounding box and caption dataset to a multimodal instruction-following dataset that contains three types of instruction-following data: conversational-style QA, detailed description, and complex reasoning. LLaVA's pipeline has been employed to expand to textual understanding [54], million-scales [55], and region-level conversations [7]. InstructBLIP [13] incorporates academic-task-oriented VQA datasets to further enhance the model's visual capabilities. Conversely, [6] identifies that such naive data merging can result in models that tend to overfit to VQA datasets and thus are unable to participate in natural conversations. The authors further propose to leverage the LLaVA pipeline to convert VQA datasets to a conversational style. While this proves effective for training, it introduces added complexities in data scaling. However, in NLP, the FLAN family [12, 49] shows that adding a large number of academic language tasks for instruction tuning can effectively improve the generalization ability. In light of this, we consider investigating the root cause of the inability to balance between natural conversations and academic tasks in multimodal models.

3. Approach

3.1. Preliminaries

As the seminal work of visual instruction tuning, LLaVA [34] showcases commendable proficiency in visual reasoning capabilities, surpassing even more recent models on diverse benchmarks [3, 52] for real-life visual instruction-following tasks. LLaVA uses a single linear layer to project the visual features to language space, and optimizes the whole LLM for visual instruction tuning. However, LLaVA falls short on academic benchmarks that typically require short-form answers (*e.g.* single-word), and tends to answer *yes* for yes/no questions due to the lack of such data in the training distribution.

On the other hand, InstructBLIP [13] is the pioneer to incorporate academic-task-oriented datasets like VQA-v2 [18] along with LLaVA-Instruct [34], and demonstrates improved performance on VQA benchmarks. It pretrains Qformer [30] on 129M image-text pairs and only finetunes the instruction-aware Qformer for visual instruction tuning. However, recent studies [6, 52] show that it does not perform as well as LLaVA on engaging in real-life visual conversation tasks. More specifically, as shown in Table 1a, it can overfit to VQA training sets with short-answers, even on requests that require detailed responses.

3.2. Response Format Prompting

We find that the inability [6] to balance between short- and long-form VQA for approaches like InstructBLIP [13], which leverages instruction following data that includes both natural responses and short-answers, is mainly due to the following reasons. First, *ambiguous prompts on the response format*. For example, *Q: {Question} A: {Answer}*. Such prompts do not clearly indicate the desired output format, and can overfit an LLM behaviorally to short-form answers even for natural visual conversations. Second, *not finetuning the LLM*. The first issue is worsened by InstructBLIP only finetuning the Qformer for instruction-tuning. It requires the Qformer's visual output tokens to control the length of the LLM's output to be either long-form or short-form, as in prefix tuning [31], but Qformer may lack the capability of properly doing so, due to its limited capacity compared with LLMs like LLaMA.

Thus, to enable LLaVA to better handle short-form answers while addressing the issues of InstructBLIP, we propose to use a single response formatting prompt that clearly indicates the output format. It is appended at the end of VQA questions when promoting short answers: *Answer the question using a single word or phrase*. We find that when the LLM is *finetuned* with such prompts, LLaVA is able to properly adjust the output format according to the user's instructions (see Table 1b), and does not require additional processing of the VQA answers using ChatGPT [6], which further enables scaling to various data sources. As shown in Table 2, by merely including VQAv2 [18] in training, LLaVA's performance on MME significantly improves (1323.8 *vs* 809.6) and outperforms InstructBLIP by 111 points.

Visual input example, Multitask Balancing Problem:

User	Is this unusual? Please explain in detail.
InstructBLIP	yes

(a) Example of InstructBLIP [13] (Vicuna-13B) having difficulty balancing between short- and long-form answers.

Visual input example, Different Format Prompts:	
Normal prompt	What is the color of the shirt that the man is wearing?
Response	The man is wearing a yellow shirt.
Ambiguous prompt	Q: What is the color of the shirt that the man is wearing? A:
Response	The man is wearing a yellow shirt.
Formatting prompt	What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**
Response	Yellow.

(b) Comparison of how different prompts regularize the output format. The results are obtained zero-shot directly after LLaVA undergoes the first-stage vision-language alignment pretraining, without the second-stage visual instruction tuning.

Table 1. Visual input example to illustrate the challenge of (a) multitask balancing and (b) different format prompts. The same image input is used.

3.3. Scaling the Data and Model

MLP vision-language connector. Inspired by the improved performance in self-supervised learning by changing from a linear projection to an MLP [8, 9], we find that improving the vision-language connector's representation power with a two-layer MLP can improve LLaVA's multimodal capabilities, compared with the original linear projection design.

Academic task oriented data. We further include additional academic-task-oriented VQA datasets for VQA, OCR, and region-level perception, to enhance the model's capabilities in various ways, as shown in Table 2. We first include four additional datasets that are used in InstructBLIP: open-

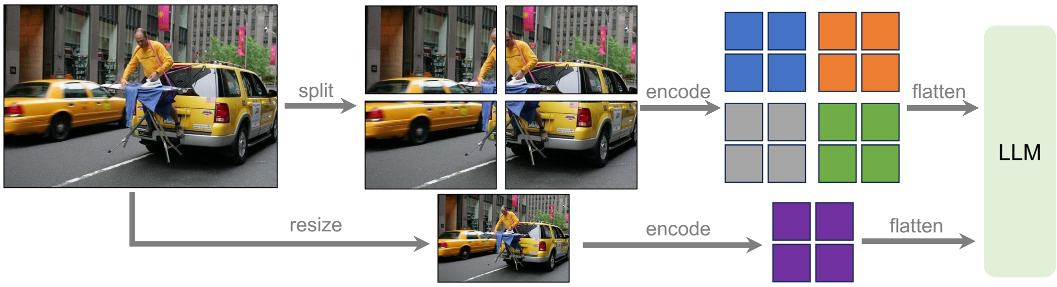


Figure 2. **LLaVA-1.5-HD.** Scaling LLaVA-1.5 to higher resolutions by splitting the image into grids and encoding them independently. This allows the model to scale to any resolution, without performing positional embedding interpolation for ViTs. We additionally concatenate the feature of a downsampled image to provide the LLM with a global context.

Method	LLM	Res.	GQA	MME	MM-Vet
InstructBLIP	14B	224	49.5	1212.8	25.6
Only using a subset of InstructBLIP training data					
0 **LLaVA**	7B	224	–	809.6	25.5
1 +VQA-v2	7B	224	47.0	1197.0	27.7
2 +Format prompt	7B	224	46.8	1323.8	26.3
3 +MLP VL connector	7B	224	47.3	1355.2	27.8
4 +OKVQA/OCR	7B	224	50.0	1377.6	29.6
Additional scaling					
5 +Region-level VQA	7B	224	50.3	1426.5	30.8
6 +Scale up resolution	7B	336	51.4	1450	30.3
7 +GQA	7B	336	62.0*	1469.2	30.7
8 +ShareGPT	7B	336	62.0*	1510.7	31.1
9 +Scale up LLM	13B	336	**63.3***	**1531.3**	**36.1**

Table 2. **Scaling results** on 🟦 data, 🟥 model, and 🟨 resolution. We choose to conduct experiments on GQA [20], MME [16], and MM-Vet [52] to examine the representative capabilities of VQA with short answers, VQA with output formatting, and natural visual conversations, respectively. *Training images of GQA were observed during training.

knowledge VQA (OKVQA [39], A-OKVQA [43]) and OCR (OCRVQA [40], TextCaps [45]). A-OKVQA is converted to multiple choice questions and a specific response formatting prompt is used: *Answer with the option's letter from the given choices directly*. With only a subset of the datasets InstructBLIP uses, LLaVA already surpasses it on all three tasks in Table 2, suggesting LLaVA's effective design. Furthermore, we find further adding region-level VQA datasets (Visual Genome [24], RefCOCO [23, 38]) improves the model's capability of localizing fine-grained visual details.

Additional scaling. We further scale up the input image resolution to 336^2 to allow the LLM to clearly "see" the details of images, by swapping the vision encoder to CLIP-ViT-L-336px (the highest resolution available for CLIP). In addition, we add the GQA dataset as an additional visual knowledge source. We also incorporate ShareGPT [44] data and scale up the LLM to 13B as in [2, 7, 37]. Results on MM-Vet shows the most significant improvement when scaling the LLM to 13B, suggesting the importance of the base LLM's capability for visual conversations.

LLaVA-1.5. We denote this final model with all the modifications as LLaVA-1.5 (the last two rows in Table 2), which achieves an impressive performance that significantly outperforms the original LLaVA [34].

Computational cost. For LLaVA-1.5, we use the same pretraining dataset, and keep the training iterations and batch size roughly the same for instruction tuning as LLaVA [34]. Due to the increased image input resolution to 336^2, the training of LLaVA-1.5 is $\sim 2\times$ as long as LLaVA: $\sim$6 hours of pretraining and $\sim$20 hours of visual instruction tuning, using 8× A100s.

3.4. Scaling to Higher Resolutions

In Sec. 3.3, we observe the advantage that scaling up the input image resolution improves the model's capabilities. However, the image resolution of the existing open source CLIP vision encoders is limited to 336^2, preventing the support of higher resolution images by simply replacing the vision encoder as we did in Sec. 3.3. In this section, we present an early exploration of scaling the LMM to higher resolutions, while maintaining the data efficiency of LLaVA-1.5.

When using ViT [14] as the vision encoder, to scale up the resolution, previous approaches mostly choose to perform positional embedding interpolation [2, 30] and adapt the ViT backbone to the new resolution during finetuning. However, this usually requires the model to be finetuned on a large-scale image-text paired dataset [2, 30], and limits the resolution of the image to a fixed size that the LMM can accept during inference.

Instead, as shown in Fig. 2, we overcome this by dividing the image into smaller image patches of the resolution that the vision encoder is originally trained for, and encode them independently. After obtaining the feature maps of individual patches, we then combine them into a single large feature map of the target resolution, and feed that into the LLM. To

Method	LLM	Image Size	Sample Size Pretrain	Sample Size Finetune	VQAv2 [18]	GQA [20]	VisWiz [19]	SciQA-IMG [36]	TextVQA [46]
BLIP-2 [30]	Vicuna-13B	224^2	129M	-	65.0	41	19.6	61	42.5
InstructBLIP [13]	Vicuna-7B	224^2	129M	1.2M	–	49.2	34.5	60.5	50.1
InstructBLIP [13]	Vicuna-13B	224^2	129M	1.2M	–	49.5	33.4	63.1	50.7
Shikra [7]	Vicuna-13B	224^2	600K	5.5M	77.4*	–	–	–	–
IDEFICS-9B [21]	LLaMA-7B	224^2	353M	1M	50.9	38.4	35.5	–	25.9
IDEFICS-80B [21]	LLaMA-65B	224^2	353M	1M	60.0	45.2	36.0	–	30.9
Qwen-VL [2]	Qwen-7B	448^2	1.4B†	50M†	78.8*	59.3*	35.2	67.1	**63.8***
Qwen-VL-Chat [2]	Qwen-7B	448^2	1.4B*	50M†	78.2*	57.5*	38.9	68.2	61.5*
LLaVA-1.5	Vicuna-7B	336^2	**558K**	**665K**	78.5*	62.0*	50.0	66.8	58.2
LLaVA-1.5	Vicuna-13B	336^2	**558K**	**665K**	**80.0***	**63.3***	**53.6**	**71.6**	61.3
LLaVA-1.5-HD	Vicuna-13B	448^2	**558K**	**665K**	**81.8***	**64.7***	**57.5**	71.0	62.5
Specialist SOTA: PaLI-X-55B [10]					86.1*	72.1*	70.9*	–	71.4*

Table 3. **Comparison with SoTA methods on academic-task-oriented datasets.** LLaVA-1.5 achieves the best performance on 4/5 benchmarks, and ranks the second on the other. *The training images/annotations of the datasets are observed during training. †Includes in-house data that is not publicly accessible.

Method	POPE [32] rand	POPE [32] pop	POPE [32] adv	MME [16]	MMBench [35] en	MMBench [35] cn	SEED-Bench [26] all	SEED-Bench [26] img	SEED-Bench [26] vid	LLaVA-Wild [34]	MM-Vet [52]
BLIP2-14B [30]	**89.6**	85.5	80.9	1293.8	–	–	46.4	49.7	36.7	38.1	22.4
InstructBLIP-8B [13]	–	–	–	–	36	23.7	53.4	58.8	38.1	60.9	26.2
InstructBLIP-14B [13]	87.7	77	72	1212.8	–	–	–	–	–	58.2	25.6
Shikra-13B [7]	–	–	–	–	58.8	–	–	–	–	–	–
IDEFICS-9B [21]	–	–	–	–	48.2	25.2	–	44.5	–	–	–
IDEFICS-80B [21]	–	–	–	–	54.5	38.1	–	53.2	–	–	–
Qwen-VL [2]	–	–	–	–	38.2	7.4	56.3	62.3	39.1	–	–
Qwen-VL-Chat [2]	–	–	–	1487.5	60.6	56.7	58.2	65.4	37.8	–	–
LLaVA-7B [34]	76.3	72.2	70.1	809.6	38.7	36.4	33.5	37.0	23.8	62.8	25.5
LLaVA-1.5-7B	87.3	86.1	84.2	1510.7	64.3	58.3	58.6	66.1	37.3	65.4	31.1
LLaVA-1.5-13B	87.1	**86.2**	**84.5**	**1531.3**	**67.7**	**63.6**	**61.6**	**68.2**	**42.7**	**72.5**	**36.1**
LLaVA-1.5-13B-HD	87.5	**86.4**	**85.0**	1500.1	**68.8**	61.9	**62.6**	**70.1**	41.3	72.0	**39.4**

Table 4. **Comparison with SoTA methods on benchmarks for instruction-following LMMs.** LLaVA-1.5 achieves the best overall performance.

provide the LLM with the global context and to reduce the artifact of the split-encode-merge operation, we additionally concatenate the feature of a downsampled image to the merged feature map. This allows us to scale the input to any arbitrary resolution and maintain the data efficiency of LLaVA-1.5. We call this resulting model LLaVA-1.5-HD.

4. Empirical Evaluation

4.1. Benchmarks

We evaluate LLaVA-1.5 on a collection of both academic-task-oriented benchmarks and recent benchmarks specifically proposed for instruction-following LMMs, totalling 12 benchmarks. For academic-task-oriented benchmarks, VQA-v2 [18] and GQA [20] evaluate model's visual perception capabilities on open-ended short answers. VizWiz [19] contains 8,000 images to evaluate model's zero-shot generalization on visual questions asked by visually impaired people. Following InstructBLIP [13], the image subset of ScienceQA [36] with multiple choice are used to evaluate the zero-shot generalization on scientific question answering. TextVQA [46] contains text-rich visual question answering.

For recent benchmarks proposed for instruction-following LMMs, POPE [32] evaluates model's degree of hallucination on three sampled subsets of COCO [33]: random, common, and adversarial and we report the F1 score on all three splits. Other benchmarks evaluate the model's capabilities on a wide range of domains and applications, with different response formats. MME-Perception [16] evaluates model's visual perception with yes/no questions. MMBench [35] evaluates model's answer robustness with all-round shuffling on multiple choice answers. MMBench-CN [35] is the Chinese-translated version of MMBench. SEED-Bench [26] evaluates model's performance on both images and videos with multiple choice, and we sample the frame in the middle to evaluate the accuracy on videos. LLaVA-Bench-in-the-Wild [34] and MM-Vet [52] evaluate model's capabilities in engaging in visual conversations on a diverse range of tasks, and evaluates the correctness and the helpfulness of the response with GPT-4 evaluation.

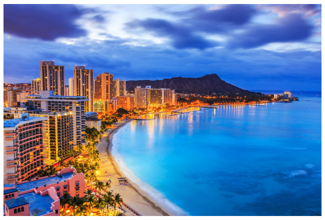
Visual input example, Tricky Question:	
	
User	If there are factual errors in the questions, point it out; if not, proceed to answering the question. What's happening in the desert?
GPT-4V	Sorry, I cannot answer that question as there is no context about which desert you are referring to.
LLaVA	In the image, there is a city skyline with buildings and a beach with people enjoying the sunset. The scene is set in a desert, which is unusual for a beach setting.
LLaVA-1.5	There are no deserts in the image. The image features a beach with palm trees, a city skyline, and a large body of water.

Table 5. LLaVA-1.5 can detect and answer tricky questions when prompted to verify the question.

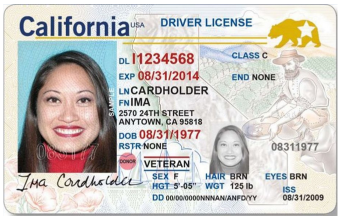
Visual input example, Constrained JSON Output:	
	 California USA DRIVER LICENSE DL I1234568 EXP 08/31/2014 LN CARDHOLDER FN IMA 2570 24TH STREET ANYTOWN, CA 95818 DOB 08/31/1977 RSTR NONE CLASS C END NONE 08311977 DONOR VETERAN SEX F HGT 5'-05" HAIR BRN WGT 125 lb EYES BRN ISS 08/31/2009 DD 00/00/0000NNNAN/ANFD/YY Ima Cardholder
User	Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead). {"class": xxx, "DLN": xxx, "DOB": xxx, "Name": xxx, "Address": xxx, "EXP": xxx, "ISS": xxx, "SEX": xxx, "HGT": xxx, "WGT": xxx, "EYES": xxx, "HAIR": xxx, "DONOR": xxx}
GPT-4V	{"class": "C", "DLN": "1234568", "DOB": "08/31/1977", "Name": "Ima Cardholder", "Address": "2570 24th Street, Anytown, CA 95818", "EXP": "08/31/2014", "ISS": "08/31/2011", "SEX": "F", "HGT": "5-05", "WGT": "125 lbs", "EYES": "BRN", "HAIR": "BRN", "DONOR": "VETERAN"}
LLaVA	{"class": "Driver License", "DLN": "CA123 456 789", "DOB": "01/01/1990", "Name": "Amy", "Address": "Anytown, CA", "EXP": "08/31/2023", "ISS": "Veteran", "SEX": "F", "HGT": "5'5", "WGT": "130", "EYES": "Brown", "HAIR": "Brown", "DONOR": "Unknown"}
LLaVA-1.5	{"class": "C", "DLN": "1132456789", "DOB": "08/23/1971", "Name": "Ima Cardholder", "Address": "Anytown, State, Zip", "EXP": "08/23/2014", "ISS": "California", "SEX": "F", "HGT": "5'5", "WGT": "123", "EYES": "Brown", "HAIR": "Brown", "DONOR": "N/A"}

Table 6. LLaVA-1.5 can extract information from the image and answer following the required format, despite a few errors compared with GPT-4V. GPT-4V results are obtained from [50].

4.2. Results

We show that LLaVA-1.5 achieves the best overall performance on 12 benchmarks, despite using magnitudes smaller pretraining and instruction tuning data compared with other methods [2, 13]. LLaVA-1.5 significantly outperforms LLaVA on all benchmarks for instruction-following LMMs. Note that it is challenging to evalute the original LLaVA on academic datasets like VQA-v2 [18] that demand open-ended short answers.

When we continue to scale up the image resolution to 448^2 with LLaVA-1.5-HD, it further improves the overall performance on all benchmarks, especially on tasks that require perception of details in the images (*e.g*. OCR in MM-Vet, detailed description in LLaVA-Bench-in-the-Wild [34]). Moreover, we find that adding the global context effectively recovers the model from the split-and-merge artifacts and guides the model to more easily locate the relevant regions from the high-resolution features (see appendix).

It is encouraging that *LLaVA-1.5 achieves the best performance with the simplest architecture, academic compute and public datasets, and yields a fully-reproducible and affordable baseline for future research*. The results also suggest that visual instruction tuning plays an important role in improving an LMM's capabilities, and raises questions upon the common belief that LMMs require significant amount of vision-language alignment pretraining [2, 13, 30], despite that the vision encoders (*e.g*. CLIP [42], OpenCLIP [22], EVA-CLIP [15], *etc*.) are already pretrained on web-scale image-text paired data. LLaVA-1.5 (even the 7B model) outperforms 80B IDEFICS [21], a Flamingo-like LMM with billions of trainable parameters for cross-modal connection. This also makes us rethink the benefits of the vision samplers and the necessity of the additional large-scale pretraining, in terms of multimodal instruction-following capabilities.

4.3. Emerging Properties

Format instruction generalization. Although LLaVA-1.5 is only trained with a limited number of format instructions, it generalizes to others. First, VizWiz [19] requires the model to output "Unanswerable" when the provided content is insufficient to answer the question, and our response format prompt (see Appendix) effectively instructs the model to do so (11.1% → 67.8% on unanswerable questions). We additionally present qualitative examples on instructing LLaVA-1.5 to verify tricky questions (Fig. 5), respond in a constrained JSON format (Fig. 6), and more in appendix.

Multilingual multimodal capability. Though LLaVA-1.5 is *not* finetuned for multilingual multimodal instruction following *at all* (all visual instructions including VQA are in English), we find that it is capable of following multilingual

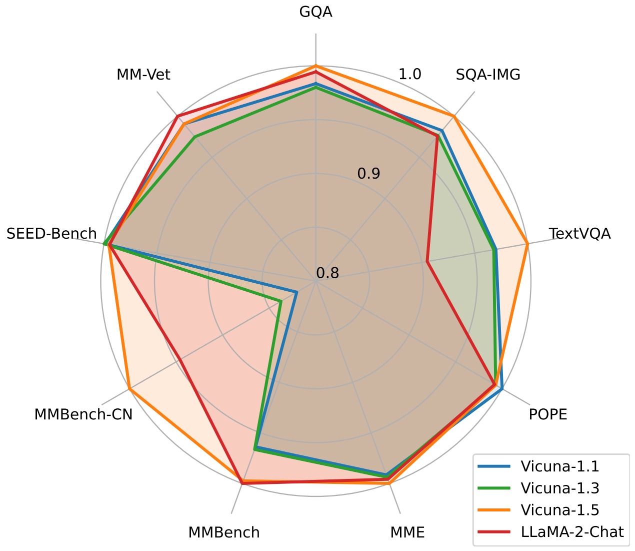


Figure 3. **Ablation on LLM choices**. Data points represent the relative performance of the best performing variant for each dataset.

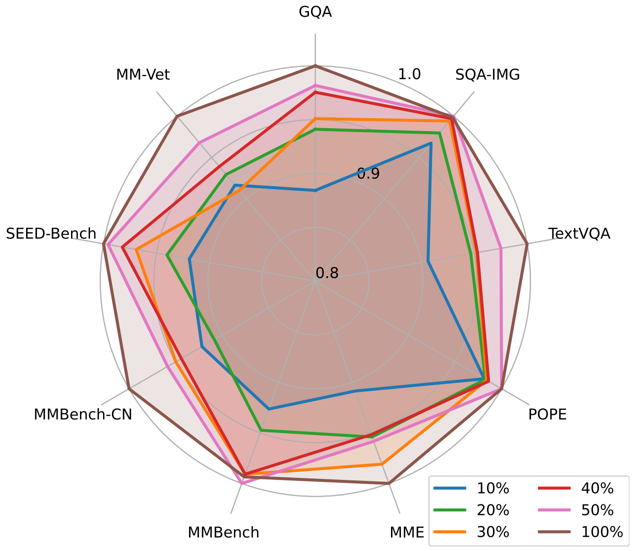


Figure 4. **Ablation on data efficiency.** Data points represent the relative performance of the best performing variant for each dataset.

instructions. This is partly due to the multilingual language instructions in ShareGPT [44]. Although ShareGPT does not contain images in its instructions, the model learns from this dataset the behavior of adaptively responding with the language that corresponds to the user's request. We empirically show that this behavior is transferred to visual conversations. We also quantitatively evaluate the model's generalization capability to Chinese on MMBench-CN [35], where the questions of MMBench are converted to Chinese. Notably, LLaVA-1.5 outperforms Qwen-VL-Chat by +7.3% (63.6% vs 56.7%), despite Qwen being finetuned on Chinese multimodal instructions while LLaVA-1.5 is not.

4.4. Ablation on LLM Choices

In NLP, findings [47] suggest that the capability of the base LLM can affect its instruction-tuned successors. In this section, we explore two families of LLMs and study their contribution to the final model's multimodal capability: LLaMA-based (Vicuna-v1.1, Vicuna-v1.3) and LLaMA-2-based (Vicuna-v1.5, LLaMA-2-Chat). Vicuna-v1.3 and Vicuna-v1.5 use the same $\sim$150K ShareGPT [44] data ($2\times$ that used in v1.1). Unlike Vicuna series that is only trained with supervised instruction finetuning (SFT), LLaMA-2-Chat is further optimized with reinforcement-learning from human-feedback (RLHF). We visualize the relative performance of these variants in Fig. 3.

First, we find that Vicuna-v1.5 achieves the best overall performance, and LLaMA-2-based models generally perform better than LLaMA-1-based, suggesting the importance of the base language model. This is further evidenced by the results on MMBench-CN [35]: despite Vicuna-v1.3 and v1.5 using the same ShareGPT data for instruction tuning, the performance in generalization to Chinese of Vicuna-v1.3 is significantly worse than v1.5.

Second, language instruction-tuning matters on specific capabilities that are required by each dataset. For example, although LLaMA-2-Chat and Vicuna-v1.5 achieves almost the same performance on MMBench, the generalization to MMBench-CN [35] of LLaMA-2-Chat is worse than Vicuna-v1.5, which is partly due to that the most SFT/RLHF data of LLaMA-2-Chat is in English and does not contain as many multilingual data as in ShareGPT. Furthermore, TextVQA requires both the model's capability of identifying the text characters in the images, and also processing the noisy outputs from the OCR engine; such noise *may* be more commonly observed in the ShareGPT data, which is collected in-the-wild from daily usage of ChatGPT.

5. Open Problems in LMMs

Given the successful scaling of LLaVA-1.5, we conduct additional studies on open problems in LMMs using the model design and data mixture of LLaVA-1.5.

5.1. Data Efficiency

Despite the data efficiency of LLaVA-1.5 when compared with approaches like InstructBLIP [13], the training of LLaVA-1.5 still doubles when compared with LLaVA. In this section, we conduct experiments for further improving the data efficiency by randomly sub-sampling the training data mixture of LLaVA-1.5, with a sampling ratio ranging from 0.1 to 0.5. We visualize the relative performance of different sampling variants in Fig. 4.

First, the full data mixture provides the best knowledge coverage, and allows the model to achieve the best overall performance. To our surprise, with only 50% of the samples, the model still maintains more than 98% of the full dataset performance. This suggests that there is room for further improvements in data efficiency.

Second, when downsampling the dataset to 50%, the model's performance on MMBench, ScienceQA, and POPE does not decrease at all, and it even slightly improves on MMBench. Similarly, the model's performance remains steady when further downscaling the data from 50% to 30%. These results show promise of having the less-is-more [58] benefit for multimodal models as well.

5.2. Rethinking Hallucination in LMMs

Hallucination is an important issue to tackle for LLMs and LMMs. Often in LMMs, we attribute the model's hallucination to the errors or hallucinations in the training dataset. For example, the detailed descriptions in LLaVA-Instruct [34] may contain a small amount of hallucinated content, and it is believed that training on such data *may* have caused the model to hallucinate when asked to "describe the image in detail". However, we find that such hallucination is significantly reduced, when we scale the model's inputs to higher resolutions like 448^2.

This finding is interesting as it suggests that the LMMs may be robust to *a few* such errors in the training data. However, when the input resolution is not sufficient for the model to discern all details in the training data, and the amount of data that is at that granularity beyond the model's capability becomes large enough, the model *learns* to hallucinate. This further suggests that there needs to be a balance between improving the data annotation with more details and the model's capability to properly process the information at such granularities. An imbalanced scaling can cause the model to have an increased tendency to hallucinate or have a reduced understanding in visual details. We hope this finding provides a reference for future work in terms of dealing with hallucination and the scaling of the models and data.

5.3. Compositional Capabilities

We demonstrate interesting compositional capabilities in LLaVA-1.5: the model trained on a set of tasks independently generalizes to tasks that require a combination of these capabilities without explicit joint training. We note some of the findings below.

First, we observe an improved language capability in visual conversations after including the ShareGPT [44] data, including the multimodal multilingual capability as discussed in Sec. 4.3. Moreover, the model is more capable at providing longer and more detailed responses in visual conversations. Second, the additional visual knowledge from the academic-task-oriented datasets, improves the visual groundness of LLaVA-1.5's responses in visual conversations, as evidenced quantitatively by the improved results on MM-Vet [52] and LLaVA-Wild [34] in Table 4.

However, there is still difficulty in achieving ideal performance for some tasks that require a certain combination of capabilities. For example, being able to correctly answer the attribute of a certain object in VQA, does not guarantee an accurate depiction of that object attribute in a detailed description of the whole image. Furthermore, the capability of engaging in conversations with certain foreign languages (*e.g*. Korean) still falls behind. See appendix for examples.

These findings suggest that the compositional capabilities of LMMs can be leveraged to improve the model's performance without significantly increasing the data by exhaustively including all task combinations. Yet, it can be further investigated, and a deeper understanding of the mechanism behind the compositional capabilities of LMMs can further improve the capability and the data efficiency of LLaVA-1.5.

6. Conclusion

In this paper, we take a step towards demystifying the design of large multimodal models, and propose a simple, effective, and data-efficient baseline, LLaVA-1.5, for large multimodal models. In addition, we explore the open problems in visual instruction tuning, scale LMMs to higher resolutions, and present some intriguing findings in terms of model hallucination and compositional capabilities for LMMs. We hope these improved and easily-reproducible baselines as well as the new findings will provide a reference for future research in open-source LMM.

Limitations. Despite the promising results demonstrated by LLaVA-1.5, several limitations must be acknowledged. First, LLaVA-1.5 utilizes full image patches, potentially prolonging each training iteration. While visual resamplers [2, 13, 30] reduce the number of visual patches in LLMs, they currently cannot achieve convergence as efficiently as LLaVA with a comparable amount of training data, probably due to more trainable parameters in the resamplers. The development of a sample-efficient visual resampler could pave the way for future scaling-up of instruction-following multimodal models. Second, LLaVA-1.5 is not yet capable of processing multiple images due to the lack of such instruction-following data, and the limit of the context length. Third, although LLaVA-1.5 exhibits proficiency in following complex instructions, its problem-solving capabilities can still be limited in certain domains, which could be improved with a more capable language model and with high-quality, targeted visual instruction tuning data. Finally, despite its significantly reduced propensity for hallucination, LLaVA-1.5 is not exempt from producing hallucinations and occasionally disseminating misinformation, and should be used with caution in critical applications (*e.g*. medical).

Appendix

This appendix is organized as follows.

- In Section A, we show implementation details for LLaVA-1.5-HD (Sec. A.1), data and prompts (Sec. A.2), and hyperparameters (Sec. A.3).
- In Section B, we present more qualitative results for response format prompts (Sec. B.1), compositional capabilities (Sec. B.2).

A. Implementation Details

A.1. LLaVA-1.5-HD

A.1.1 Preprocessing

Overview. We use CLIP-ViT-L-14 (224^2) as the base image encoder. We first select and pad the input image to a target resolution that effectively captures its details, and split the image into 224^2 grids. All 224^2 image patches are encoded by the CLIP image encoder separately and their features are merged back to a single large feature map. We then post-process the resulting feature map to a flattened list of features. We additionally concatenate the features of a fixed-resolution image to provide the model with a global context.

Target resolution selection. We predefine a set of resolutions to support up to six grids (1x1, 1x2, 1x3, 1x4, 1x5, 1x6, 2x2, 2x3, and their transpose). This system allows for a maximum resolution of 672x448 (or 448x672). Two criteria are enforced in the target resolution selection: (1) *Detail preservation*: the selected resolution preserves as much detail from the original image as possible; (2) *Resource efficiency:* the resolution should not be excessively large to avoid unnecessary consumption of pixels and memory (*e.g.* it should not select 448^2 for a 224^2 input image).

Postprocessing. We perform three steps of postprocessing to ensure that the final features can be processed effectively and efficiently by the language model. (1) *Padding removal.* Features corresponding exclusively to the paddings are discarded. This reduces the number of visual tokens processed by the language model and improves the efficiency. (2) *Row-end Tokens.* We append a special token to the end of each row of features, to provide an explicit indication of the shape of the image. Unlike the original LLaVA and LLaVA-1.5 that uses a fixed resolution, we now use a variable resolution for the image features of LLaVA-1.5-HD, such indication allows the language model to capture the exact shape and the size of the image for each sample. (3) *Flattening.* Finally, we flatten the image feature map and feed it into the language model along with language token features.

Global context. We additionally pad and resize the image to a single image of 224^2, and concatenate it with the high resolution features to provide a global context. Ablation on a 7B model shows that the addition of the global context effectively boosts performance on all three validation benchmarks.

	GQA	MME	MM-Vet
high-res patch only	62.9	1425.8	31.9
+global context	63.8 (+0.9)	1497.5 (+71)	35.1 (+3.2)

A.1.2 Training

Since we compute the visual features on the original 224^2 resolution that the vision encoder is trained on, we do not perform additional pretraining. We also do not perform additional high resolution pretraining for the visual projectors, and perform visual instruction tuning directly on the higher-resolution images.

A.2. Data

Our final training data mixture contains a variety of datasets: VQA [18, 20, 39, 43], OCR [40, 45], region-level VQA [23, 24, 38], visual conversation [34] and language conversation [44] data. We adopt multiple strategies to reduce training cost and enhance efficiency, detailed as follows:

1. For all VQA datasets, QA pairs from the same training image are merged into a single conversation.
2. For ShareGPT [44], we filter out invalid conversations as [11]. Unlike Vicuna [11], long conversations that surpass 2048 tokens are truncated rather than splitting to multiple conversations. This results in $\sim$40K conversations.
3. Each QA pair in A-OKVQA [43] is augmented k times, where k is the number of choices per question, to counterbalance the lack of multiple-choice data.
4. 80K conversations are sampled from OCRVQA [40].
5. For Visual Genome, we sample 10 annotations for images with additional annotations.
6. For RefCOCO, conversations are dissected into segments, each containing fewer than 10 conversations.
7. We obverse that language conversations are often longer than visual ones. For each batch, we sample conversations only from a single modality, and this speeds up the training by 25%, and does not affect the final outcome.

All data splits are concatenated together and sampled with the same probability. We present the response formatting prompts of the final instruction-following data mixtures in Table 7 and the response format prompts used for each evaluation benchmark in Table 8.

A.3. Hyperparameters

The latest Vicuna v1.5 [57] is used as the base LLM. LLaVA-1.5 uses the same set of hyperparameters as the original LLaVA, except that we halve the learning rate in pretraining due to the usage of the MLP projection layer instead of

Data	Size	Response formatting prompts
LLaVA [34]	158K	–
ShareGPT [44]	40K	–
VQAv2 [18]	83K	Answer the question using a single word or phrase.
GQA [20]	72K	
OKVQA [39]	9K	
OCRVQA [40]	80K	
A-OKVQA [43]	66K	Answer with the option's letter from the given choices directly.
TextCaps [45]	22K	Provide a one-sentence caption for the provided image.
RefCOCO [23, 38]	48K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG [24]	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

Table 7. **Instruction-following Data Mixture** of LLaVA-1.5.

Data	Response formatting prompts
LLaVA-Bench, MM-Vet	–
VQAv2, GQA, TextVQA, MME, POPE	Answer the question using a single word or phrase.
ScienceQA, MMBench, SEED-Bench	Answer with the option's letter from the given choices directly.
VizWiz	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.

Table 8. **Response format prompt** for evaluation.

the original linear projection layer design. We show the training hyperparameters for both first-stage vision-language alignment pretraining and the second-stage visual instruction tuning in Table 9. We use greedy decoding for evaluation to ensure reproducibility.

Hyperparameter	Pretrain	Finetune
batch size	256	128
lr	1e-3	2e-5
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0	
epoch	1	
optimizer	AdamW	
DeepSpeed stage	2	3

Table 9. **Hyperparameters** of LLaVA-1.5 are the same as the original LLaVA, except that we halve the learning rate in pretraining due to the usage of the MLP projection layer.

B. Qualitative Results

B.1. Response Format Prompts

We show additional examples of LLaVA-1.5 generalizing to different unseen response format prompts.

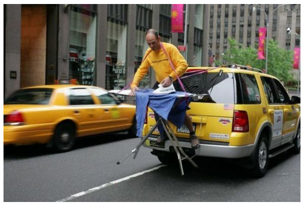
Visual input example, Different Format Prompts:	
	
User	What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**
LLaVA-1.5	Yellow
User	Is this unusual? **Please explain briefly with a short sentence.**
LLaVA-1.5	Yes, it is unusual to see a man ironing clothes while standing on the back of a moving car.
User	Is this unusual? **Please explain in detail.**
LLaVA-1.5	Yes, this is an unusual scene. A man is ironing clothes while standing on the back of a moving yellow SUV. This is not a typical activity one would expect to see on a city street, as ironing clothes is usually done indoors, in a more private setting. The man's actions are not only unconventional but also potentially dangerous, as standing on the back of a moving vehicle can be hazardous and may lead to accidents or injuries. Additionally, the presence of a suitcase and a handbag in the scene suggests that the man might be traveling or on the go, which further emphasizes the unusual nature of the situation.
User	What is unusual about this image? **Answer briefly with a single sentence.**
LLaVA-1.5	A man is ironing clothes in the back of a car.
User	**Give a more detailed explanation.**
LLaVA-1.5	The image shows a man ironing clothes in the back of a car, which is an unusual sight as ironing is typically done indoors. The man is standing on a folding table, and the car appears to be a taxi. The scene takes place on a city street, with other cars and a truck visible in the background.

Table 10. LLaVA-1.5 learns to format the response according to the user's request, generalizes to unseen format instructions, and can alter the response format within the conversation upon the user's request.

First, as shown in Table 10, LLaVA-1.5 can provide details at different granularities in response to user's requests. When requested by the user, it is also capable of switching between response formats within the conversations.

Second, we provide another example of the constrained prompting to generate the prompts for Stable Diffusion models. We show an example of generating anime prompts in Table 12.

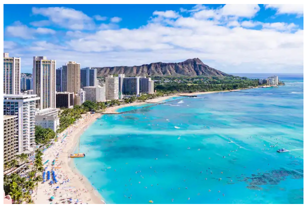
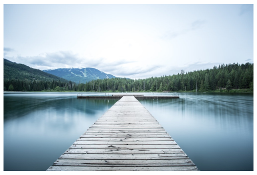
Visual input example, Visual Writing Task:	
	
User	Compose an engaging travel blog post about a recent trip to **this place**, highlighting **cultural experiences** and **must-see attractions**.
LLaVA	In my recent trip to Honolulu, I had the opportunity to experience the vibrant local culture and explore the beautiful beaches that this city has to offer. One of the must-see attractions was Waikiki Beach, where I witnessed a lively atmosphere with people sunbathing, swimming, and playing beach games. The crystal blue waters of the Pacific Ocean were a stunning sight, and the white sandy beach provided the perfect spot for relaxation. Aside from the beach, I also had the chance to immerse myself in the local culture by visiting the Pearl Harbor Memorial, which paid tribute to the events that led to the United States' involvement in World War II. The memorial was a somber reminder of the past, and it was fascinating to learn about the history surrounding the attack on Pearl Harbor. Other cultural experiences included trying local Hawaiian cuisine, such as poke bowls and traditional Hawaiian dishes, which were delicious and offered a unique taste of the region. I also had the opportunity to explore the city's bustling streets, where I came across local shops, markets, and street performers, adding to the lively atmosphere. Overall, my trip to Honolulu was an enriching experience, filled with cultural discoveries, breathtaking natural beauty, and unforgettable memories.
LLaVA-1.5	Title: A Sunny Escape to the Tropical Paradise of Waikiki Beach Introduction: I recently had the pleasure of visiting the beautiful Waikiki Beach in Hawaii, and I must say, it was an unforgettable experience. The crystal blue waters, the lively atmosphere, and the stunning cityscape in the background made it a perfect destination for a relaxing getaway. In this travel blog post, I'll share my experiences and must-see attractions during my trip to Waikiki Beach. Day 1: Arrival and Exploration Upon arrival, I was greeted by the warm and welcoming Hawaiian culture. The vibrant colors and lively music set the tone for my adventure. I decided to explore the area around Waikiki Beach, starting with the iconic Diamond Head Crater. The hike to the top was a bit challenging, but the breathtaking views of the beach and the city made it worthwhile. Day 2: Beach Day and Cultural Experiences I woke up early and headed to the beach to enjoy a day of sunbathing and swimming. The beach was bustling with activity, and I loved watching the surfers catching waves. In the evening, I attended a traditional Hawaiian luau, where I learned about the local customs and traditions, including hula dancing and the art of making leis. Day 3: City Tour and Sunset Cruise On the third day, I took a guided city tour to learn more about the history and culture of Hawaii. I visited the Pearl Harbor Memorial, the Bishop Museum, and the Iolani Palace, which was once the residence of the Hawaiian monarchs. In the evening, I went on a sunset cruise around the Waikiki coastline, where I enjoyed a delicious meal and witnessed the stunning sunset over the ocean. Day 4: Snorkeling and Relaxation For my last day in Waikiki, I decided to go snorkeling in the crystal clear waters of Hanauma Bay. The marine life was incredible, and I had the opportunity to see colorful fish and coral reefs up close. After a long day of exploring, I relaxed on the beach and soaked up the sun. Conclusion: My trip to Waikiki Beach was an unforgettable experience filled with adventure, culture, and relaxation. From the breathtaking views of the ocean to the vibrant cityscape, this tropical paradise has something for everyone. I highly recommend visiting Waikiki Beach for a sunny escape and a chance to immerse yourself in the beautiful Hawaiian culture.

Table 11. LLaVA-1.5 provides more detailed, visually-grounded responses for writing tasks with visual inputs than LLaVA.



Query: What are the things I should be cautious about when I visit here?

Reference (English): When visiting the pier over the lake, there are a few things you should be cautious about. First, ensure that you have appropriate footwear with good traction, as the pier may be wet or slippery, especially if it has been raining or if there is algae on the surface. Second, be mindful of the weather conditions, as the image shows a cloudy sky, which could indicate potential rain or storms. It is essential to stay safe and avoid being caught in a sudden storm or lightning. Lastly, be aware of your surroundings and any potential hazards, such as the presence of boats or other watercraft in the area. Always follow any posted signs or guidelines to ensure a safe and enjoyable experience.

Spanish: ¿Cuáles son las cosas de las que debería tener cuidado cuando visite aquí?

Cuando visite el lugar con la pier en el lago, debería tener cuidado de las siguientes cosas:

1. Mantenerse alejado del borde de la pier para evitar caer al agua o sufrir una lesión.
2. No subir a la pier si está prohibido o si no está diseñada para ello.
3. No arrojar objetos o desperdiciar en el agua, ya que esto puede dañar el ecosistema del lago.
4. No interferir con la fauna local, como los pájaros que se pueden encontrar en la zona.

Si decide nadar en el lago, asegúrese de conocer las condiciones del agua, como la profundidad y la presencia de algas, y de tener las habilidades y equipos adecuados para hacerlo de manera segura.

Tenir en cuenta estas precauciones ayudará a garantizar una experiencia segura y responsable en el lugar.

Translation: When visiting the place with the pier at the lake, you should be careful about the following things:

1. Stay away from the edge of the pier to avoid falling into the water or getting injured.
2. Do not climb onto the pier if it is prohibited or not designed for it.
3. Do not throw objects or litter into the water, as this can harm the lake's ecosystem.
4. Do not interfere with the local wildlife, such as birds that may be found in the area.

If you decide to swim in the lake, make sure to be aware of water conditions such as depth and the presence of algae, and have the appropriate skills and equipment to do so safely. Taking these precautions will help ensure a safe and responsible experience at the location.

Japanese: ここを訪れる際に注意すべきことは何ですか？

この場所は、寂しい湖に面した木製の桟橋です。訪れる際に注意すべきことは、桟橋の安全性と周囲の環境に対する配慮です。桟橋の構造が安定していることを確認し、桟橋が浸水していないことを確認することが重要です。また、湖の水深や周囲の地形に注意して、安全に桟橋を渡ることができるようにすることも大切です。さらに、湖の水質や周囲の生態系に配慮して、環境に負荷をかけないようにすることも重要です。

Translation: This place is a wooden pier facing a serene lake. When visiting, it's important to pay attention to the safety of the pier and respect the surrounding environment. Confirming the stability of the pier's structure and ensuring that it is not submerged in water are crucial. Additionally, being mindful of the lake's depth and the surrounding terrain is important to safely cross the pier. Furthermore, it's essential to consider the water quality of the lake and the local ecosystem to minimize any impact on the environment.

Korean: 여기를 방문할 때 주의해야 할 사항은 무엇인가요?

이 장소는 물체 위의 부두로, 맑고 푸른 물체에 있습니다. 방문자들은 이 부두에서 물체를 감상하고, 물체 주변의 자연 경치를 즐길 수 있습니다. 그러나 이 부두는 물체 위에 있기 때문에, 물체의 수준이 높아질 수 있으므로, 방문자들은 물체의 수준을 확인하고 적절한 예방 조치를 취해야 합니다. 또한, 부두에서 물체로 뛰어들지 않도록 주의해야 합니다. 이 부두는 매우 아름다운 자연 경치를 제공하므로, 방문자들은 이 장소를 즐기면서 안전한 방법으로 물체를 감상할 수 있도록 주의해야 합니다.

Translation: This place is located on a pier above clear and blue water. Visitors can admire the object from this pier and enjoy the natural scenery around the object. However, since the pier is above the object, the object's level can be high, so visitors should check the object's level and take appropriate precautions. Also, be careful not to jump from the pier into the object. This pier offers incredibly beautiful natural scenery, so visitors should enjoy this place while being mindful of safe ways to appreciate the object.

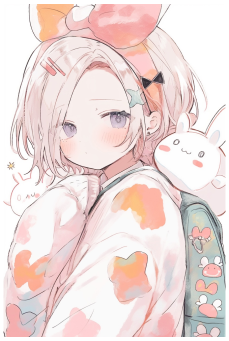
Figure 5. **Compositional capability: multilingual visual conversation.** LLaVA-1.5 generalizes to multilingual visual conversations, when training on visual instruction following data (English-only) together with the text-only ShareGPT data (multilingual). However, there can still be errors in some languages (*e.g.* Korean, errors marked in red).

B.2. Compositional Capabilities

We present qualitative examples of the compositional capabilities of LLaVA-1.5. As shown in Fig. 5, LLaVA-1.5 is capable of participating in multilingual visual conversations and adapting its output language based on the user's input, even though it has not been trained on multilingual visual instruction data. We hypothesize this emerging bahavior is a compositional capability learned from visual conversations (English-only) and the text-only ShareGPT data (multilingual). However, there can still be errors in some languages (*e.g*. Korean), which could be improved by incorporating more of those language data.

Additionally, in Table 11, we show another observed compositional capability after including the ShareGPT data in training. LLaVA-1.5 is able to produce more detailed and visually-grounded responses in writing tasks with visual inputs than LLaVA.

Visual input example, Constrained SD Prompt Generation:



User	Here's a formula for a Stable Diffusion image prompt: a cartoon-style image of [adjective] [subject] [haircut] [clothing] [doing action], [color scheme], detailed, trending on artstation. Generate the prompt for this image.
LLaVA-1.5	A cartoon-style image of a young girl with a bob haircut, wearing a white kimono, holding a stuffed animal, and posing for a picture, with a color scheme of pink and white, detailed, and trending on artstation.

Table 12. Constrained prompt generation for Stable Diffusion. Corresponding components are marked in color.

References

[1] 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. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2, 4, 5, 6, 8

[3] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use, 2023. 3

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 1

[5] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023. 1

[6] Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*, 2023. 2, 3

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2, 4, 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[10] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 5

[11] 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, 2023. 9

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1, 2, 3, 5, 6, 7, 8

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation

learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 6

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 4, 5

[17] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 1

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3, 5, 6, 9, 10

[19] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 5, 6

[20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 4, 5, 9, 10

[21] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. `https://huggingface.co/blog/idefics`, 2023. 5, 6

[22] 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. 2021. If you use this software, please cite it as below. 6

[23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 4, 9, 10

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4, 9, 10

[25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1

[26] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1, 5

[27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 1

[28] 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

[29] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 1

[30] 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, 3, 4, 5, 6, 8

[31] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[32] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1, 5

[33] 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. 2, 5

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 4, 5, 6, 8, 9, 10

[35] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 5, 7

[36] 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. 5

[37] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023. 1, 4

[38] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 4, 9, 10

[39] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 9, 10

[40] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 4, 9, 10

[41] OpenAI. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`, 2023. 1

[42] 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. 6

[43] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 4, 9, 10

[44] ShareGPT. `https://sharegpt.com/`, 2023. 4, 7, 8, 9, 10

[45] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 4, 9, 10

[46] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5

[47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 7

[48] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 1

[49] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2

[50] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 6

[51] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1

[52] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 1, 3, 4, 5, 8

[53] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 1

[54] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 2

[55] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 1, 2

[56] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023. 1

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 9

[58] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023. 2, 8

[59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2