Visual Instruction Tuning

Haotian Liu[1*], **Chunyuan Li**[2*], **Qingyang Wu**[3], **Yong Jae Lee**[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

Abstract

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has improved zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. In this paper, we present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLaVA: **L**arge **L**anguage **a**nd **V**ision **A**ssistant, an end-to-end trained large multimodal model that connects a vision encoder and LLM for general-purpose visual and language understanding. Our early experiments show that LLaVA demonstrates impressive multimodel chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model and code base publicly available.

1 Introduction

Humans interact with the world through many channels such as vision and language, as each individual channel has a unique advantage in representing and communicating certain world concepts, and thus facilitates a better understanding of the world. One of the core aspirations in artificial intelligence is to develop a general-purpose assistant that can effectively follow multi-modal vision-and-language instructions, aligned with human intent to complete various real-world tasks in the wild [4, 24].

To this end, the community has witnessed an emergent interest in developing language-augmented foundation vision models [24, 14], with strong capabilities in open-world visual understanding such as classification [36, 18, 53, 50, 35], detection [26, 58, 29], segmentation [23, 59, 54] and captioning [46, 25], as well as visual generation and editing [38, 39, 52, 13, 40, 27]. We refer readers to the *Computer Vision in the Wild* reading list for a more up-to-date literature compilation [11]. In this line of work, each task is solved independently by one single large vision model, with the task instruction implicitly considered in the model design. Further, language is only utilized to describe the image content. While this allows language to play an important role in mapping visual signals to language semantics—a common channel for human communication, it leads to models that usually have a fixed interface with limited interactivity and adaptability to the user's instructions.

Large language models (LLM), on the other hand, have shown that language can play a wider role: a universal interface for a general-purpose assistant, where various task instructions can be explicitly represented in language and guide the end-to-end trained neural assistant to switch to the task of interest to solve it. For example, the recent success of ChatGPT [31] and GPT-4 [32] have demonstrated the power of aligned LLMs in following human instructions, and have stimulated tremendous interest in developing open-source LLMs. Among them, LLaMA [44] is an open-source LLM that matches the performance of GPT-3. Alpaca [43], Vicuna [45], GPT-4-LLM [34]

[*]Equal contribution

Preprint. Work in progress

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

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

- *Multimodal instruction-following data*. One key challenge is the lack of vision-language instruction-following data. We present a data reformation perspective and pipeline to convert image-text pairs into the appropriate instruction-following format, using ChatGPT/GPT-4.
- *Large multimodal models*. We develop a large multimodal model (LMM), by connecting the open-set visual encoder of CLIP [36] with the language decoder LLaMA, and fine-tuning them end-to-end on our generated instructional vision-language data. Our empirical study validates the effectiveness of using generated data for LMM instruction-tuning, and suggests practical tips for building a general-purpose instruction-following visual agent. With GPT-4, we achieve state-of-the-art performance on the Science QA [30] multimodal reasoning dataset.
- *Open-source*. We release the following assets to the public: the generated multimodal instruction data, the codebase for data generation and model training, the model checkpoint, and a visual chat demo.

2 Related Work

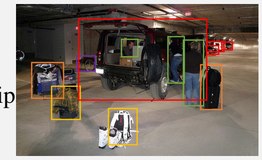
Multimodal Instruction-following Agents. In computer vision, existing works that build instruction-following agents can be broadly categorized into two classes: (i) End-to-end trained models, which are separately explored in each specific research topic. For example, the vision-language navigation task [3, 16] and Habitat [42] require the embodied AI agent to follow natural language instructions and take a sequence of actions to complete goals in visual environments. In the image editing domain, given an input image and a written instruction that tells the agent what to do, InstructPix2Pix [6] edits images by following the human instructions. (ii) A system that coordinates various models via LangChain [1] / LLMs [31], such as Visual ChatGPT [49], X-GPT [59], MM-REACT [51]. While sharing the same north star in building instruction-following agents, we focus on developing an end-to-end trained multimodal model for *multiple* tasks.

Instruction Tuning. In the natural language processing (NLP) community, to enable LLMs such as GPT-3 [7], T5 [37], PaLM [9], and OPT [56] to follow natural language instructions and complete real-world tasks, researchers have explored methods for LLM instruction-tuning [33, 48, 47], leading to instruction-tuned counterparts such as InstructGPT [33]/ChatGPT [31], FLAN-T5 [10], FLAN-PaLM [10], and OPT-IML [19], respectively. It turns out this simple approach can effectively improve the zero- and few-shot generalization abilities of LLMs. It is thus natural to borrow the idea from NLP to computer vision. Flamingo [2] can be viewed as the GPT-3 moment in the multimodal domain, due to its strong performance on zero-shot task transfer and in-context-learning. Other LMMs trained on image-text pairs include BLIP-2 [25], FROMAGe [22], and KOSMOS-1 [17]. PaLM-E [12] is an LMM for embodied AI. Based on the recent "best" open-source LLM LLaMA, OpenFlamingo [5] and LLaMA-Adapter [55] are open-source efforts that enable LLaMA to use image inputs, paving a way to build open-source multimodal LLMs. While promising task transfer generalization performance is presented, these models are not explicitly instruction-tuned with vision-language instruction data. In this paper, we aim to fill this gap and study its effectiveness. To clarify, visual instruction tuning is different from visual prompt tuning [20]: the former aims to improve the model's instruction-following abilities, while the latter aims to improve the parameter-efficiency in model adaptation.

3 GPT-assisted Visual Instruction Data Generation

The community has witnessed a surge in the amount of public multimodal data such as image-text pairs, ranging from CC [8] to LAION [41]. However, when it comes to multimodal instruction-following data, the available amount is limited, partially because the process is time-consuming and less well-defined when human crowd-scouring is considered. Inspired by the success of recent

Context type 1: Captions



A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

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

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description

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

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

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

GPT models in text-annotation tasks [15], we propose to leverage ChatGPT/GPT-4 for multimodal instruction-following data collection, based on the widely existing image-pair data.

For an image $\mathbf{X}_{\texttt{v}}$ and its associated caption $\mathbf{X}_{\texttt{c}}$, it is natural to create a set of questions $\mathbf{X}_{\texttt{q}}$ with the intent to instruct the assistant to describe the image content. We prompt GPT-4 and curate such a questions list in Table 8 in the Appendix. Therefore, a simple way to expand an image-text pair to its instruction-following version is `Human` : $\mathbf{X}_{\texttt{q}}$ $\mathbf{X}_{\texttt{v}}$`<STOP>\n` `Assistant` : $\mathbf{X}_{\texttt{c}}$`<STOP>\n`. Though cheap to construct, this simple expanded version lacks diversity and in-depth reasoning in both the instructions and responses.

To mitigate this issue, we leverage language-only GPT-4 or ChatGPT as the strong teacher (both accept only text as input), to create instruction-following data involving visual content. Specifically, in order to encode an image into its visual features to prompt a text-only GPT, we use two types of symbolic representations: (i) *Captions* typically describe the visual scene from various perspectives. (ii) *Bounding boxes* usually localize the objects in the scene, and each box encodes the object concept and its spatial location. One example is shown in the top block of Table 1.

This symbolic representation allows us to encode the image as an LLM-recognizable sequence. We use COCO images [28] and generate three types of instruction-following data. One example per type is shown in the bottom block of Table 1. For each type, we first manually design a few examples. They are the only human annotations we have during data collection, and are used as seed examples in in-context-learning to query GPT-4.

- *Conversation*. We design a conversation between the assistant and a person asking questions about this photo. The answers are in a tone as if the assistant is seeing the image and answering the question. A diverse set of questions are asked about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects. Only questions that have definite answers are considered. Please see Table 10 for the detailed prompt.
- *Detailed description*. To include a rich and comprehensive description for an image, we create a list of questions with such an intent. We prompt GPT-4 then curate the list, which is show in Table 9 in the Appendix. For each image, we randomly sample one question from the list to ask GPT-4 to generate the detailed description.
- *Complex reasoning*. The above two types focus on the visual content itself, based on which we further create in-depth reasoning questions. The answers typically require a step-by-step reasoning process by following rigorous logic.

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

4 Visual Instruction Tuning

4.1 Architecture

The primary goal is to effectively leverage the capabilities of both the pre-trained LLM and visual model. The network archtecture is illustrated in Figure 1. We choose LLaMA as our LLM $f_{\phi}(\cdot)$ parameterized by ϕ, as its effectiveness has been demonstrated in several open-source language-only instruction-tuning works. [43, 45, 34].

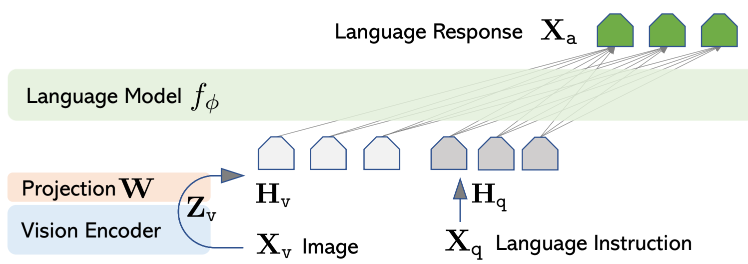


Figure 1: LLaVA network architecture.

For an input image $\mathbf{X}_{\mathrm{v}}$, we consider the pre-trained CLIP visual encoder ViT-L/14 [36], which provides the visual feature $\mathbf{Z}_{\mathrm{v}} = g(\mathbf{X}_{\mathrm{v}})$. The grid features before and after the last Transformer layer are considered in our experiments. We consider a simple linear layer to connect image features into the word embedding space. Specifically, we apply a trainable projection matrix $\mathbf{W}$ to convert $\mathbf{Z}_{\mathrm{v}}$ into language embedding tokens $\mathbf{H}_{\mathrm{q}}$, which have the same dimensionality of the word embedding space in the language model:

$$\mathbf{H}_{\mathrm{v}} = \mathbf{W} \cdot \mathbf{Z}_{\mathrm{v}}, \text{ with } \mathbf{Z}_{\mathrm{v}} = g(\mathbf{X}_{\mathrm{v}}) \tag{1}$$

Thus we have a sequence of visual tokens $\mathbf{H}_{\mathrm{v}}$. Note that our simple projection scheme is lightweight and cost-effective, which allows us to iterate data centric experiments quickly. More sophisticated (but expensive) schemes to connect the image and language representations can also be considered, such as gated cross-attention in Flamingo [2] and Q-former in BLIP-2 [25], or other vision encoders such as SAM [21] that provide object-level features. We leave exploring possibly more effective and sophisticated architecture designs for LLaVA as future work.

$\mathbf{X}_{\texttt{system-message}}$ <STOP> \n
Human : $\mathbf{X}^1_{\texttt{instruct}}$ <STOP> \n Assistant: $\mathbf{X}^1_{\texttt{a}}$ <STOP> \n
Human : $\mathbf{X}^2_{\texttt{instruct}}$ <STOP> \n Assistant: $\mathbf{X}^2_{\texttt{a}}$ <STOP> \n $\cdots$

Table 2: The input sequence used to train the model. Only two conversation turns are illustrated here; in practice, the number of turns varies based on the instruction-following data. In our current implementation, $\mathbf{X}_{\texttt{system-message}}$ = `A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.` and <STOP> = ###. The model is trained to predict the assistant answers and where to stop, and thus only green sequence/tokens are used to compute the loss in the auto-regressive model.

4.2 Training

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

$$\mathbf{X}^t_{\texttt{instruct}} = \begin{cases} \text{Random choose } [\mathbf{X}^1_{\texttt{q}}, \mathbf{X}_{\texttt{v}}] \text{ or } [\mathbf{X}_{\texttt{v}}, \mathbf{X}^1_{\texttt{q}}], & \text{the first turn } t = 1 \\ \mathbf{X}^t_{\texttt{q}}, & \text{the remaining turns } t > 1 \end{cases} \tag{2}$$

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

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

$$p(\mathbf{X}_{\texttt{a}}|\mathbf{X}_{\texttt{v}}, \mathbf{X}_{\texttt{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(x_i|\mathbf{X}_{\texttt{v}}, \mathbf{X}_{\texttt{instruct},<i}, \mathbf{X}_{\texttt{a},<i}), \tag{3}$$

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

Stage 1: Pre-training for Feature Alignment. To strike a balance between concept coverage and training efficiency, we filter CC3M to 595K image-text pairs. Please see the Appendix for details of the filtering process. These pairs are converted to the instruction-following data using the naive expansion method describe in Section 3. Each sample can be treated as a single-turn conversation. To construct the input $\mathbf{X}_{\texttt{instruct}}$ in (2), for an image $\mathbf{X}_{\texttt{v}}$, a question $\mathbf{X}_{\texttt{q}}$ in Table 8 is randomly sampled, which is a language instruction to request the assistant to describe the image briefly. The prediction answer $\mathbf{X}_{\texttt{a}}$ is the original caption. In training, we keep both the visual encoder and LLM weights frozen, and maximize the likelihood of (3) with trainable parameters $\boldsymbol{\theta} = \mathbf{W}$ (the projection matrix) only. In this way, the image features $\mathbf{H}_{\texttt{v}}$ can be aligned with the pre-trained LLM word embedding. This stage can be understood as training a compatible visual tokenizer for the frozen LLM.

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

- *Multimodal Chatbot*. We develop a Chatbot by fine-tuning on the 158K unique language-image instruction-following data collected in Section 3. Among the three types of responses, conversation is multi-turn while the other two are single-turn. They are uniformly sampled in training.
- *Science QA*. We study our method on the ScienceQA benchmark [30], the first large-scale multimodal science question dataset that annotates the answers with detailed lectures and

explanations. Each question is provided a context in the form of natural language or an image. The assistant provides the reasoning process in natural language and selects the answer from multiple choices. For training in (2), we organize the data as a single turn conversation, the question & context as $\mathbf{X}_{\texttt{instruct}}$, and reasoning & answer as $\mathbf{X}_{\texttt{a}}$.

5 Experiments

5.1 Multimodal Chatbot

We have developed a Chatbot demo to show image understanding and conversation abilities of LLaVA. To further study how well LLaVA is able to digest visual input and exhibits instruction-following capability. We first use the examples in the original GPT-4 paper [32], shown in Table 4 and Table 5. The prompt requires in-depth image understanding. For comparisons, we quote the prompt and response of the multimodal GPT-4 from their paper, and query BLIP-2 and OpenFlamingo model checkpoints to get their response.

Surprisingly, though LLaVA is trained with a small multimodal instruction-following dataset ($\sim$80K unique images), it demonstrates quite similar reasoning results with multimodal GPT-4 on these two examples. Note that both images are out-of-domain for LLaVA, LLaVA is able to understand the scenes and follow the question instruction to respond. In contrast, BLIP-2 and OpenFlamingo focus on describing the image, instead of following the user instruction to answer in an appropriate manner. More examples are shown in Figure 3, Figure 4 and Figure 5. We recommend readers to interact with LLaVA to study its performance.

Quantitative Evaluation. To have a systematic understanding of the performance of LLaVA, we aim to leverage a quantitative metric in measuring the model's instruction-following capability. Motivated by [45], we leverage GPT-4 to measure the quality of our model's generated responses. Specifically, we randomly select 30 images from the COCO validation split, and generate three types of question (conversation, detailed description, complex reasoning) using the proposed data generation pipeline. LLaVA predicts the answers based on the question and the visual input image. GPT-4 makes a reference prediction based on the question, and the ground-truth bounding boxes and captions, marking an upper bound of the teacher model. After obtaining the response from both models, we feed the question, visual information (in the format of captions and bounding boxes), and the generated responses from both assistants, to the GPT-4. GPT-4 evaluates the helpfulness, relevance, accuracy, and level of details of the responses from the assistants, and give an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. GPT-4 is also asked to provide a comprehensive explanation the evaluation, for us to better understand the models.

We vary the training datasets to study the effectiveness of different types of the instruction-following data. We show the results in Table 3. First, with instruction tuning, the model's capability of following the user instructions improves significantly by over 50 points. Second, adding a small amount of the detailed description and complex reasoning questions contributes to a considerable improvement of the model's overall capability by 7 points. Furthermore, it also improves the model's performance on conversational questions, suggesting that the overall improvement in the reasoning capability is

	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)

Table 3: Relative scores for different settings *w.r.t.* GPT-4 (text-only) on 30 randomly sampled images from COCO Val 2014. Each image is associated one short question, one detailed question, and one complex reasoning question, resulting in a total of 90 questions. Following the same setting as our data generation pipeline, GPT-4 uses ground truth image captions and bounding boxes as visual input. We prompt GPT-4 with the answers from our model outputs and the answers by GPT-4 (text-only), and let it compare between both responses and give a rating with an explanation.

Visual input example, Extreme Ironing:	
	
	Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [32]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	Can you explain this meme in detail?
BLIP-2	a man is sitting on the back of a yellow cab
User	Can you explain this meme in detail?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Table 4: Example prompt demonstrating LLaVA and GPT-4's visual input capability. The prompt requires image understanding.

complementary to its conversational capabilities. Finally, we show that having all three types of the data yields the best performance of 85.1%. We hope this evaluation protocol serves as a starting point for a comprehensive evaluation and understanding of the capability of large multimodal models.

5.2 ScienceQA

ScienceQA [30] contains 21k multimodal multiple choice questions with rich domain diversity across 3 subjects, 26 topics, 127 categories, and 379 skills. The benchmark dataset is split into training, validation, and test splits with 12726, 4241, and 4241 examples, respectively. We consider two representative methods, including GPT-3.5 model (`text-davinci-002`) with and without chain-of-thoughts (CoT), LLaMA-Adapter [55], as well as multimodal chain-of-thoughts (MM-CoT) [57], which is the current SoTA method on this dataset. For more baseline numbers, please see [30].

The results are reported in Table 6. For LLaVA, we use the visual features before the last layer, ask the model to first predict reasons then the answer, and train it for 12 epochs. It yields 90.92% accuracy, which is quite close to the SoTA 91.68%. To explore the limit of LLMs, we also prompt GPT-4 using 2-shot in-context-learning and achieve 82.69% accuracy, which is a 7.52% absolute gain compared with 75.17% from GPT-3.5. For a substantial number of questions, we note that GPT-4 fails simply because it reports that there is insufficient context such as images or plots. We consider two schemes to combine the outcomes from our model and GPT-4. (i) *A GPT-4 complement*.

Visual input example, Chicken Nugget Map:	
	
User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.
User	Can you explain this meme?
GPT-4 [32]	This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
User	Can you explain this meme in detail?
BLIP-2	sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User	Can you explain this meme in detail?
OpenFlamingo	It's a picture of a chicken nugget on the International Space Station.

Table 5: Example prompt demonstrating LLaVA and GPT-4's visual input capability. The prompt requires image understanding.

Whenever GPT-4 fails to provide answers, we use the prediction from our method. This schemes yields 90.97% accuracy, which is almost the same as applying our method alone. (ii) *GPT-4 as the judge*. Whenever GPT-4 and LLaVA produce different answers, we prompt GPT-4 again, asking it to provide its own final answer based on the question and two outcomes. The spirit is similar with CoT, but with the external knowledge from the other model. Surprisingly, GPT-4 is able to provide consistent improvement over all question classes, and achieves a new SoTA accuracy of 92.53%. To the best of our knowledge, this is the first time that GPT-4 is used for model ensembling. We hope this finding can encourage future research to explore more effective methods to leverage LLMs for model ensembling.

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)	-

Table 7: Design choice ablations (%). The difference with the best variant is reported in red text.

Ablations. We ablate several design choices on ScienceQA in Table 7. (i) *Visual features*. We tried using the last layer feature from CLIP vision encoder, which yields 89.96% and is 0.96% lower than the feature before the last year. We hypothesize that this is because CLIP's last year features may focus more on global image properties compared to the layer before it, which can focus more on localized properties that can be more useful for understanding specific image details. (ii) *Chain-of-thoughts*. To decide the order between the answer and reasoning process in the model prediction, we run both variants and observe that answer-first reports the best number 89.77%

Method	Subject			Context Modality			Grade		Average
	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Representative & SoTA methods with numbers reported in the literature									
Human [30]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [30]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [30]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [55]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [57]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [57]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
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**

Table 6: Results (accuracy %) on Science QA dataset. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12.

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

6 Discussions

This paper demonstrates the effectiveness of visual instruction tuning using language-only GPT-4. We have presented an automatic pipeline to create language-image instruction-following data, based on which we train LLaVA, a multimodal model to follow human intent to complete visual tasks. It achieves the new SoTA accuracy when fine-tuned on ScienceQA, and excellent visual chat experience when fine-tuned on multimodal chat data.

This project is a work in progress, and several directions can be explored: (i) *Data scale*. The pre-training data is limited to a subset of CC3M, and the fine-tuning data is a subset of COCO. We think it will be worthwhile to pre-train on larger image-text data to increase concept coverage (*e.g.,* entities and OCR). It will also be promising to apply the data generation pipeline to a larger set of language-image grounding data (e.g., as used in GLIP [26] and GLGEN [27]) to generate more instruction-following data to fine-tune the multimodal chat assistant. (ii) *Connecting with more visual models*. Our promising results indicate near multimodal GPT-4 performance in some cases. Besides trying to match its performance through data/model scaling, it may be more interesting for academia to connect other powerful vision models such as SAM [21] into LLaVA, to enable new capabilities that multimodal GPT-4 currently might not be equipped with.

Acknowledgements. We thank Baolin Peng and Pan Lu for valuable discussions on instruction-tuning language models and Science QA, respectively. We thank the LLaMA team for giving us access to their models. This work was supported in part by NSF CAREER IIS2150012, and Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training).

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] 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] 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

[10] 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

[11] CVinW. Computer vision in the wild. `https://github.com/Computer-Vision-in-the-Wild/CVinW_Readings`, 2022. 1

[12] 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

[13] 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

[14] 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

[15] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023. 3

[16] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, 2020. 2

[17] 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

[18] 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

[19] 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-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 2

[20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4, 9

[22] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023. 2

[23] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ICLR*, 2022. 1

[24] 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

[25] 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

[26] 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, 9

[27] 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, 9

[28] 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. 4

[29] 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

[30] 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, 7, 9

[31] OpenAI. ChatGPT. `https://openai.com/blog/chatgpt/`, 2023. 1, 2

[32] OpenAI. Gpt-4 technical report, 2023. 1, 6, 7, 8

[33] 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

[34] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*, 2023. 1, 4

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] Vicuna. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org/, 2023. 1, 4, 6

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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, 7, 9

[56] 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

[57] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 7, 9

[58] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei 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

[59] 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 Data

Instructions for brief image description. The list of instructions used to briefly describe the image content are shown in Table 8. They present the same meaning with natural language variance.

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo's key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

Table 8: The list of instructions for brief image description.

Instructions for detailed image description. The list of instructions used to describe the image content in detail are shown in Table 9. They present the same meaning with natural language variance.

- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"
- "Characterize the image using a well-detailed description"
- "Break down the elements of the image in a detailed manner"
- "Walk through the important details of the image"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"

Table 9: The list of instructions for detailed image description.

CC3M. We extract noun-phrases using Spacy for each caption over the whole cc3m dataset, and count the frequency of each unique noun-phrase. We skip noun-phrases whose frequency is smaller than 3, as they are usually rare combinations concept and attributes that has already been covered by other captions. We then start from the noun-phrases with lowest remaining frequency, add the captions that contain this noun-phrase to the candidate pool. If the frequency of the noun-phrase is larger than 100, we randomly choose a subset of size 100 out of all its captions. This results in around 595K image-text pairs.

The comparison of noun-phrase statistics before and after filtering CC3M is shown in Figure 2. The filtered dataset shows a good coverage of concepts whose frequency is higher from 3, but with a smaller number of image-text pairs.

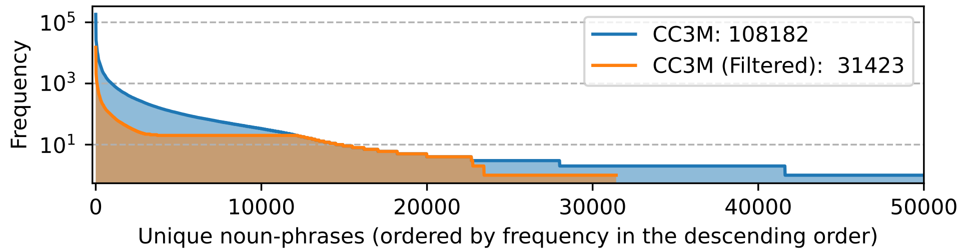


Figure 2: Comparison of noun-phrase statistics before and after filtering CC3M. The total number of unique noun-phrases are reported in the legend.

B Prompts

The prompt used to generate image-based conversation from ChatGPT/GPT-4 is shown in Table 10.

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
]
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)})
```

Table 10: For each query, we illustrate the prompt construction process for ChatGPT/GPT-4 to collect `query['response']` from `query['context']`, using few-shot in-context-learning, where examples are from `fewshot_samples`, each example including input `sample['context']` and output `sample['response']`. Note that `messages` is the final prompt. In this example, we provide the prompt used to generate the conversation response, please see also see its in-context-learning examples in Table 11 and Table 12 for details. We recommend readers to check out the codebase for the prompts to generated two other types of responses, including detailed decription and complex reasoning.

Captions
There is a movie theater that displays the show times above the doors.
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

Conversation
Question:
What color is the fire hydrant in the image?
===
Answer:
The fire hydrant in the image is red.
===
Question:
Is the fire hydrant fully covered by snow or partially visible?
===
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
===
Question:
What is the condition of the sidewalk near the fire hydrant?
===
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
===
Question:
What does the image suggest about the winter season in the city?
===
Answer:
The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents.

For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.

Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes. Additionally, the heavy snow can put extra strain on infrastructure, such as roofs and power lines, increasing the likelihood of structural damage or power outages.

In conclusion, the image of the red fire hydrant deep in the snow and the recently plowed sidewalk suggest that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure.

Table 11: One example used in in-context-learning to construct visual conversation data.

Captions
A man is skiing in the open snow covered hills
A skier is making his way into the snow.
A skier on the foothills of a huge mountain range.
A skier looks at mountains while standing near a trail sign.
a single person skiing by an area with a lot of bushes

Conversation
Question:
What is the skier doing?
===
Answer:
The skier is skiing in the snow-covered hills, making his way through the snow and enjoying the beautiful mountain scenery. They are also standing near a trail sign, which indicates their location on the mountain.
===
Question:
Are there any other people or skiers visible in the image?
===
Answer:
No, there is only a single person skiing in the image, and they are surrounded by an area with a lot of bushes.
===
Question:
What kind of skiing is the person doing? Cross-country or downhill?
===
Answer:
Based on what I can see in the photo, it is not possible to definitively determine whether the skier is doing cross-country or downhill skiing. However, given the presence of a trail sign and the fact that they are skiing in the foothills, it is more likely that the skier is engaged in cross-country skiing. Downhill skiing typically takes place on more steep slopes and involves ski lifts, whereas cross-country skiing often occurs on relatively flatter terrain with marked trails.
===
Question:
Can you explain cross-country and downhill skiing a bit more?
===
Answer:
Cross-country skiing and downhill skiing are two distinct types of skiing, each with its own techniques, equipment, and terrain preferences.

1. Cross-Country Skiing:
Also known as Nordic skiing, cross-country skiing involves traversing relatively flat terrain or gentle slopes using skis and poles. It is a popular recreational and competitive sport that focuses on endurance, balance, and aerobic fitness. Cross-country skiers use their own locomotion to move across snow-covered terrain, rather than relying on ski lifts or other forms of assistance.
There are two main techniques in cross-country skiing:
- Classic: This technique involves skiing in parallel tracks, using a forward-and-backward gliding motion while keeping the skis parallel to each other. The skier uses their poles to propel themselves forward.
- Skate skiing: This technique is similar to ice skating, where the skier pushes off one ski at an angle and glides on the other. It requires a smooth, groomed surface and is typically faster than the classic technique.
Cross-country ski equipment consists of lightweight skis, boots, bindings, and poles. The boots are more flexible than those used in downhill skiing, allowing for better ankle movement and control.

2. Downhill Skiing:
Also known as alpine skiing, downhill skiing involves descending slopes at high speeds using skis and poles for balance and control. This sport is more focused on speed, technique, and navigating challenging terrain, including steep slopes, moguls, and even jumps.
Downhill skiing can be further categorized into several disciplines, such as slalom, giant slalom, super-G, and downhill racing. Each discipline has its own set of rules, courses, and ski equipment.
Downhill ski equipment includes heavier and stiffer skis, boots, bindings, and poles than those used in cross-country skiing. The boots are more rigid to provide better support and control during high-speed descents and sharp turns.

In summary, cross-country skiing is an endurance-based sport that involves traveling across flat or gently sloping terrain, while downhill skiing is focused on speed and technique as skiers navigate steeper slopes and challenging terrain. Both sports require specialized equipment and techniques, but they offer different experiences and challenges to participants.

Table 12: One example used in in-context-learning to construct visual conversation data.

What are the meals that I can cook with these?

With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, you can make a delicious smoothie using the fruits and vegetables, which would be a healthy and nutritious snack or breakfast option. The milk and eggs can be used to create various dishes, such as a milk-based smoothie, a baked casserole, or scrambled eggs. The juice can also be used to make a fresh fruit and juice mix or as a refreshing homemade popsicle. The possibilities are vast, and the meals you can create will depend on your preferences and dietary needs.

I would like to do the fruit salad. Show me the recipe.

Certainly! Here's a simple fruit salad recipe using the items in your refrigerator:

Ingredients:
- 2 cups fresh strawberries, sliced
- 1 cup fresh blueberries
- 1 large carrot, grated
- Juice of half a lemon
- A handful of chopped fresh parsley or mint
- A few ice cubes
- Optional: honey or maple syrup (adjust to taste)

Instructions:
1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix.
2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor.
3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more.
4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences.
5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad.
6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.

This fruit salad recipe is perfect for a healthy snack or side dish, and with the variety of ingredients available in your refrigerator, it should be easy to put together quickly.

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

When visiting this peaceful location with a pier extending over a serene lake, surrounded by mountains, there are several things to be cautious about. First, pay attention to the weather conditions, as sudden changes in weather, especially strong winds or storms, can cause the water to become rough and pose risks to the stability of the pier. Second, be mindful of the structural integrity of the pier; avoid overloading it with too many people or excessive weight, as this could potentially damage the pier or cause it to collapse. Additionally, be aware of the surrounding wildlife and respect their space; do not harm or disturb them in any way. Finally, if you decide to go fishing, ensure that you adhere to any local fishing regulations and use appropriate fishing techniques to keep the lake ecosystem healthy and maintain the balance of the food chain.

Compose an engaging travel blog post about my trip to this place, highlighting the beautiful sceneries, joyful experiences, and include possible anecdotes.

Traveling to this picturesque location with a serene lake and mountains in the background was a refreshing and rejuvenating experience. The pier extending over the calm water provided an excellent vantage point from which to appreciate the breathtaking scenery. As I stood on the pier, I couldn't help but feel a deep connection with nature, taking a moment to soak in the peaceful atmosphere.

One of the highlights of my trip was undoubtedly the fishing experience. The area is teeming with life, and I was thrilled to land a sizeable catch, despite the modest size of the lake. The joy of fishing in such a beautiful setting was truly rewarding, and it allowed me to take a break from my busy life and truly immerse myself in the tranquility of nature.

While exploring the surrounding area, I came across a family of birds that had built a nest close to the pier, adding to the charm of the location. Interacting with the wildlife and observing their behavior was an enriching experience that I'll cherish for years to come.

In conclusion, this idyllic destination provided me with an opportunity to unwind and recharge, surrounded by mesmerizing natural beauty and the gentle sounds of water and wildlife. If you're looking for a serene escape from the hustle and bustle of everyday life, this pier overlooking the lake is the perfect place to find solace and reconnect with nature.

Figure 3: LLaVA is capable of recognizing the visual content following the user's intent, without directly prompting for visual recognition. It also provides a detailed response when prompted with a follow-up request, and the generated response is closely related to the provided visual content.

What's the ending of this movie?

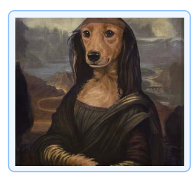
In the image, a young woman is leaning on a man in front of a beautiful sunset, with both of them standing on a boat. They appear to be enjoying a romantic moment together. The scene suggests that they could be depicted as the characters Jack and Rose from the popular movie Titanic. In the movie, their romance blossoms aboard the fictional ocean liner Titanic, with the pair ultimately sharing a passionate kiss as the ship sinks into the ocean. The ending of the movie is known for its emotional and heartfelt scenes, where the characters are forced to say goodbye to each other as they are separated by the sinking ship. The image evokes a sense of romance and nostalgia, paying tribute to the iconic love story portrayed in the movie Titanic.

Figure 4: LLaVA relates the movie scenes to the textural knowledge from the pretrained LLM.

Do you know who drew this painting?

The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

Start a new conversation, and the history is cleared.



Do you know who drew this painting?

The painting depicts a dog in a humorous situation, where it appears to be wearing a woman's dress and standing in front of a landscape. Although the artist of the painting is not explicitly identified, it could be a reference to the famous artist Da Vinci, known for his artistic works, including the Mona Lisa. The painting is likely a creative and comical take on the traditional portrait style, replacing a human figure with a dog dressed in a woman's clothing, making it an interesting and amusing visual piece.

Figure 5: LLaVA recognizes the famous art work, Mona Lisa, by Leonardo da Vinci. When we start a new conversation, it also explains the humourous artwork created on the web, mimicking the Mona Lisa.