Visual Instruction Tuning

Instruction Tuning

LLMs trained without supervision are not user-aligned, i.e. are not capable to follow the natural language instructions.

Prompt	Explain the moon landing to a 6 year old in a few sentences.
GPT-3	Explain the theory of gravity to a 6 year old. Explain the theory of relativity to a 6 year old in a few sentences. Explain the big bang theory to a 6 year old. Explain evolution ot a 6 year old.

$$\Downarrow$$

Finetune LLMs with collection of a {instruction, output} pairs.

$$\Downarrow$$

Improve zero and few shot generalization!

Motivation

Why not expand the instruction tuning to Large Multimodal Models?

$\Downarrow$

Use *visual instruction tuning*!

GPT-assisted Visual Instruction Data Generation

Given an image $\mathbf{X}_v$ and its corresponding caption $\mathbf{X}_c$, natural way to create a visual instruction data is to create a set of questions $\mathbf{X}_q$ that asks to describe the given image.

Ex. $\mathbf{X}_q$ generated by GPT-4

Brief	Describe the image concisely.
	Provide a brief description of the given image.
Detailed	Describe the following image in detail.
	Provide a detailed description of the given image.

This approach is cheap but lacks diversity and in-depth reasoning.

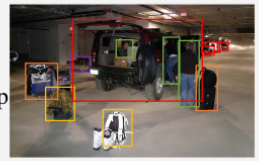
GPT-assisted Visual Instruction Data Generation

To mitigate the previous issue, the authors use GPT-4 (which only takes text as input) as a strong teacher. First authors create two types of context, or symbolic representation of images to encode visual as an LLM-recognizable sequence. Then using context and few human designed examples, authors prompt GPT-4 to create 3 types of instruction-following data through in-context learning.

1. Context: symbolic representation of image
   - Captions
   - Bounding boxes
2. Instruction-following data
   - Conversation
   - Detailed description
   - Complex reasoning

GPT-assisted Visual Instruction Data Generation

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.

Figure 1: Example of instruction-following data

Architecture

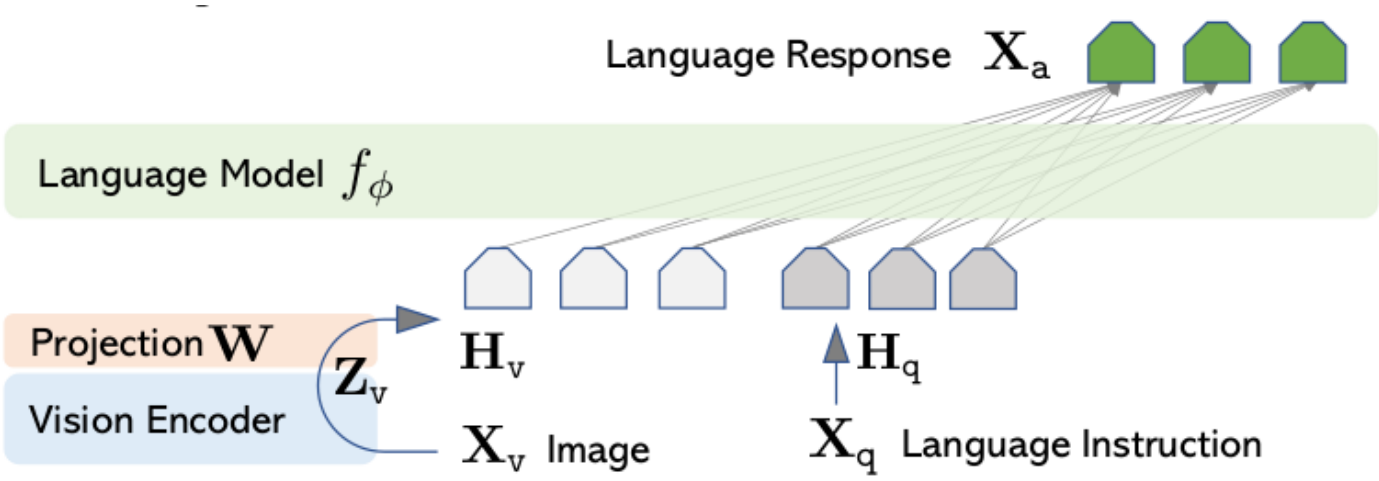


Figure 2: LLaVA network architecture

- Language model f_ϕ: Vicuna
- Vision encoder g: CLIP ViT-L/14
- Projection W: linear layer

Training

Given an image $\mathbf{X}_v$, generate multi-turn conversation data $(\mathbf{X}_q^1, \mathbf{X}_a^1, \cdots, \mathbf{X}_q^T, \mathbf{X}_a^T)$, create a sequence of instructions $\mathbf{X}_{\text{instruct}}^t$ as

$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1] & t = 1 \\ \mathbf{X}_q^t & t > 1 \end{cases}$$

and train the model with

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

Training

Stage 1 **Pretraing for feature alignment.** Freeze both visual encoder and LLM weights, and train just W.

Stage 2 **Fine-tuning end-to-end.** Freeze only visual encoder weights and update both W and ϕ.

- Multimodal Chatbot
- Science QA

Multimodal Chatbot

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 [35]	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	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Figure 3: Multimodal chatbot example

LLaVA-Bench

1. Create {image, textual description, question} triplet
2. Feed {image, question} to multimodal model
3. Feed {textual description, question} to GPT-4
4. Feed textual description and response to a *judge* GPT-4 and ask to evaluate the score from 1 to 10 with a comprehensive explanation of such evaluation
5. Report the *relative* score w.r.t to text-only GPT-4

LLaVA-Bench

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [27]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

Figure 4: LLaVA-Bench (In-the-Wild)

Demo

LLaVA Demo

Thank You

Q & A