NeurIPS 2023 · Oral

arXiv:2304.08485

LLaVA: Large Language *and* Vision Assistant

Visual Instruction Tuning

Haotian Liu · Chunyuan Li · Qingyang Wu · Yong Jae Lee

UW–Madison · Microsoft Research · Columbia University

Instruction tuning transformed language models — but the multimodal world was untouched.

Instruction tuning

Fine-tuning a pretrained model on examples phrased as *instruction → response*, so it learns to follow open-ended human commands rather than merely continue text — and generalizes **zero-shot** to tasks it never saw during training.

For text-only LLMs this recipe was well established — machine-generated instruction data made them broadly helpful.

For **images** it was barely explored. No large, diverse dataset existed to teach a model to *see and converse* at once.

Four contributions, one open recipe.

i **Multimodal instruct data**

The first use of language-only GPT-4 to generate image–text instruction-following data at scale.

ii **The LLaVA model**

An end-to-end multimodal model connecting a vision encoder to an LLM for general visual chat.

iii **Strong performance**

GPT-4-level chat behaviors and a new state of the art on Science QA.

iv **Fully open source**

Data, model weights, and code released for the research community.

GPT-4 writes the conversations; COCO images ground them.

CONTEXT — WHAT GPT-4 IS GIVEN

CAPTIONS

A group of people standing outside a black SUV with various luggage.

People try to fit all of their luggage in an SUV parked in a public garage.

BOUNDING BOXES

```
person [0.68, 0.24, 0.77, 0.69]
backpack [0.38, 0.70, 0.49, 0.91]
suitcase [0.05, 0.66, 0.27, 0.81]
```

LANGUAGE-ONLY GPT-4 →

GENERATED — AN INSTRUCTION–RESPONSE PAIR

Q What are the people in the image doing?

A They are trying to fit all of their luggage into the SUV, likely preparing for a trip.

GPT-4 sees only these **symbolic descriptions** — never the pixels. That keeps generation cheap and lets the pipeline scale to **158K** samples across COCO images.

The 158K samples span three response styles.

Conversation

58K samples

Q What type of vehicle is featured? A A black sport utility vehicle (SUV), parked in an underground garage.

Detailed Description

23K samples

The image shows an underground parking area with a black SUV. Three people work together to pack their luggage for a trip, suitcases and backpacks surrounding the vehicle.

Complex Reasoning

77K samples

Q What challenges do these people face? A Fitting all their luggage into the SUV — with many suitcases and backpacks, they must use the limited space efficiently.

A frozen vision encoder, a language model, and one projection matrix between them.

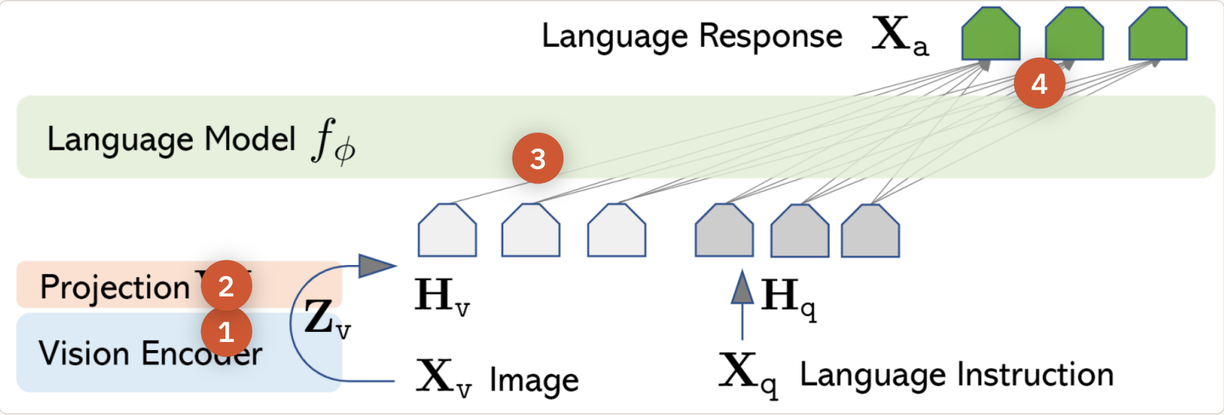

Image X_v and instruction X_q in → response X_a out

1 **Vision encoder** `~0.3B · frozen`

CLIP ViT-L/14 encodes the image into visual features Z_v. Pretrained and kept **frozen**.

2 **Projection** W `~5M · trained`

A single linear layer maps those features into the LLM's word-embedding space as visual tokens H_v — the only weights trained in stage 1.

3 **Language model** $f\phi$ `7B / 13B · Vicuna`

Vicuna reads the visual tokens alongside the tokenized instruction H_q.

4 **Response** X_a

The answer is generated **autoregressively**, one token at a time.

Align the features first, then fine-tune end-to-end.

1 Pre-training for feature alignment

CC3M · 595K image–text pairs

The projection maps CLIP's visual tokens into the LLM's word-embedding space.

Trains · projection W | Frozen · vision encoder + LLM

INPUT	→	TARGET OUTPUT
Image + a brief “describe the image” instruction	→	The image's original caption

2 Fine-tuning end-to-end

158K LLaVA-Instruct · or Science QA

Visual tokens H_v join the instruction H_q in one sequence; the LLM is now trained to respond to the image.

Trains · projection W + LLM | Frozen · vision encoder

INPUT	→	TARGET OUTPUT
Image + the user's multi-turn instructions	→	Assistant replies — next-token loss on the **answer tokens only**

Decoupling alignment from instruction-following keeps training stable — stage 2 runs in **~24 GPU-hours on one 8×A100 node**.

85.1%

LLaVA's answers score **85.1% as high** as a strong text-only GPT-4 reference — graded by GPT-4 itself.

1. **90 prompts.** 30 unseen COCO images × three question types — conversation, detailed description, complex reasoning.
2. **Two answers each.** LLaVA replies from the image; the reference is a text-only GPT-4 given that image's captions and boxes.
3. **GPT-4 grades 1–10.** The relative score is LLaVA's total ÷ the reference's total — **85.1%** overall.

LLaVA alone is competitive; paired with GPT-4 it sets the record.

Science QA is a multimodal benchmark of multiple-choice science questions paired with images, diagrams and text context. Accuracy is the share answered correctly.

90.92%
LLAVA, FINE-TUNED

+ GPT-4 as judge →

92.53%
NEW SOTA

When the two disagree, a text-only GPT-4 re-reads both its own answer and LLaVA's and picks the final one. This **"GPT-4 as judge"** ensemble reaches 92.53% — past every prior method at the time.

On GPT-4's own showcase images, LLaVA reads the scene and the joke.

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

Extreme ironing — unusual-scene reasoning · excerpt

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

Chicken-nugget map — reading visual humor · excerpt

A small, open recipe for teaching models to see and talk.

Language models can label vision

GPT-4 generated useful multimodal instruction data from symbols alone — no image input required.

Open beats closed for research

Released data, weights and code made LLaVA a foundation for a wave of follow-up work.

A linear bridge is enough

One projection matrix connects frozen CLIP features to an LLM and produces fluent visual chat.

Cheap to reproduce

LLaVA-1.5 reaches SoTA on 11 benchmarks training in ~1 day on a single 8×A100 node.

Everything is public.

PAPER	Visual Instruction Tuning	arXiv:2304.08485 · NeurIPS 2023
CODE	End-to-end training & inference	github.com/haotian-liu/LLaVA
DEMO	Live visual chat	llava.hliu.cc
IMPROVED	LLaVA-1.5 baselines	arXiv:2310.03744
DATA	LLaVA-Instruct-150K	huggingface.co/datasets/liuhaotian
PROJECT	Overview & figures	llava-vl.github.io