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

HUMANE

박현빈

2025_01_03

Outline

- proposing Instruction Tuning using visual and textual information to improve instruction following
- generate multimodal language-image instruction-following data
- introducing LLaVA
- construct evaluation benchmarks

Instruction Tuning

- LLM instruction-tuning can effectively improve the zero- and few-shot abilities of LLMs
- MLLM is not tuned with vision-language instruction data

GPT-assisted Visual Instruction Data Generation

- How to generation?

COCO images Context

(Captions, Bounding Boxes)

provide image context instead of image

→

Text-only GPT-4

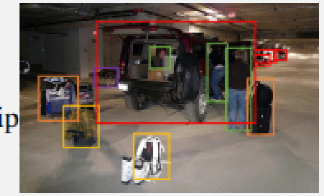
→

Responses

GPT-assisted Visual Instruction Data Generation

- Data Composition

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], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

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.

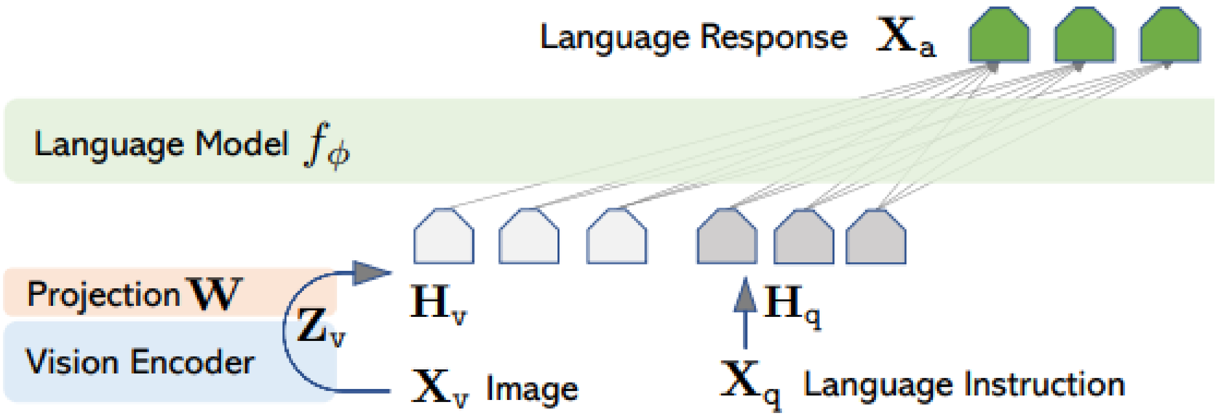
Total : 158k language-image pairs (58k conversations + 23k detailed description + 77k complex reasoning)

Visual Instruction Tuning

- Architecture

Decoder-only

- Vision Encoder : CLIP ViT-L/14
- LLM : Vicuna 13B



Visual Instruction Tuning

- Training

  - For each image X_V, we generate multi-turn data $(X_q^1, X_a^1, \ldots, X_q^T, X_a^T)$, where T is the total number of turns
  - Instruction and question can be considered as same

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

  - The input sequence used to train the model

$$\begin{aligned} &\mathbf{X}_{\texttt{system-message}} \ \texttt{<STOP>} \\ &\texttt{Human}: \mathbf{X}_{\texttt{instruct}}^1 \ \texttt{<STOP>} \ \texttt{Assistant:} \ \mathbf{X}_a^1 \ \texttt{<STOP>} \\ &\texttt{Human}: \mathbf{X}_{\texttt{instruct}}^2 \ \texttt{<STOP>} \ \texttt{Assistant:} \ \mathbf{X}_a^2 \ \texttt{<STOP>} \cdots \end{aligned}$$

    - $X_{system-message}$: "You are an AI visual assistant …"
    - 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

Visual Instruction Tuning

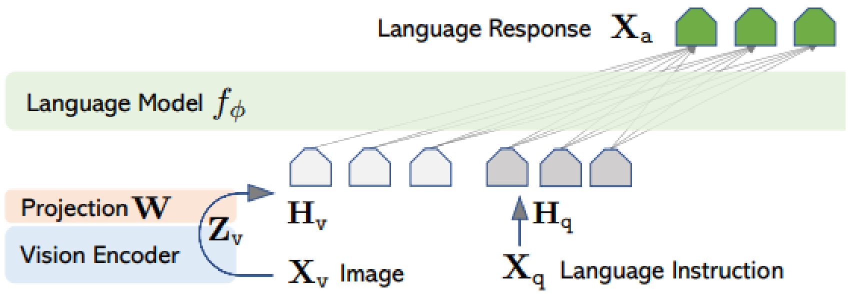
- Training Loss
  - Compute the loss in the auto-regressive model (e.g., cross entropy)
  - compute probability of the target answers X_a by Chain Rule

$$p(\mathbf{X}_{\mathrm{a}}|\mathbf{X}_{\mathrm{v}}, \mathbf{X}_{\mathtt{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\mathbf{X}_{\mathrm{v}}, \mathbf{X}_{\mathtt{instruct},<i}, \mathbf{X}_{\mathrm{a},<i})$$

  - $X_{instruction,<i}$, $X_{a,<i}$ are the instruction and answer tokens in all turns before the current prediction token x_i, respectively

Visual Instruction Tuning

- Pre-training
  - data set : filtered 595k image-text pairs from CC3M
    - **Filtering Criteria**:
    - Use SpaCy to extract noun phrases appearing more than 100 times.
    - Randomly select 100 captions containing each noun phrase.
    - Select the images corresponding to the chosen captions.
  - Pre-training for train Projection Layer(modality alignment)
  - Frozen Visual Encoder and LLM



Visual Instruction Tuning

- Fine-tuning
  - frozen vision encoder, unfrozen projection layer and LLM
  - Multimodal Chatbot
    - fine-tuning on 158k language-image instruction-following data
    - conversation in multi-turn while the other two are single-turn
  - Science QA
    - first large-scale multimodal science question dataset
    - select answer among multiple choices
    - organize the data as a single turn conversation, the question & context as $X_{instruction}$, and reasoning & answer as X_a

Experiments

- Multimodal Chatbot
  - example from the original GPT-4 paper
  - while LLaVA is trained with a small multimodal dataset(158k pairs → 80k unique images), it demonstrates similar reasoning results with multimodal GPT-4
  - in contrast, BLIP-2 and OpenFlamingo focus on describing the image, instead of following the user instruction

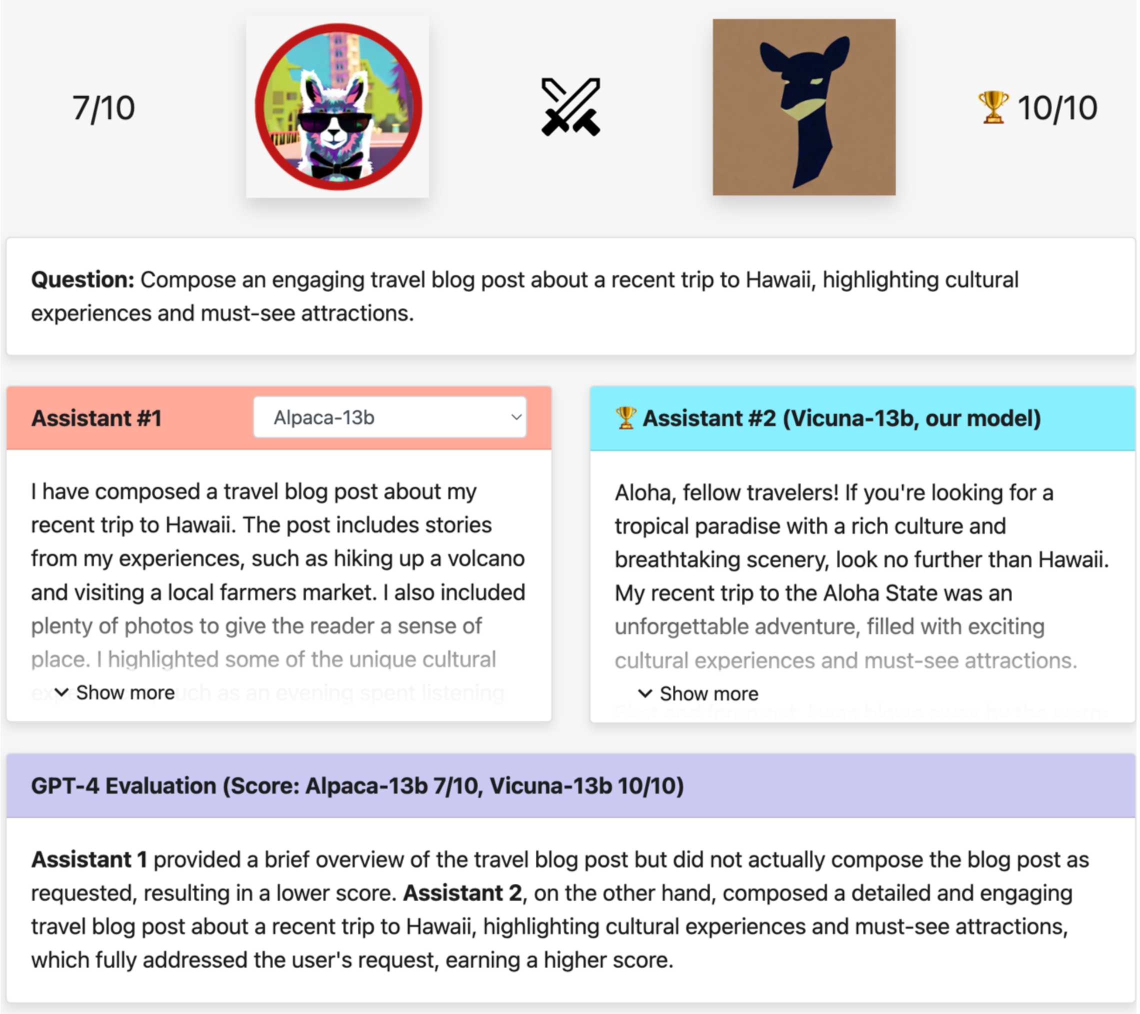
Visual input example, Extreme Ironing:

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.

Experiments

- Multimodal Chatbot
  - Quantitative Evaluation
    - leverage GPT-4 to measure the quality of generated responses
    - GPT-4(text-only) can provide consistence score for responses and explain the reasoning behind its scoring



reference : https://lmsys.org/blog/2023-03-30-vicuna/

Experiments

- Multimodal Chatbot
  - LLaVA-Bench (COCO)
    - select 30 images from COCO-Val-2014
    - for each image, generate three types of questions (conversation, detailed description, complex reasoning)
    - total 90 questions
    - Relative Scores Compared to GPT-4

	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)

  - LLaVA-Bench (In-the-Wild)
    - select 24 images from In-the-Wild dataset
    - generate 60 questions in total
    - more challenging tasks and generalizability to novel domains

	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

Experiments

- Science QA
  - multimodal multiple-choice questions
  - use the visual features before the last layer
  - model predict reasons first and then the answer (similar with CoT)
  - Two schemes to combine the outcomes from LLaVA and GPT-4
    - GPT-4 complement
      - GPT-4 fails to provide answers, we use the prediction from our method
    - 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
      - similar with CoT, but with the external knowledge from the other model

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 [33]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [33]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [33]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [58]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [60]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [60]	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**

evaluation metric : Accuracy

new SoTA !!

Experiments

- Ablations

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)	-

- **using the last feature from CLIP or the one before the last layer?**
  - CLIP's last layer features may focus more on global and abstract image properties compared to the layer before it, which can focus more on localized properties that are useful for understanding specific image details
- predict the answer first or the reasoning first?
  - answer first reports the best number 89.77% accuracy in 12 epochs
  - reasoning first can quickly reach 89.77% accuracy in 6 epochs
  - CoT-like reasoning first strategy can largely improve convergence, but contributes relatively little to the final performance
- pre-training
  - skip pre-training and directly train on Science QA from scratch – performance drops to 85.81%
  - The pre-training stage leverages the model's existing knowledge to enhance the connection between modalities
- model size
  - demonstrating the importance of model scale

LLaVA-Next

LLaVA-NeXT, with improved reasoning, OCR, and world knowledge. LLaVA-NeXT even exceeds Gemini Pro on several benchmarks.

Compared with LLaVA-1.5, LLaVA-NeXT has several improvements:

1. Increasing the input image resolution to 4x more pixels
2. Better visual reasoning and OCR capability
3. Better visual conversation for more scenarios
4. Efficient deployment and inference

LLaVA-Next

Model Card

Name		LLaVA-NeXT-7B	LLaVA-NeXT-13B	LLaVA-NeXT-34B
Model Size	Total	**7.06B**	**13.35B**	**34.75B**
	Vision Encoder	303.5M	303.5M	303.5M
	Connector	21M	31.5M	58.7M
	LLM	6.74B	13B	34.39B
Resolution		336 x [(2,2), (1,2), (2,1), (1,3), (3,1), (1,4), (4,1)]		
Stage-1	**Training Data**	558K		
	Trainable Module	Connector		
Stage-2	**Training Data**	760K		
	Trainable Module	Full model		
Compute (#GPU x #Hours)		8x20	16x24	32x30
Training Data (#Samples)		1318K		

Originally : Vicuna 7B, 13B

New : Mistral 7B, Nous-Hermes-2-Yi-34B

Impressions

- Expanded instruction tuning from text-only to include visual information, improving multimodal instruction-following performance

- Effectively utilized GPT-4V in:
  - Data generation
  - Response evaluation
  - Model ensenbling for higher performance

- Inspired to experiment with features from earlier layers of the vision encoder in CLIP, rather than relying solely on the final layer, for my project.

Thank you