LLaVA Family of Models

LLaVA - NeurIPS 2023
LLaVA 1.5 - CVPR 2024
LLaVA One Vision - TMLR 2025

Outline

1. Introduction
2. Background
3. Problem Statement
4. Related Works
5. Approach
6. Experiments & Results
7. Limitations, Societal Implications
8. Summary of Strengths, Weaknesses, Relationship to Other Papers

Introduction

Niki Vasan
MSA
Interests: *Agentic AI, MoE Models*

William Stevens
ML PhD
Interests: *Vision, GenAI, Multimodality*

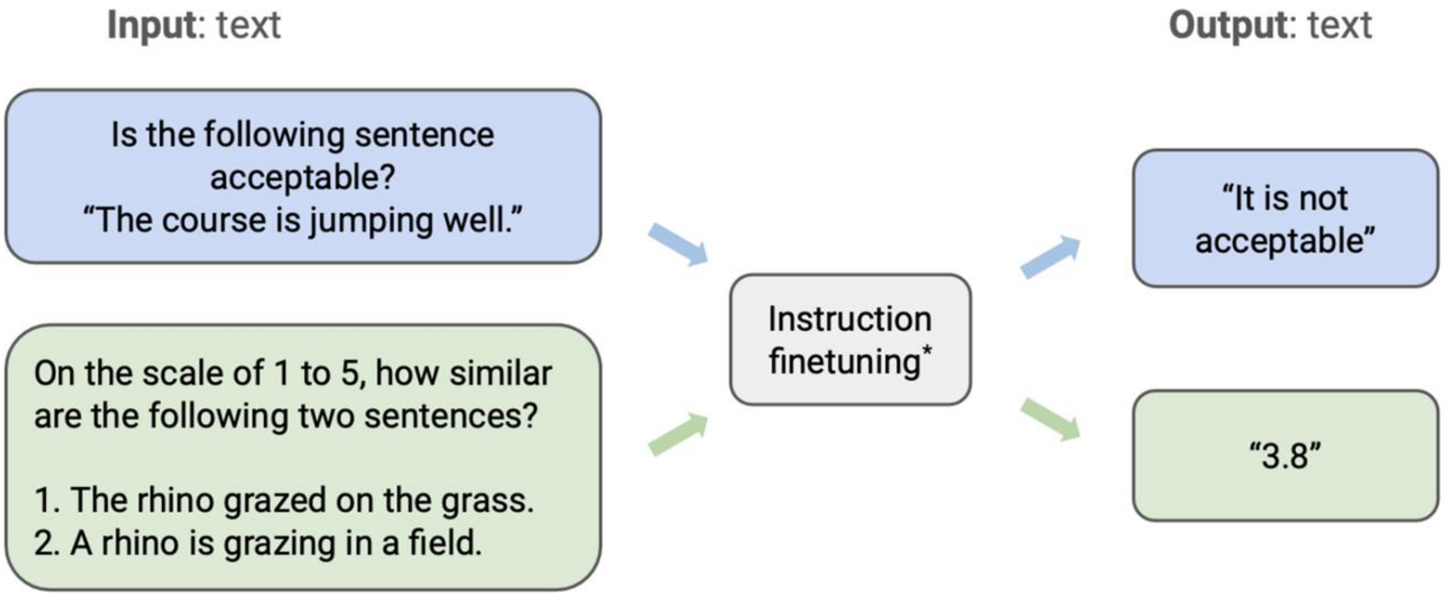
Ethan Haarer
MSCS
Interests: *Symbolic AI, Gen AI, Agentic Systems*

Georgia Tech

LLaVA

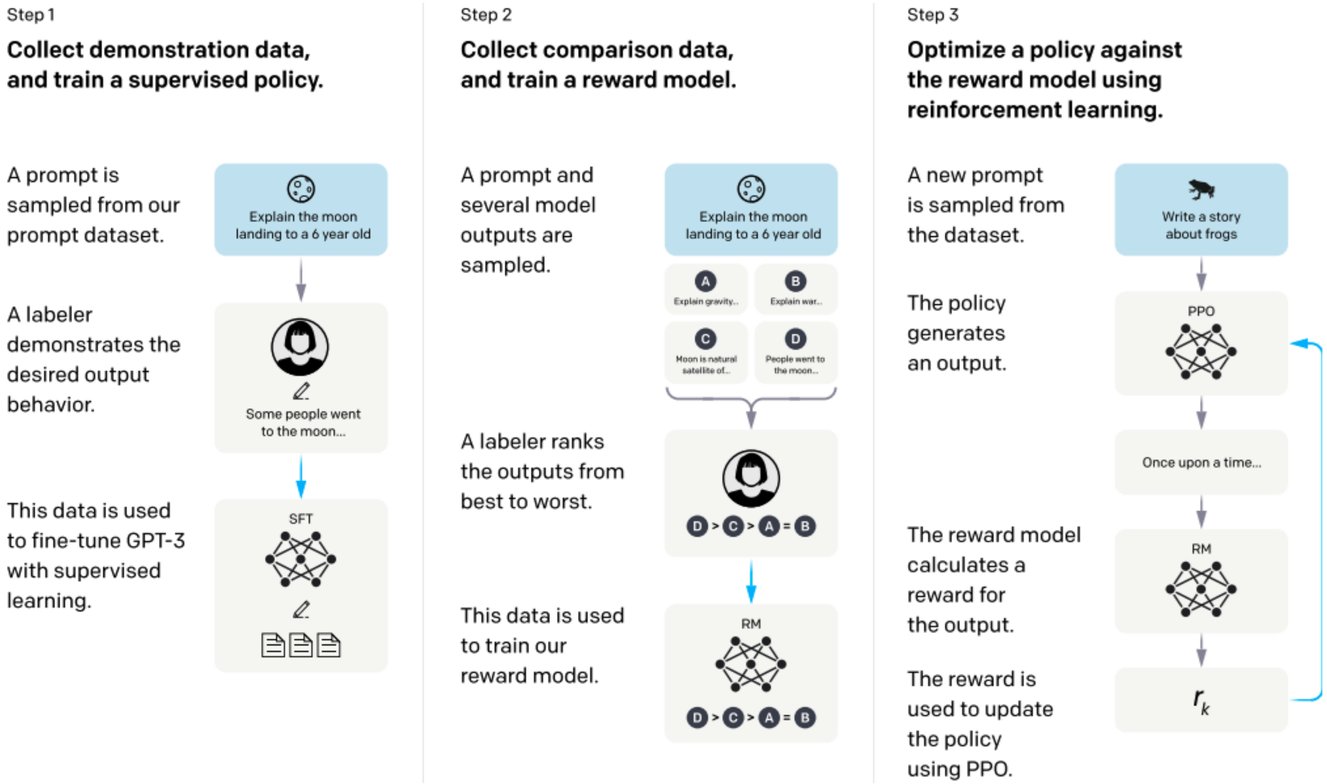
Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee

Background



Input: text
Output: text
Is the following sentence acceptable?
"The course is jumping well."
On the scale of 1 to 5, how similar are the following two sentences?
1. The rhino grazed on the grass.
2. A rhino is grazing in a field.
Instruction finetuning*
"It is not acceptable"
"3.8"

Background



Step 1
Collect demonstration data, and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain the moon landing to a 6 year old
A labeler demonstrates the desired output behavior.
Some people went to the moon...
This data is used to fine-tune GPT-3 with supervised learning.
SFT
Step 2
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled.
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
A labeler ranks the outputs from best to worst.
D > C > A = B
This data is used to train our reward model.
RM
D > C > A = B
Step 3
Optimize a policy against the reward model using reinforcement learning.
A new prompt is sampled from the dataset.
Write a story about frogs
The policy generates an output.
PPO
Once upon a time...
The reward model calculates a reward for the output.
RM
r_k
The reward is used to update the policy using PPO.

Problem Statement

Problems:

1. Traditional VLMs: language only describes images, limited interactivity.
2. Lack of vision-language instruction-following data and benchmarks.

Goal: Build a *general-purpose visual assistant* with instruction-following.

Key Idea: Visual Instruction Tuning (LLaVA-Main).

Open Source!

https://github.com/LLaVA-Annonymous/LLaVA/tree/master

Problem Statement

Caption-based supervision (traditional)

A boy rides a red bicycle in a field.

Instruction-following supervision (LLaVA)

User: "What color is the bicycle?"

Model: "The bicycle is red."

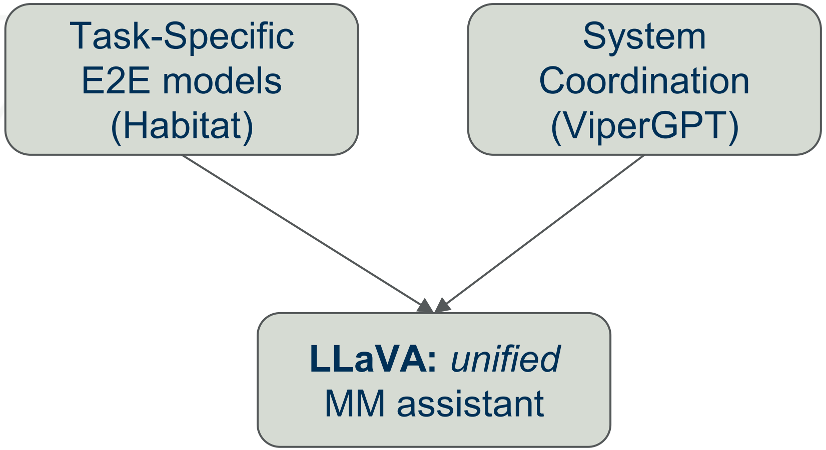
User: **"**Write a short story about this picture."

Model: "A young boy speeds through the field on his shiny red bike, enjoying the sunny weather…"

Georgia Tech

Related Works

Multimodal Instruction-following Agents



Multimodal Instruction-Tuning

- No explicit tuning on vision-language instruction data
- Image-Text Pair Training: Flamingo, BLIP-2

Instruction-Tuning Data Generation

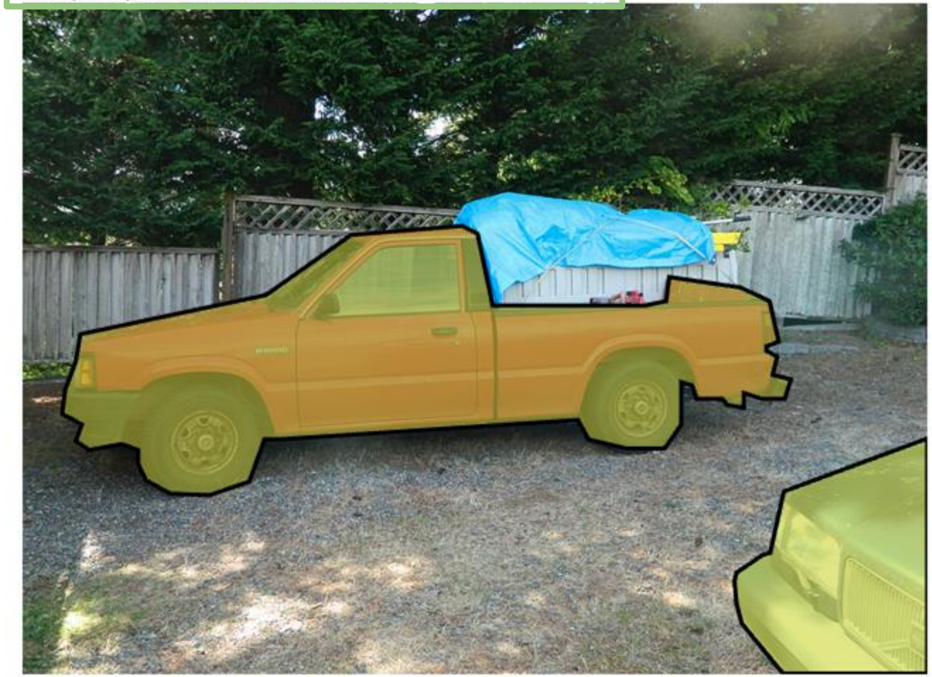
Problem:
- Most public multimodal data is in the form of image-text pairs
- Lack of multimodal instruction-following data

Solution: Use **text-only** GPT-4 for to generate instruction-following data

Instruction-Tuning Data Generation

COCO

a red pick up truck with a large blue object in it's back.
a small red pickup truck loaded with things in the back.
a truck is parked in the driveway with something in the back.
an older model red truck is parked in the gravel.
a red pick up truck next to a fence.



Instruction-Tuning Data Generation

"Symbolic Representation" of images:

1. Captions
2. Bounding Boxes

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) ...<omitted>

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. ...<omitted>

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 ...<omitted>

Instruction-Tuning Data Generation

"Symbolic Representation" of images:

1. Captions
2. **Bounding Boxes**

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) ...<omitted>

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. ...<omitted>

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 ...<omitted>

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

---

1 **Response type 1: conversation**

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

2 **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. ...<omitted>

3 **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 ...<omitted>

Instruction-Tuning Data Generation

```
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)})
```

System Prompt

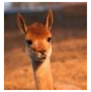
- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before 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"

Detailed Description Questions

Instruction-Tuning Data Generation

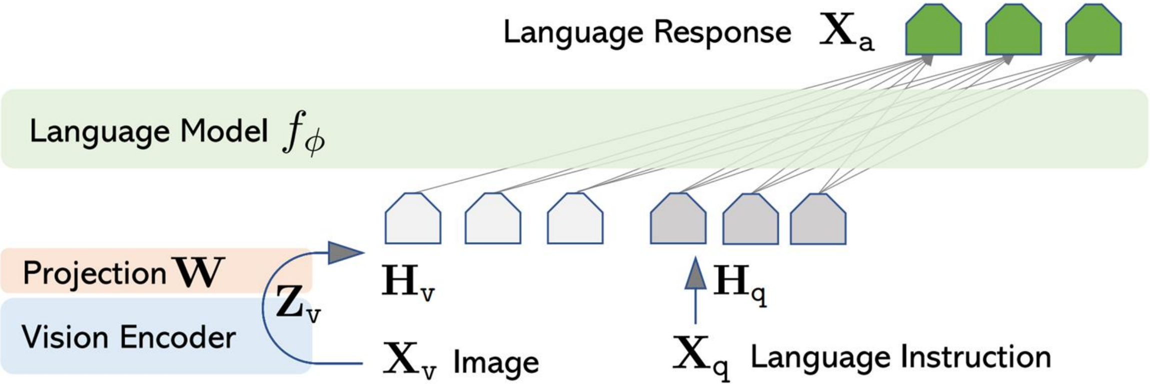
LLaVA-Instruct-158K Dataset	
Type	**Length**
Conversation	58K
Detailed Description	23K
Complex Reasoning	77K
Total	158K

Architecture



Vicuna V0

CLIP ViT-L/14



Encoder Visual Features: $\mathbf{Z}_v = g(\mathbf{X}_v)$

Language Embedding Tokens: $\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v$

Training - Image to Conversation Tokens

Each Input Image is converted into multi-turn conversation data with question-answer pairings

$$X_v = (X_q^1, X_a^1, \dots, X_q^T, X_a^T)$$

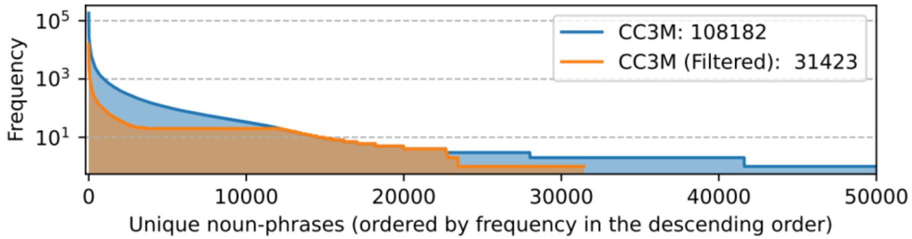
Perform instruction-tuning for sequence of length L, computing probability of target answer X_a:

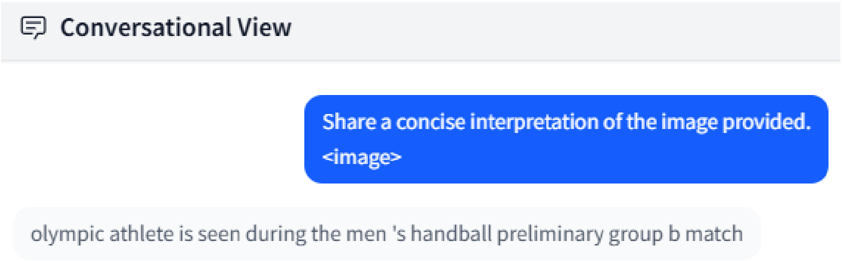
$$p(\mathbf{X}_a \mid \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^{L} p_\theta(x_i \mid \mathbf{X}_v, \mathbf{X}_{\text{instruct},<i}, \mathbf{X}_{a,<i}) .$$

Training - Stage 1: Pretraining

Dataset: CC3M
Images distilled for coverage: 3.3 million to 595K image-text pairs

- Noun-Phrases are extracted and unique frequencies recorded
- Lowest frequencies with less than 3 instances are removed





Georgia Tech

Training - Stage 2: Fine tune End-to-End

Multimodal Chatbot

158K language-image instruction-following data

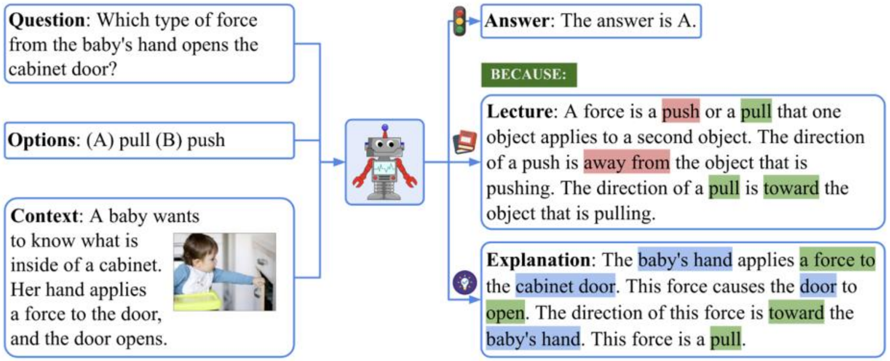
Captions
A man is skiing in the open snow covered hills
A skier is making his way into the snow.
A skier on the foothills of a huge mountain range.
A skier looks at mountains while standing near a trail sign.
a single person skiing by an area with a lot of bushes

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.

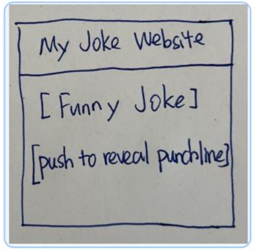
Science QA

Trains predominantly on detailed multimodal science lectures (X_instruce)

Model selects answer to multiple choice questions (X_a)



Results - Multimodal Chatbot

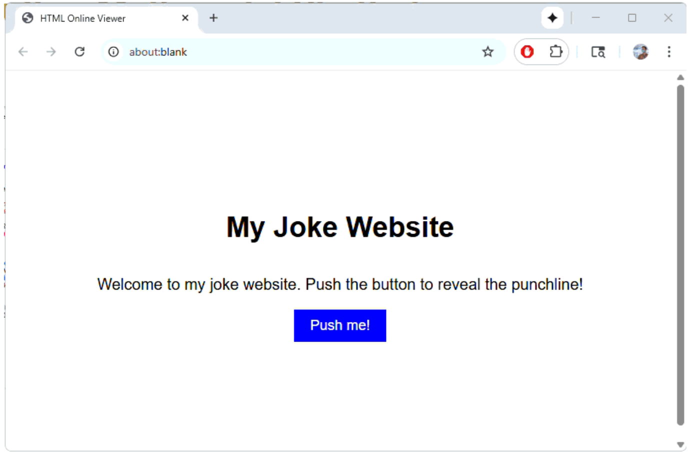


User

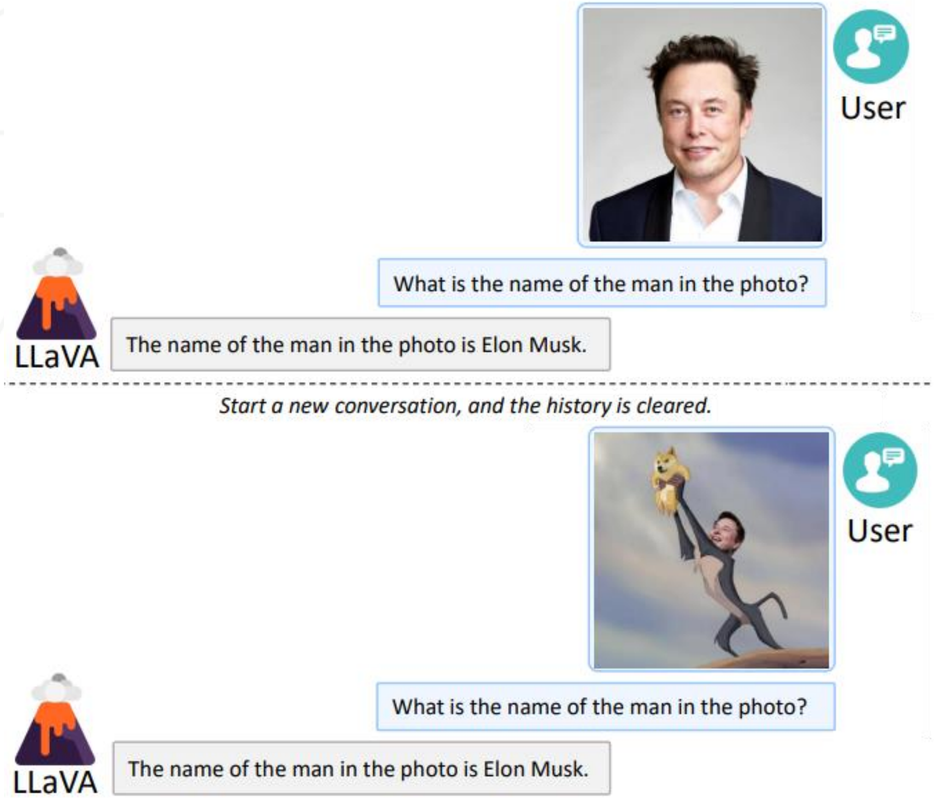
Turn this mock-up into an interactive website using html/js, replace the joke with a real joke.

LLaVA

```
<!DOCTYPE html>
<html>
 <head>
  <title>My Joke Website</title>
  <style>
   body { background-color: white; font-family: Arial, sans-serif; }
   .container { display: flex; flex-direction: column; align-items: center; justify-content: center; height: 100vh; }
   h1 { font-size: 36px; margin-bottom: 20px; }
   p { font-size: 20px; }
   button { background-color: blue; color: white; border: none; padding: 10px 20px; font-size: 18px; cursor: poi
  </style>
 </head>
 <body>
  <div class="container">
   <h1>My Joke Website</h1>
   <p id="punchline">Welcome to my joke website. Push the button to reveal the punchline!</p>
   <button onclick="revealPunchline()">Push me!</button>
  </div>
  <script>
   function revealPunchline() {
    document.getElementById("punchline").innerHTML = "The punchline is: 'Why was the math book sad?
problems.'";
   }
  </script>
 </body>
</html>
```



Results - Multimodal Chatbot



Unseen Identification

- Elon is able to be identified between image contexts
- CLIP encoder may have seen Musk beforehand
- Musk does NOT appear in training data for either visual feature alignment or visual instruction tuning
- LLaVA's Vicuna LM generalizes to unseen encoded visual concepts

Results - LLM "as a judge"

Triplet Evaluation:

Image

Question

Question:
How many dogs in this image have red leashes?

Textual Visual Info

Caption: "4 small dogs in a row running directly at view…"

Bounding boxes:
Black Dog:(40, 35, 122, 167), Terrier…

LLM as a Judge

- A separate text-only GPT-4 is then given the question, the same textual visual info, and both answers side-by-side.
- It evaluates helpfulness, relevance, accuracy, and level of detail and assigns an overall score from 1–10
- This is a pairwise comparison between the candidate and the reference answer

Results - Ablation

	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 (COCO)

- Data. 30 COCO-Val-2014 images × 3 question types (Conversation, Detailed description, Complex reasoning) = 90 total prompts.
- The ablation shows instruction tuning is crucial (≈ +50 pts), and adding a bit of detail + complex reasoning data lifts overall ability by ~7 pts

Results - Quantitative Evaluation

	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 [28]	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

LLaVA-Bench (In the Wild) Results

- Models assessed on 24 images with 60 questions total
- Instruction tuning on LLaVA achieves more significant performance
- Numbers are relative scores (%) with mean ± std over multiple inference runs. They also re-run the judge three times on the same LLaVA outputs (LLaVA†) to check judge stability—scores barely move, indicating consistent judging.

Georgia Tech

Results - Quantitative Evaluation

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 [34]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [34]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [34]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [59]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [61]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [61]	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**

Accuracy % on Science QA

- Tested on NATural, SOCial, LANguage sciences, with context given in text, image, or none at all in grades 1-6 or 7-12

LLaVA-1.5

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

Some Weaknesses of LLaVA

Question: Is there strawberry-flavored yogurt in the fridge?
LLaVA: Yes.

Question: What’s the name of the restaurant?
LLaVA: ???

LLaVA-1.5: Problem Statement

Original LLaVA:

- Excelled at conversation-style visual reasoning
- Fell short in traditional visual-QA scenarios
- Like other LMMs developed around this time, it only performed well on a specific type of task

Goal: Build a stronger *general-purpose visual assistant* that is designed with a larger variety of baselines and robust benchmarks

Georgia Tech

LLaVA's Shortcomings on LMM Benchmarks

Model	GQA	MME	MM-Vet
BLIP-2	41	1293.8	22.4
LLaVA	-	809.6	25.5
InstructBLIP	49.5	1212.8	25.6
Qwen-VL-Chat	**57.5**	**1487.5**	**45.7**

Main comparison in original paper

After original LLaVA paper

Improvements on Three Fronts

Fine-tuning datasets:

- Incorporated academic-task-oriented data
- Adjusted prompt formats to prevent over-rambling
- Included multilingual data

Model Capabilities:

- MLP instead of linear layer
- LLM size scaled up to 13B from 7B

Image Resolution:

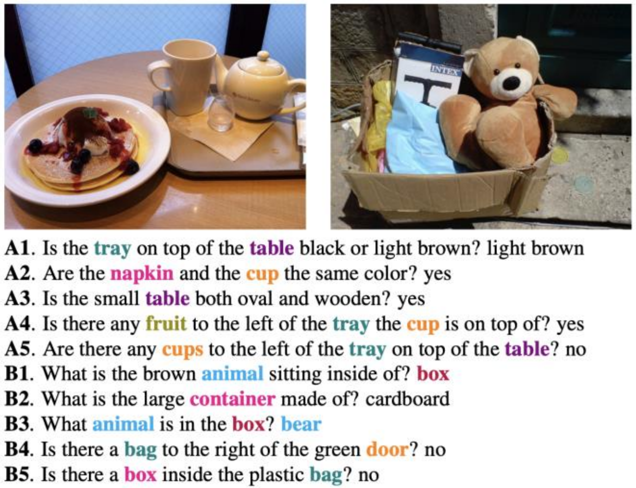
- CLIP vision encoder upgraded to highest resolution (336 x 336)

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

New Fine-tuning Datasets

- **VQA-v2 and GQA**
  - General academic tasks (Open ended short answer, true/false, etc.)
  - Prompt + "Answer the question using a single word or phrase"
- **OKVQA**
  - Academic tasks (multiple choice)
  - Prompt + "Answer with the option's letter from the given choices directly"
- **OCRVQA**
  - Images contain text vital to the academic task
- **Region-level VQA (Visual Genome, RefCOCO)**
  - Requires localization of fine-grained details
- **ShareGPT**
  - Allows for multilingual abilities (Spanish, Japanese, Korean, Chinese, etc.)

VQA-v2 Example



GQA Example

Scaling Beyond the CLIP Limit (336 x 366)

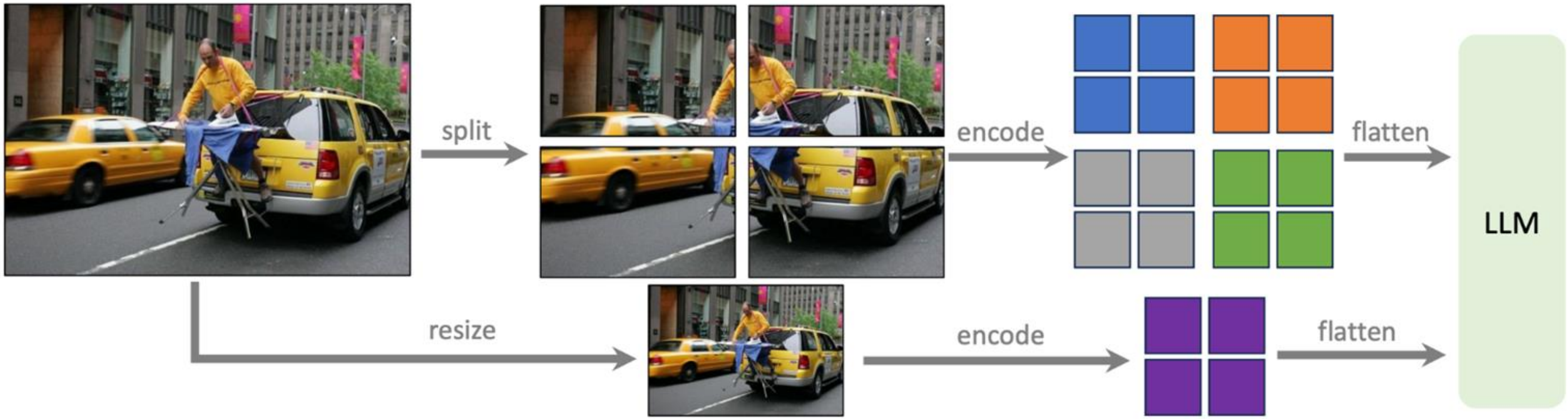


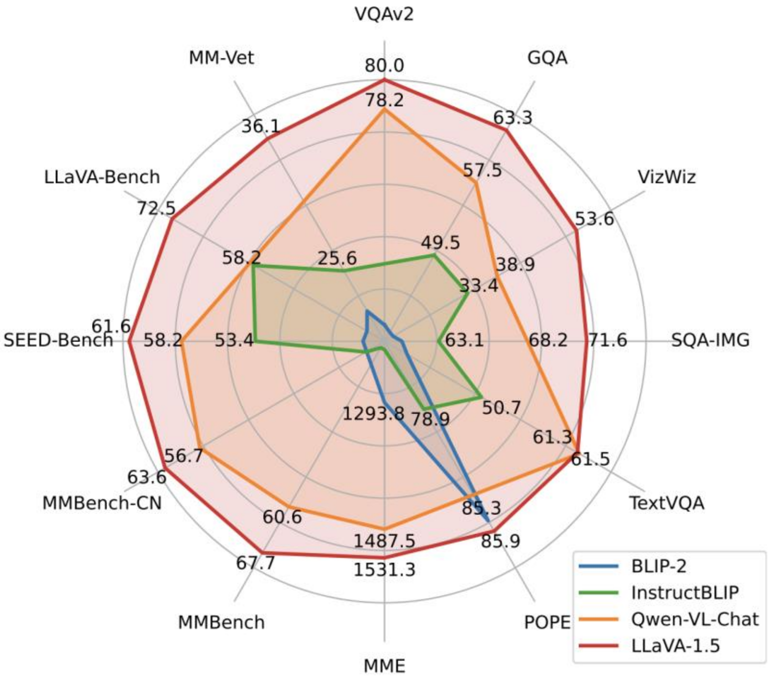
Figure 2. **LLaVA-1.5-HD.** Scaling LLaVA-1.5 to higher resolutions by splitting the image into grids and encoding them independently. This allows the model to scale to any resolution, without performing positional embedding interpolation for ViTs. We additionally concatenate the feature of a downsampled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

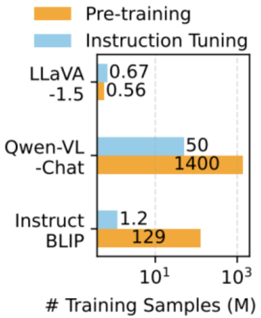
→ No need for additional fine-tuning on larger image-text pairs

Georgia Tech

LLaVA-1.5 Results

Model	GQA	MME	MM-Vet
BLIP-2	41	1293.8	22.4
LLaVA	-	809.6	25.5
InstructBLIP	49.5	1212.8	25.6
Qwen-VL-Chat	57.5	1487.5	**45.7**
LLaVA-1.5-13B	63.3	**1531.3**	36.1
LLaVA-1.5-13B-HD	**64.7**	1500.1	39.4





High training efficiency relative to comparable models

Recap: Weaknesses of LLaVA

Question: Is there strawberry-flavored yogurt in the fridge?
LLaVA: Yes.

Question: What's the name of the restaurant?
LLaVA: ???

Who thinks LLaVA-1.5 will do better than LLaVA on these examples of weaknesses they identified?

Our Experiments with LLaVA-1.5

Question: Is there strawberry-flavored yogurt in the fridge?
LLaVA: Yes.

LLaVA-1.5: Yes.

Still wrong!

Question: What's the name of the restaurant?
LLaVA: ???

LLaVA-1.5: The name of the restaurant is "Sushi Rice."

Wrong! It's "ICHI-RAN"

Interesting Ablations

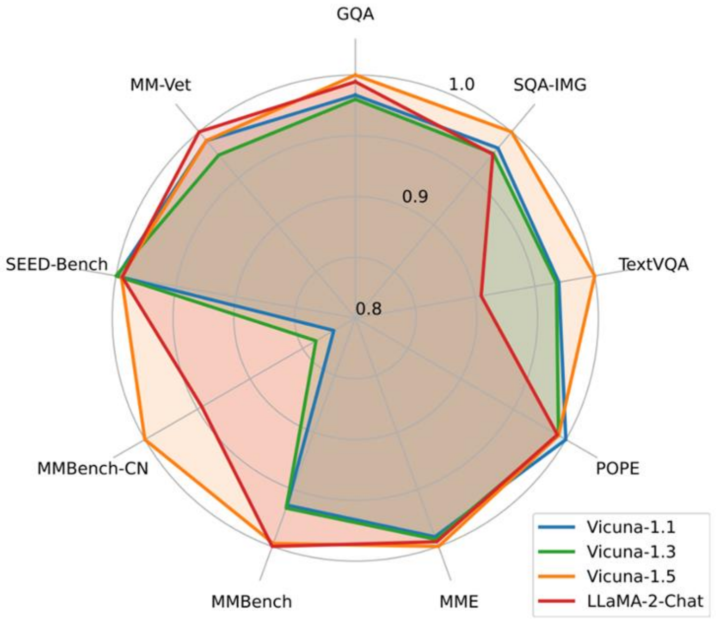


Figure 3. **Ablation on LLM choices**. Data points represent the relative performance of the best performing variant for each dataset.

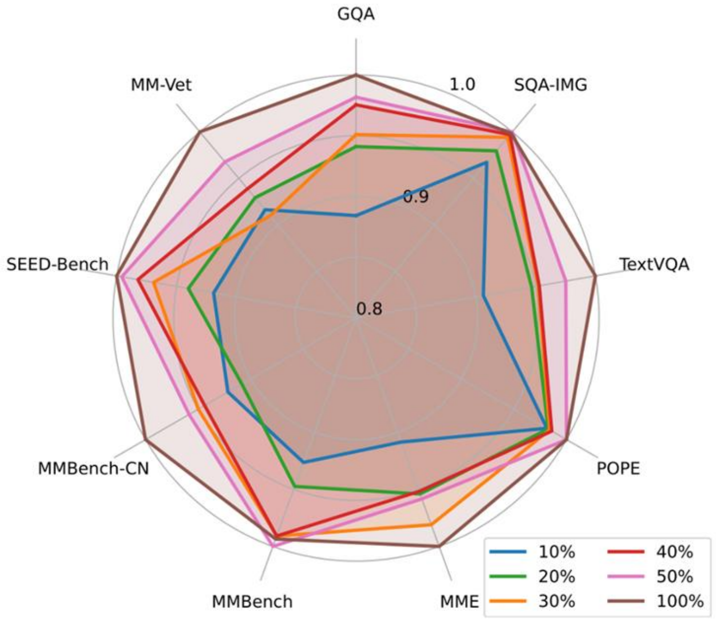


Figure 4. **Ablation on data efficiency.** Data points represent the relative performance of the best performing variant for each dataset.

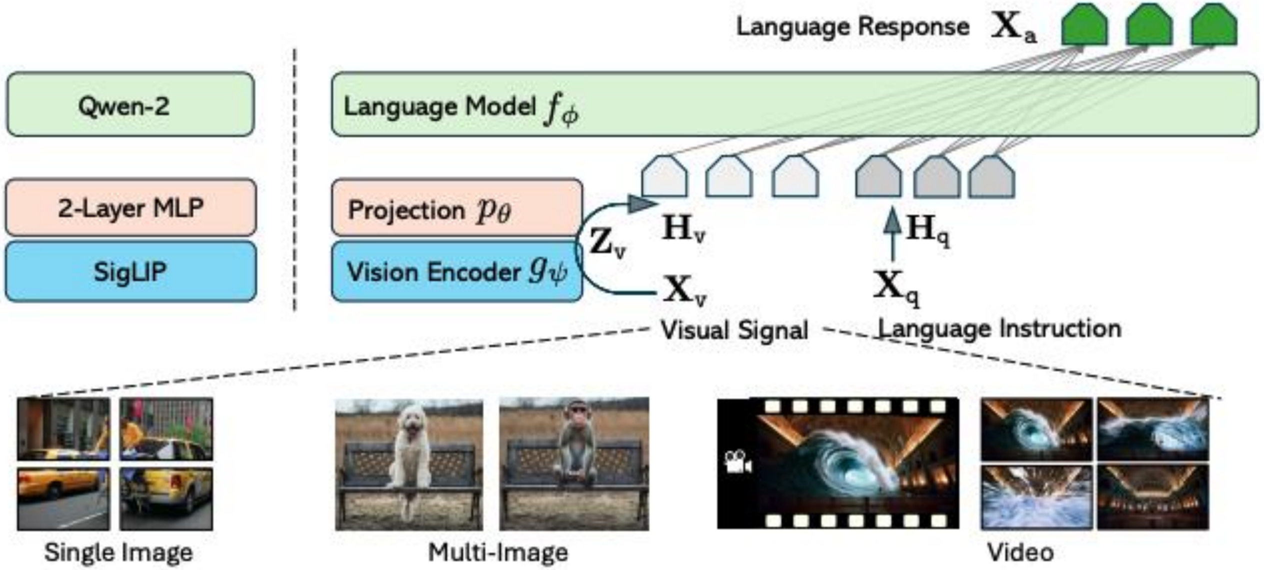
Vicuna-1.5 is strongest base LLM

Dropping training data to 50% capacity maintains 98% of the performance

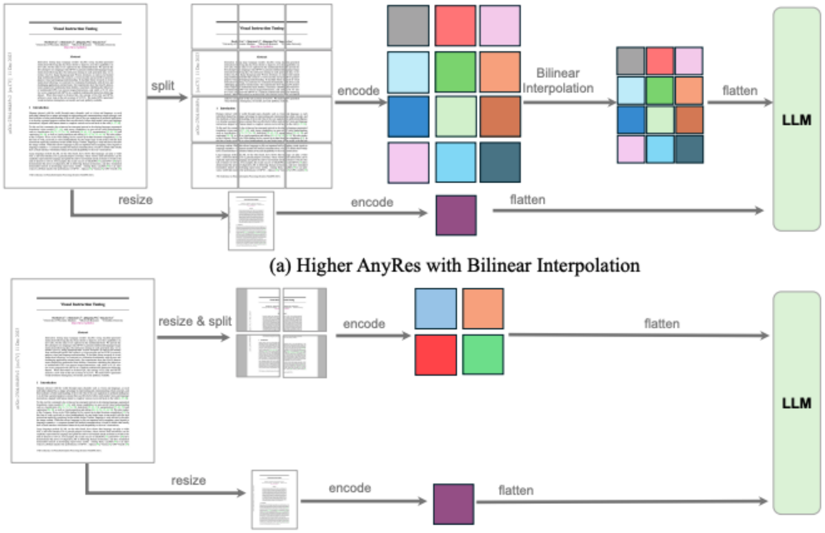
LLaVA One Vision

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang,Feng Li, Hao Zhang, Kaichen Zhang ,Peiyuan Zhang, Yanwei Li, Ziwei Liu ,Chunyuan Li

Architecture



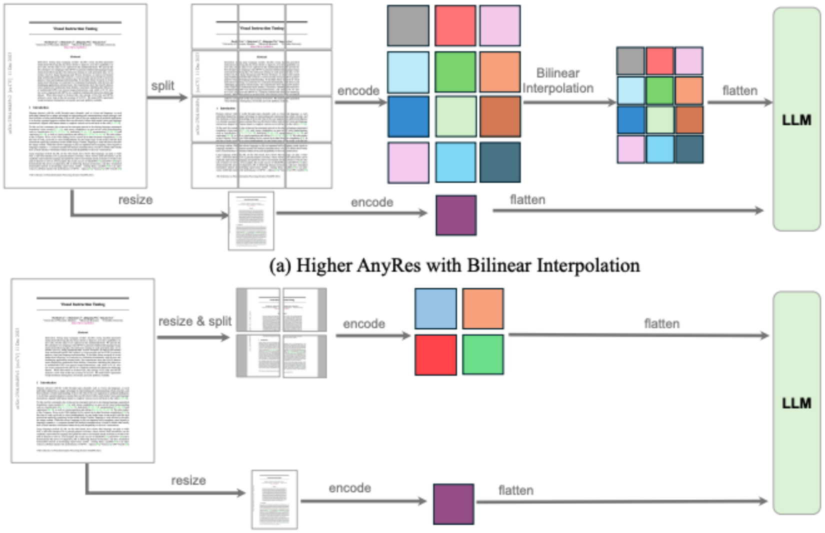
Innovations



(a) Higher AnyRes with Bilinear Interpolation

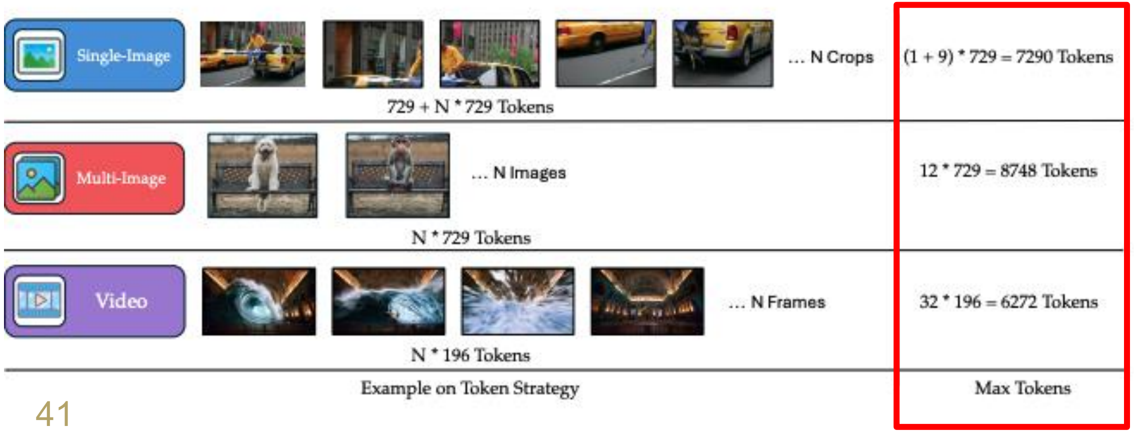
(b) The original AnyRes

Innovations

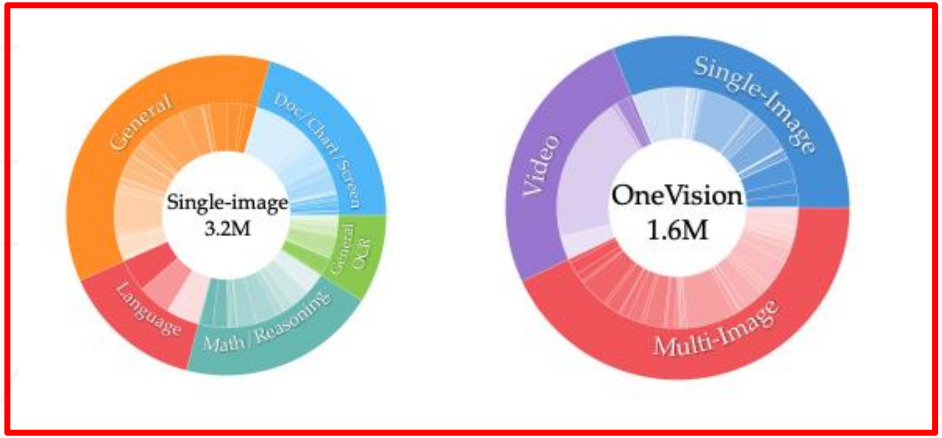


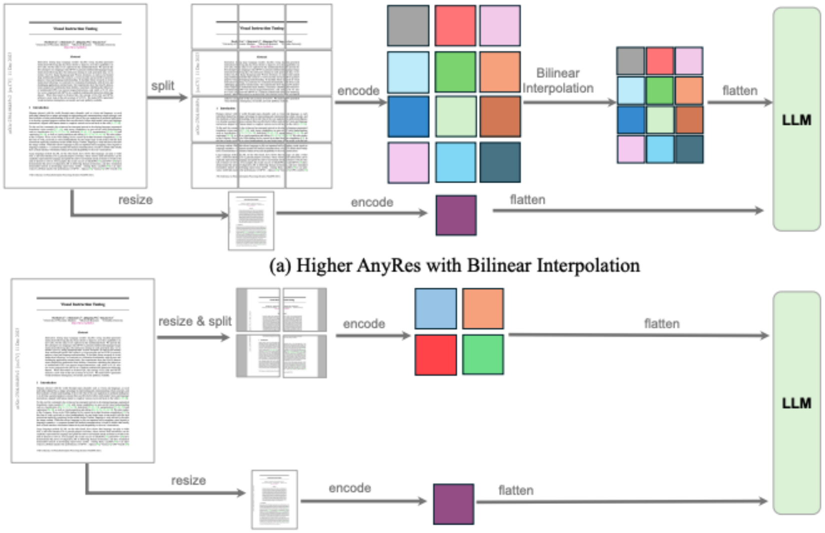
(a) Higher AnyRes with Bilinear Interpolation

(b) The original AnyRes



Innovations

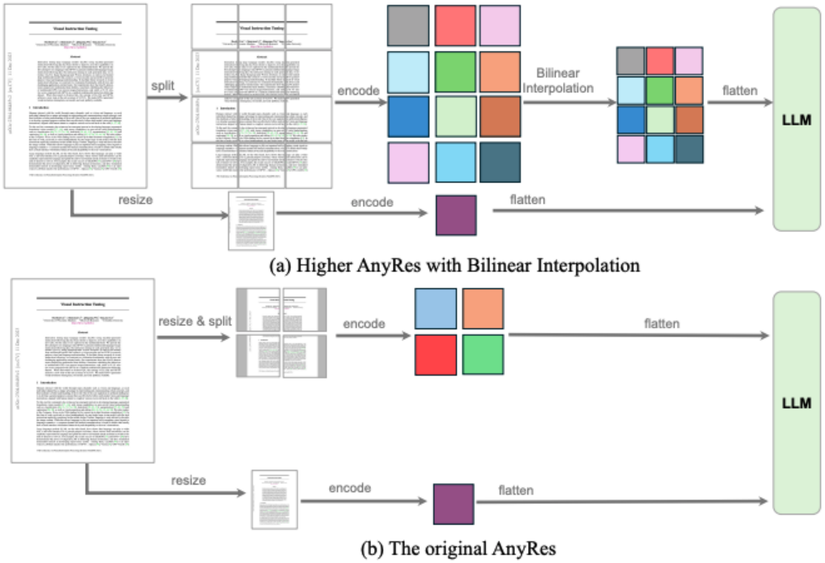


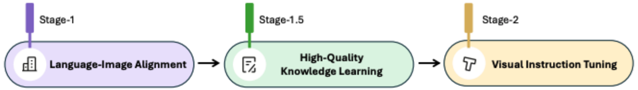


(a) Higher AnyRes with Bilinear Interpolation

(b) The original AnyRes

Innovations





Results

Capability	Benchmark	LLaVA OneVision-0.5B	LLaVA OneVision-7B	LLaVA OneVision-72B	GPT-4V (V-Preview)	GPT-4o
Single-Image	†AI2D [53] Science Diagrams	57.1%	81.4%	85.6%	78.2%	94.2%
	†ChartQA [101] Chart Understanding	61.4%	80.0%	83.7%	78.5%	85.7%
	†DocVQA [103] (test) Document Understanding	70.0%	87.5%	91.3%	88.4%	92.8%
	†InfoVQA [102] (test) Infographic Understanding	41.8%	68.8%	74.9%	-	-
	MathVerse [165] (vision-mini) Professional Math Reasoning	17.9%	26.2%	39.1%	32.8%	50.2%
	MathVista [90] (testmini) General Math Understanding	34.8%	63.2%	67.5%	49.9%	63.8%
	MMBench [86] (en-dev) Multi-discip	52.1%	80.8%	85.9%	75.0%	-
	MME [28] (cog./perp.) Multi-discip	240/1238	418/1580	579/1682	517/1409	-
	MMStar [19] Multi-discip	37.5%	61.7%	66.1%	57.1%	-
	MMMU [157] (val) College-level Multi-disp	31.4%	48.8%	56.8%	56.8%	69.1%
	MMVet [153] Multi-discip	29.1%	57.5%	63.7%	49.9%	76.2%
	SeedBench [66] (image) Multi-discip; Large-scale	65.5%	75.4%	78.0%	49.9%	76.2%
	†ScienceQA [93] High-school Science	67.2%	96.0%	90.3%	75.7%	-
	ImageDC [65] Image Detail Description	83.3%	88.2%	91.2%	91.5%	-
	RealworldQA [141] Realwold QA	55.6%	66.3%	71.9%	61.4%	-
	Vibe-Eval [112] Chanllenging Cases	33.8%	51.7%	50.7%	57.9%	63.1%
	MM-LiveBench [161] (2406) Internet Content Understanding	49.9%	77.1%	81.5%	-	92.4%
	LLaVA-Wilder [65] (small) Realworld Chat	55.0%	67.8%	72.0%	81.0%	85.9%

Multi-Image	LLaVA-Interleave [68] Out-domain	33.3%	64.2%	79.9%	60.3%	-
	MuirBench [135] Comprehensive Multi-image	25.5%	41.8%	54.8%	62.3%	-
	Mantis [47] Multi-image in the Wild	39.6%	64.2%	77.6%	62.7%	-
	BLINK [31] Unusual Visual Scenarios	52.1%	48.2%	55.4%	51.1%	-
	†Text-rich VQA [84] OCR, Webpage, Ducument	65.0%	80.1%	83.7%	54.5%	-
Video	ActivityNetQA [155] Spatio-Temporal Reasoning	50.5%	56.6%	62.3%	57.0%	-
	EgoSchema [98] Egocentric Video Understanding	26.8%	60.1%	62.0%	-	-
	PerceptionTest [115] Perception and Reasoning	49.2%	57.1%	66.9%	-	-
	SeedBench [66] (video) Multi-discip; Video	44.2%	56.9%	62.1%	60.5%	-
	LongVideoBench [138] (val) Long Video	45.8%	56.3%	63.2%	60.7%	66.7%
	MLVU [170] Long Video Understanding	50.3%	64.7%	68.0%	49.2%	64.6%
	MVBench [71] Multi-discip	45.5%	56.7%	59.4%	43.5%	-
	VideoChatGPT [97] Video Conversation	3.12	3.49	3.62	4.06	-
	VideoMME [29] Multi-discip	44.0%	58.2%	66.2%	59.9%	71.9%

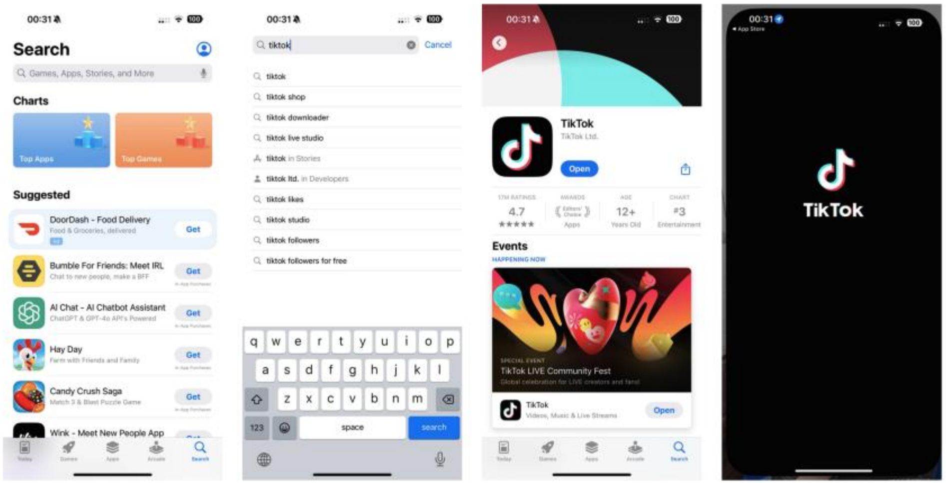
Emerging Capabilities

S1: Joint understanding of diagram and chart (Transfer from single-image to multi-image)
S2: GUI for multi-modal agent (Transfer from single-image and multi-image).
S3: Set-of-mark Prompting (Transfer from single-image task composition).
S4: Image-to-Video Editing Instruction (Transfer from single-image and video).
S5: Video-to-Video Difference (Transfer from multi-image and video)
S6: Multi-camera Video Understanding in Self-driving (Transfer from single-image and multi-image to video).
S7: Composed Sub-video Understanding (Transfer from multi-image to video).
S8: Visual prompting in video (Task transfer from single-image to video).
S9: Visual Referring in Image in Video Understanding.

Emerging Capabilities

S2: GUI for multi-modal agent (Transfer from single-image and multi-image).

S2: Multi-image GUI Understanding for Agents



User	There are four sequential mobile UI screens, and are transformed by three steps of tap operations. Please respectively describe the three tap operations.
LLaVA-OV	The image shows a sequence of four mobile application interface screens. Here's a breakdown of the three tap operations based on the visible changes: 1. First Tap Operation: - The user starts by tapping on the search bar at the top of the screen. - The keyboard is shown at the bottom of the screen, indicating that the user can now type a search query. - The user types "TikTok" in the search bar and taps the search icon on the keyboard to initiate a search. 2. Second Tap Operation: - After tapping the search icon or pressing 'Enter', the screen displays a list of search results. - The user taps on the first result, which is the TikTok application. - This action opens the TikTok's app store page, which includes the app icon, developer information ... 3. Third Tap Operation: - The user taps the "Open" button on the app store page. - This action opens the TikTok application, and the screen transitions to the app's main interface.

Emerging Capabilities

S5: Video-to-Video Difference (Transfer from multi-image and video).

S5: Video-to-Video Difference

User	Here are two videos, please provide the analysis on their differences.
LLaVA-OV	The first video shows a sequence of images where a man in a blue plaid shirt approaches a woman in a red dress, engages in a conversation, and eventually gives her a kiss on the cheek. The woman in the red dress is looking at the man with a slight smile, and another woman in a light blue top is observing the scene. The setting appears to be a street with buildings in the background. The second video shows the same man in the blue plaid shirt walking away from the woman in the red dress, who is now looking in the opposite direction. The woman in the light blue top is no longer visible in this sequence. The man is walking down the street, passing by the buildings and other pedestrians. The overall mood of the video changes from the intimate and personal interaction in the first video to a more solitary and neutral scene in the second video.

Recap

LLaVA Main
Sep 2023

Visual Instruction Tuning.

CLIP vision encoder + Vicuna LLM (linear connector)

Pioneered the idea of a **multimodal assistant**

LLaVA-1.5 / 1.5-HD
Oct 2023

Improves robustness and accuracy.

2-layer MLP connector, AnyRes scaling, academic datasets.

Achieves stronger results on standard benchmarks, but still **single-image only**.

LLaVA-Next
Jan 2024

Expanded capabilities across vision tasks.

Boosts in reasoning, **OCR**, and world knowledge, h**igher resolutions** and varied **aspect ratios**, **interleaved** image–text **inputs** and video understanding

LLaVA-OV
Aug 2024

Unification and transfer learning across modalities.

Qwen-2 + SigLIP backbones, Higher AnyRes, larger VIT dataset

One model for **single-image, multi-image** and **video**.

Thank you!