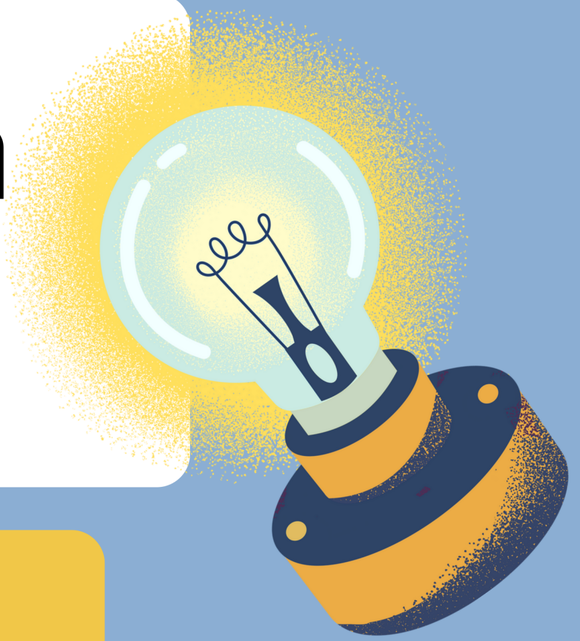
Visual Instruction Tuning

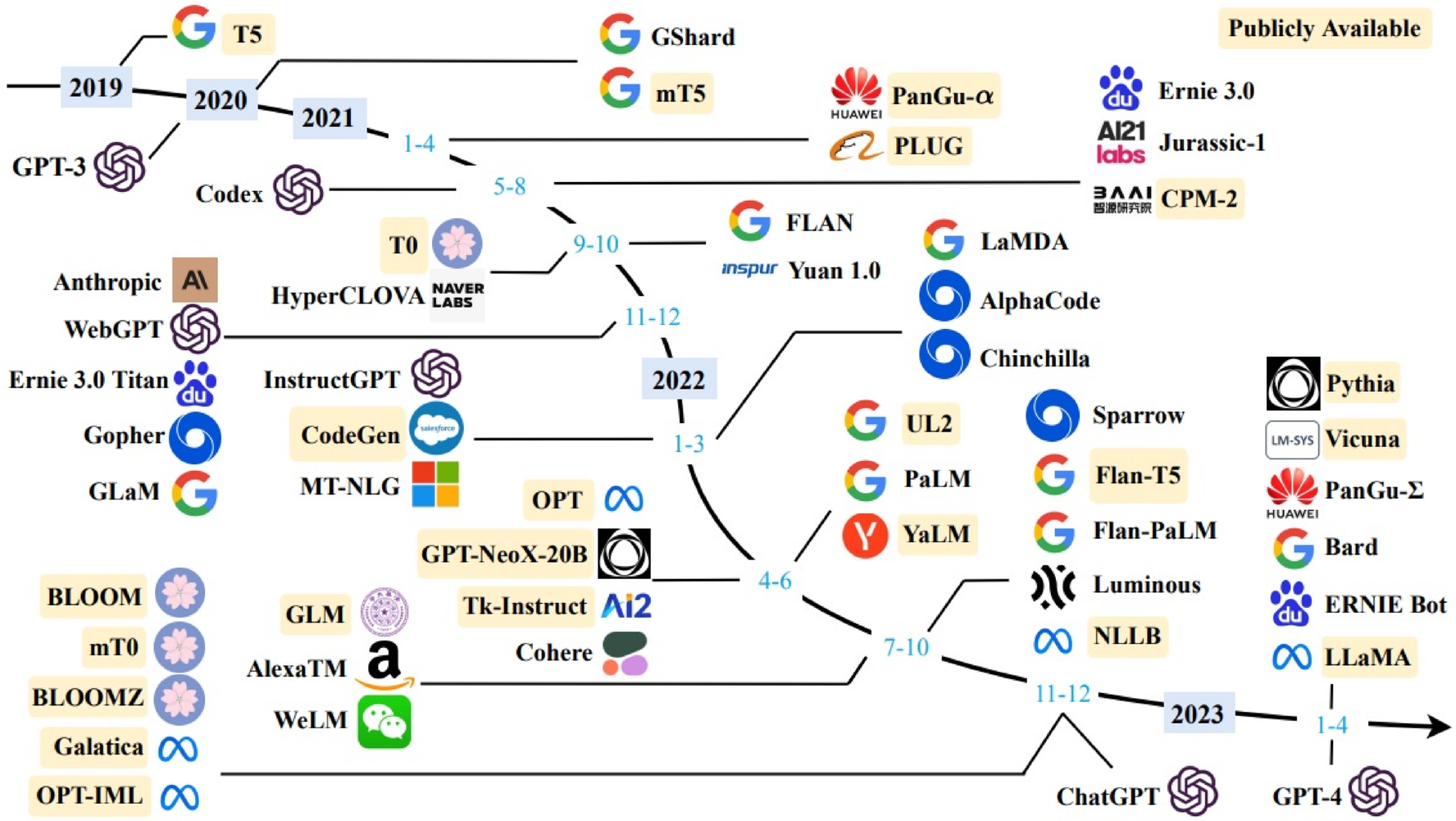


LLaVA - NeurIPS2023

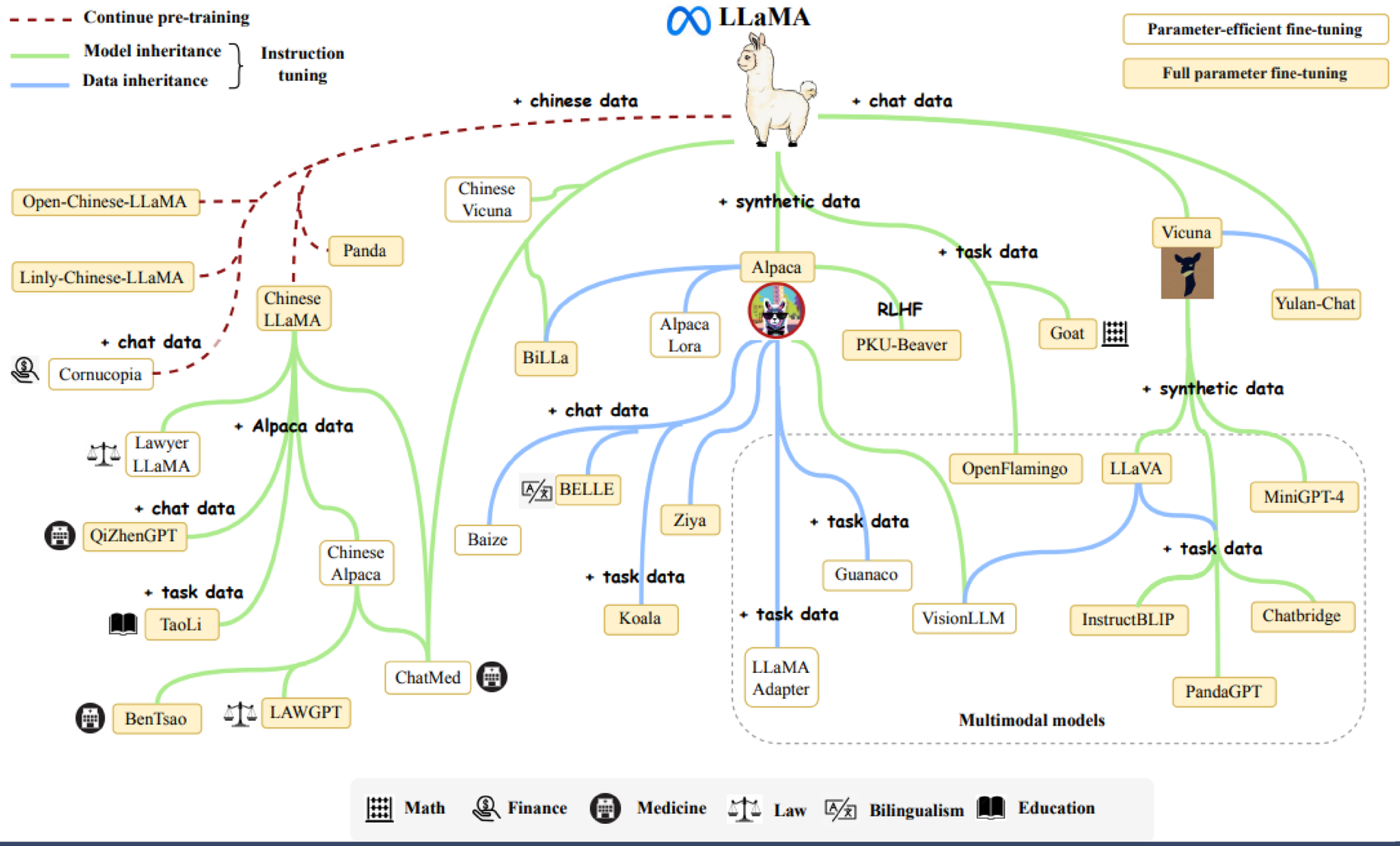
Today's Focus

- Motivation
- Instruction Tuning
- Visual Instruction Tuning
- Results

Timeline of LLM



Timeline of LLM

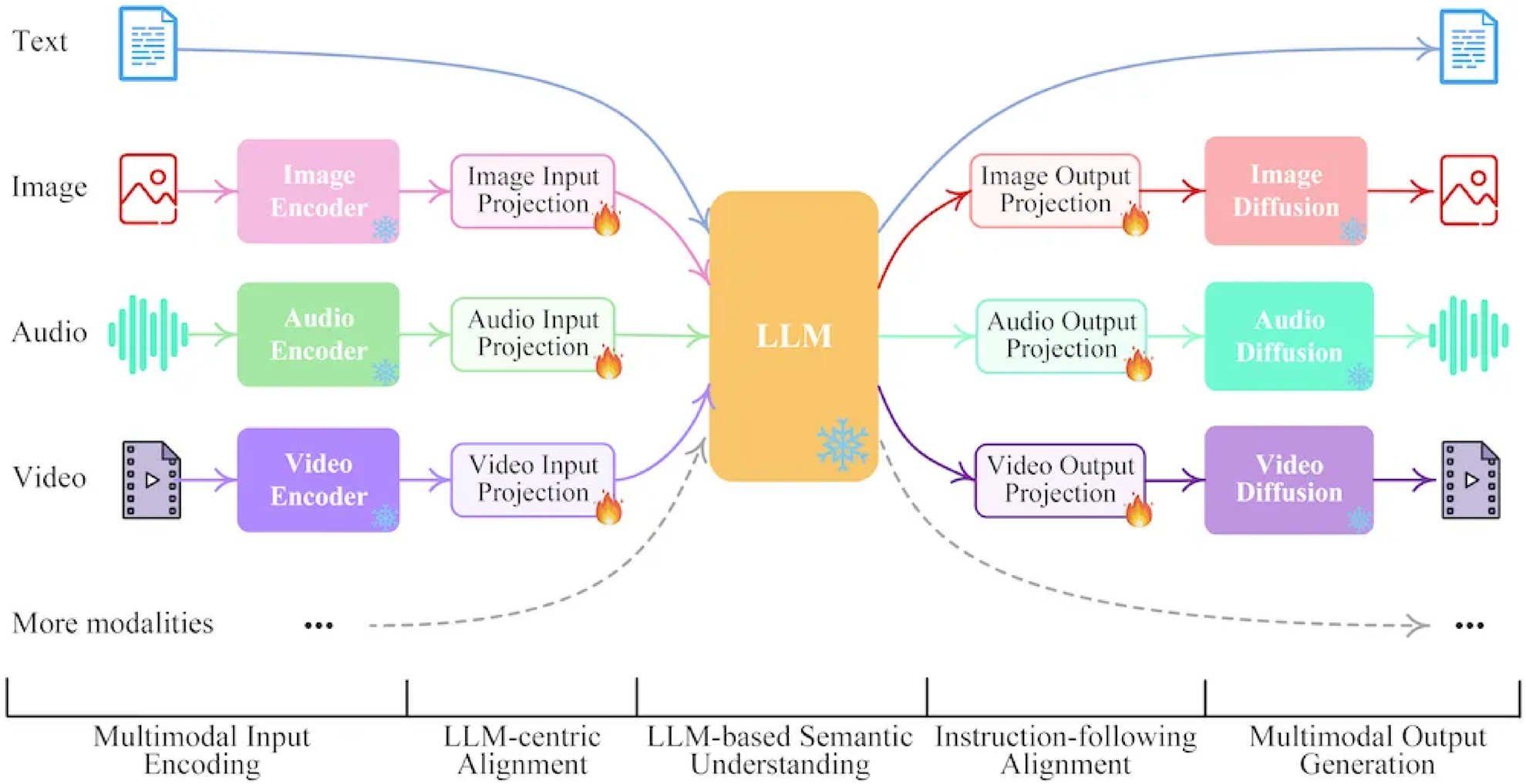


Price for training

Model	Parameters	Tokens to Train to Chinchilla Point (B)	Cerebras Model Studio CS-2 Day to Train	Cerebras Model Studio Price to Train
GPT3-XL	1.3	26	0.4	$2,500
GPT-J	6	120	8	$45,000
GPT-3 6.7B	6.7	134	11	$40,000
T-5 11B	11	34*	9	$60,000
GPT-3 13B	13	260	39	$150,000
GPT NeoX	20	400	47	$525,000
GPT 70B	70	1,400	85	$2,500,000
GPT 175B	175	3,500	Contact For Quote	Contact For Quote

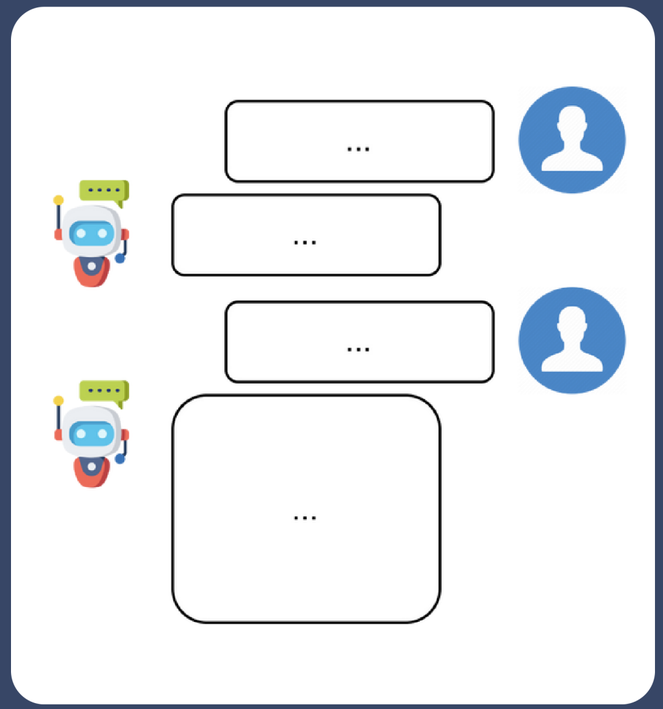
* - T5 tokens to train from the original T5 paper. Chinchilla scaling laws not applicable.

Multimodal LLM

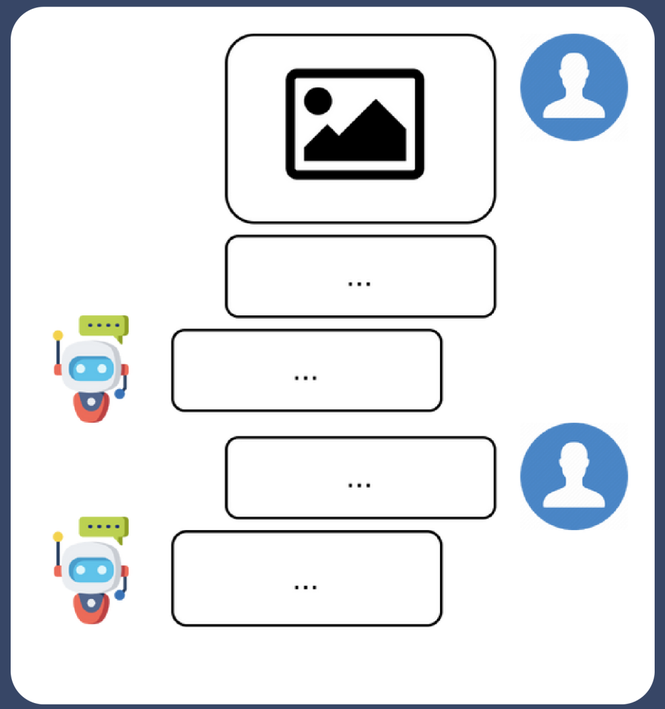


Motivation

How can we talk to a LLM about an image?



LLMs are great assistant



Extend to visual data?

Visual Instruction Tuning

But first, what is instruction tuning?

Instruction Tuning?

mistralai/**Mixtral-8x7B-Instruct-v0.1**

tiiuae/**falcon-7b-instruct**

venkycs/**phi-2-instruct**

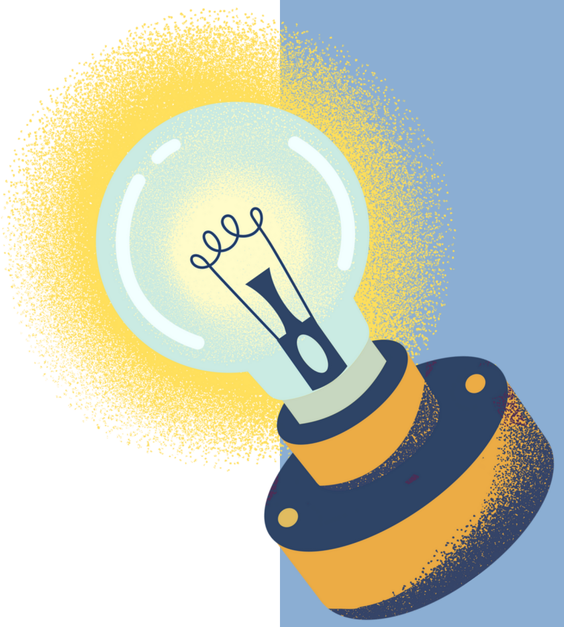
vinai/**PhoGPT-7B5-Instruct**

Instruction Tuning

- Fine-tuning with (Instruction, Output) pairs

Why instruction tuning?

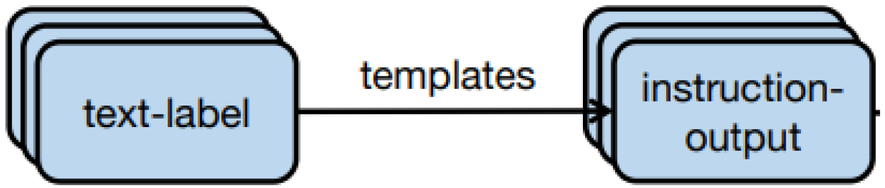
- Mismatch between training objective & user objective
- Capabilities and controllability
- Adapt to new domain without too much retraining.



Dataset Construction

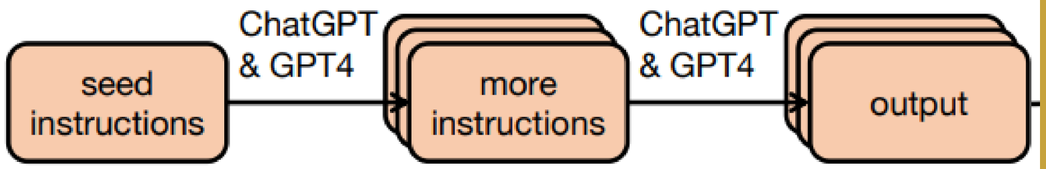
Modifying Existing Dataset

- FLAN
- P3
- xP3



Generating using LLM

- Employ LLMs such as GPT 3.5 turbo or GPT4.
- Self-Instruct
- InstructWild



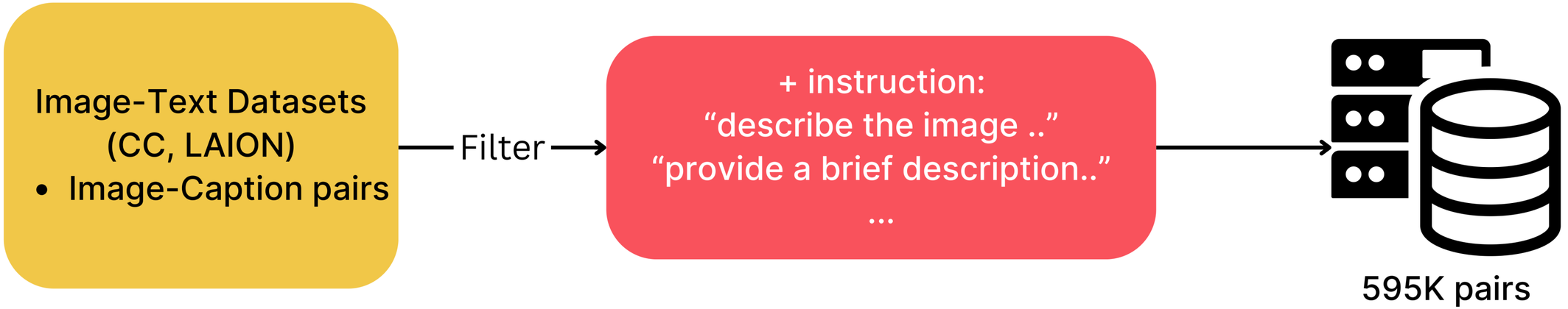
Challenges for visual instruction -following dataset

Limited data

Less well-defined

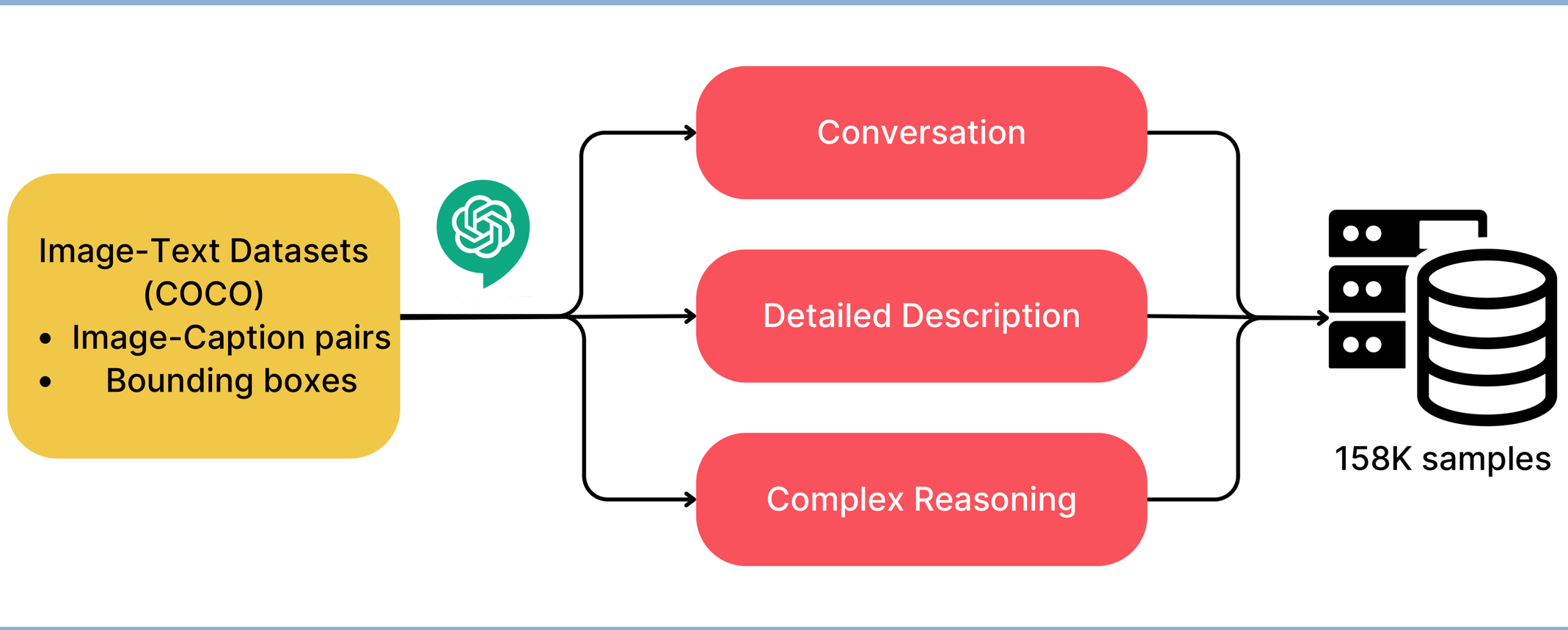
Generating Visual Instruction Dataset

The Naive way

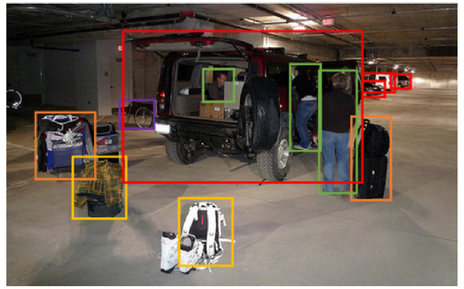


Generating Visual Instruction Dataset

GPT-assisted data generation



Example



A group of people standing ...

person: [0.6, 0.2, 0.1, 0.7]
backpack: [0.38, ...

ChatGPT

Few-shot prompting

Conversation:
Q: What type of vehicle is featured in the image
A: The image features a black sport ...

Detailed Description:
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people ...

Complex Reasoning:
Q: What challenges do these people faces?
A: .... fitting all the luggage, including....

Prompt

`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.`"""}`

Architecture

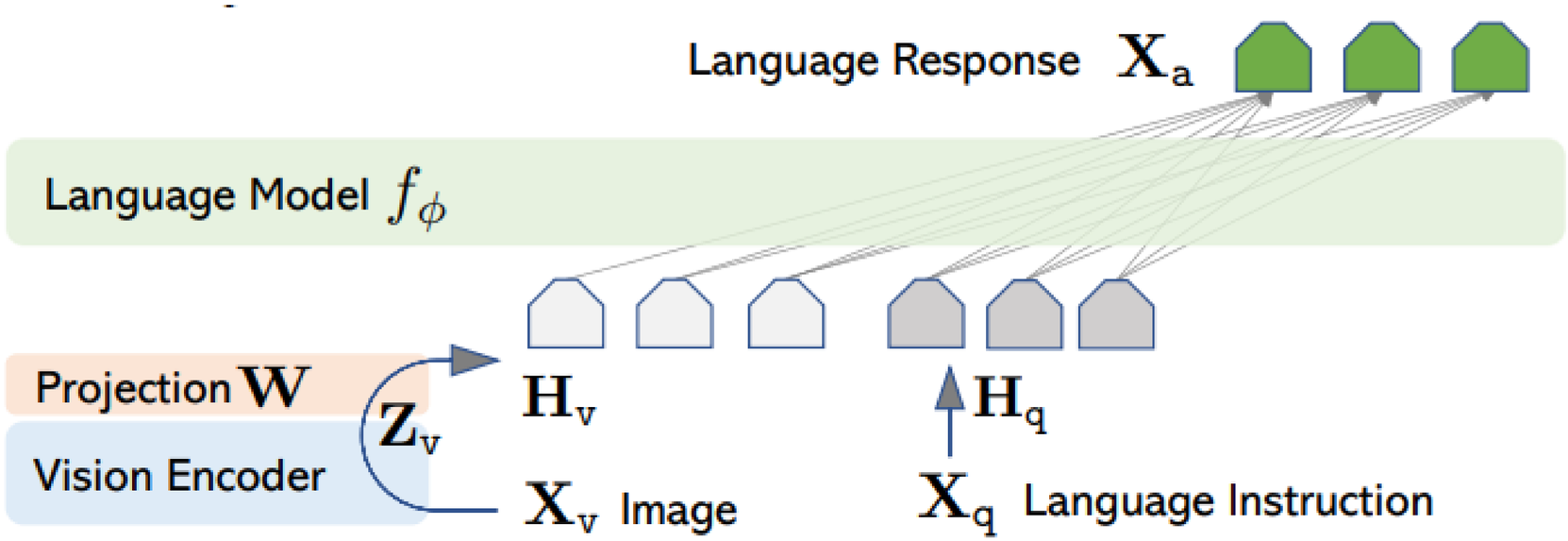
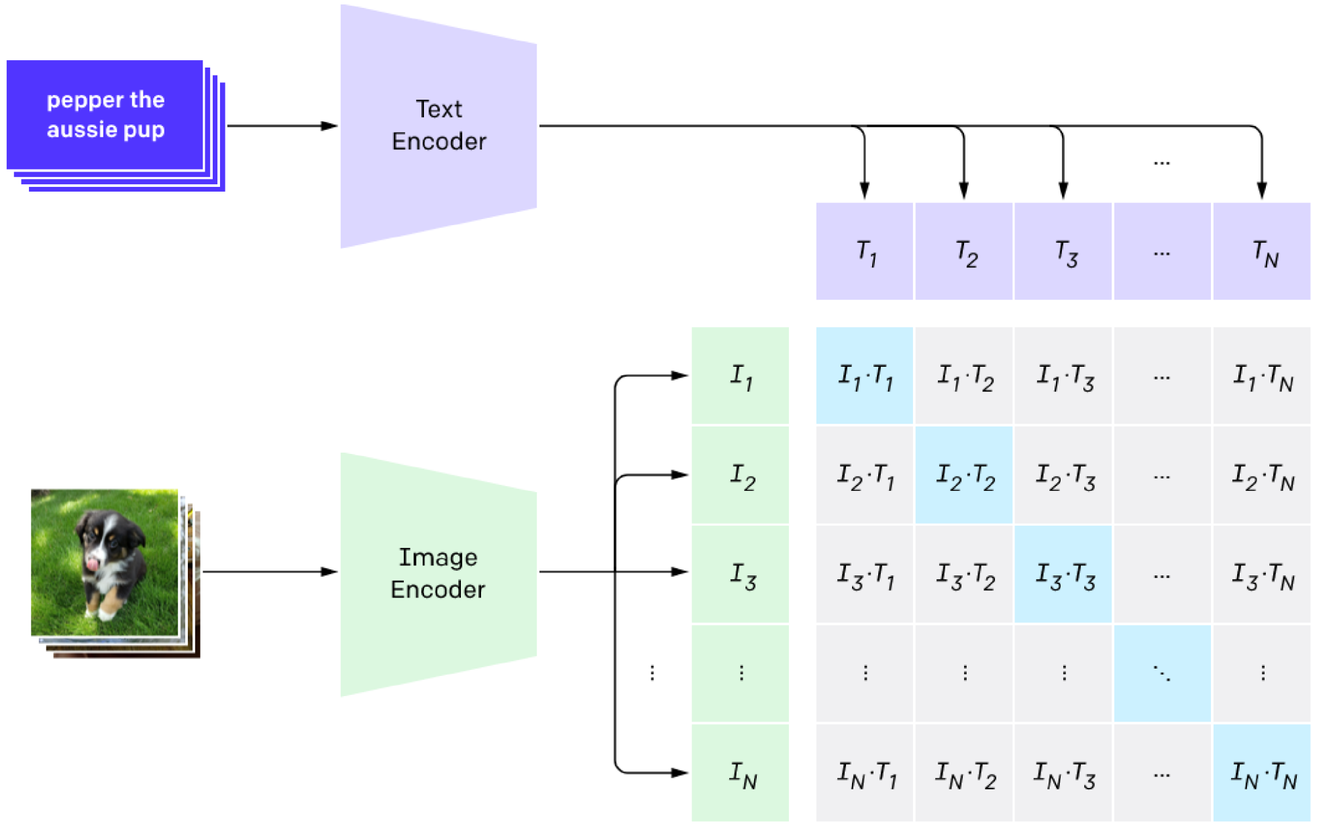


Figure 1: LLaVA network architecture.

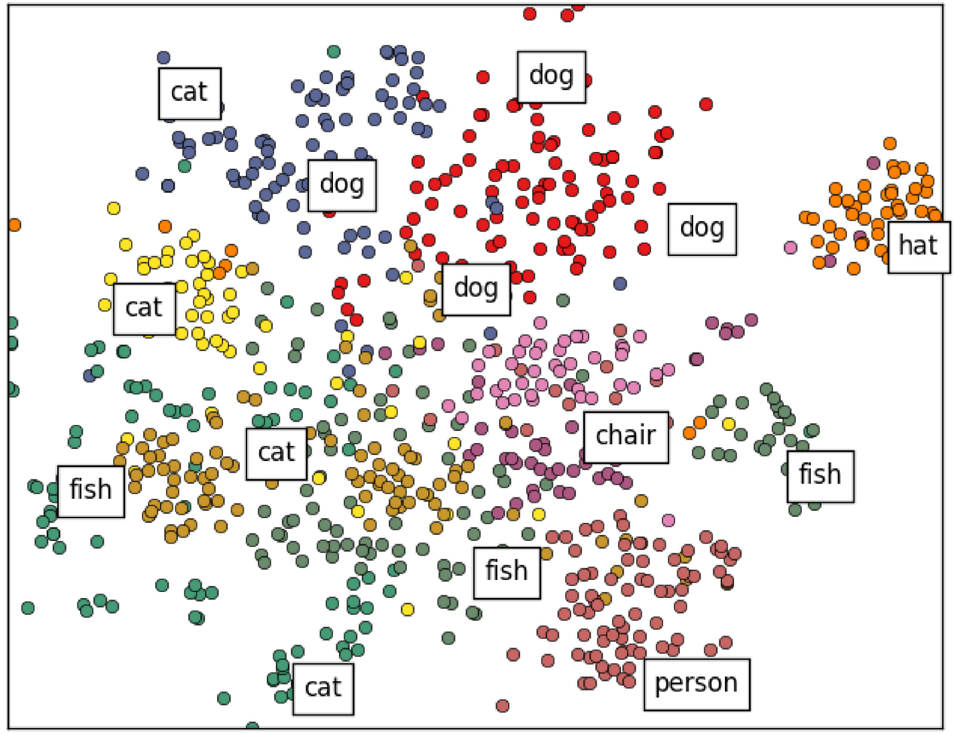
Vision Encoder



Why CLIP?

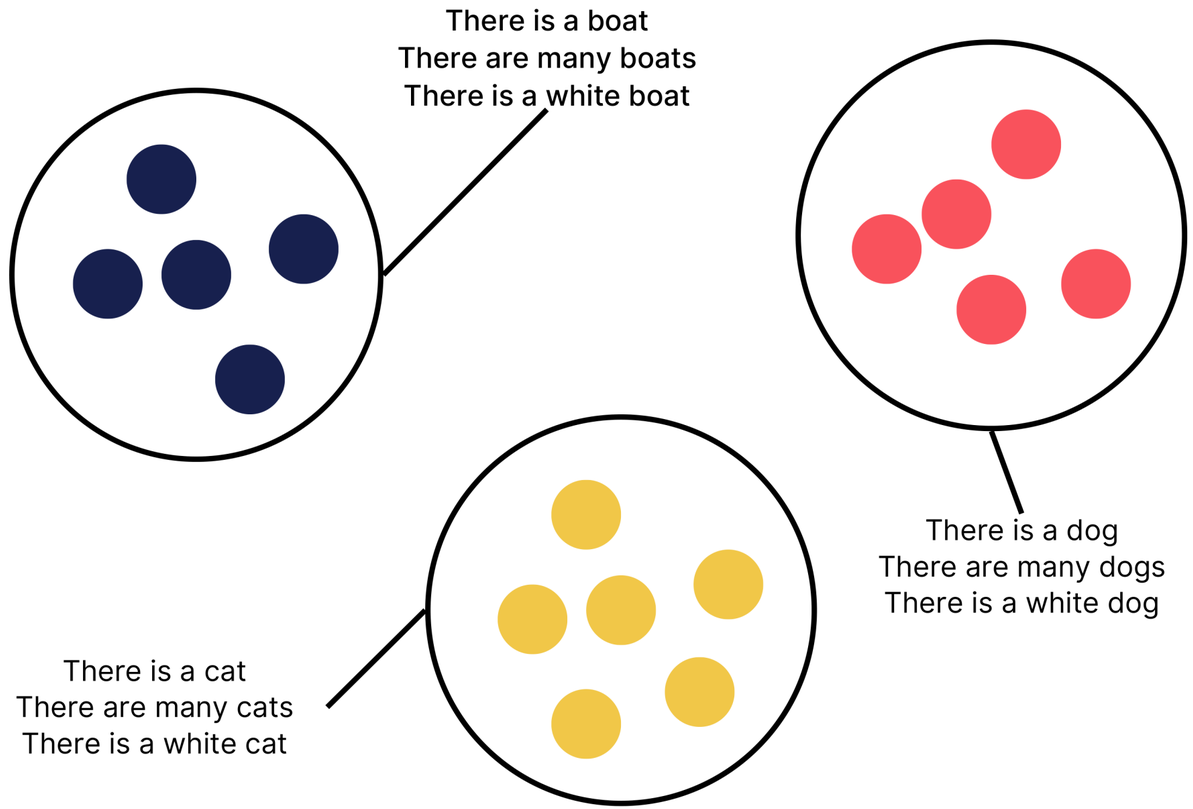
CAN WE USE ENCODER TRAINED ON IMAGENET?

These encoders are not trained to capture the complex relationship in text!



Why CLIP?

CLIP is trained to
align image and text



Architecture

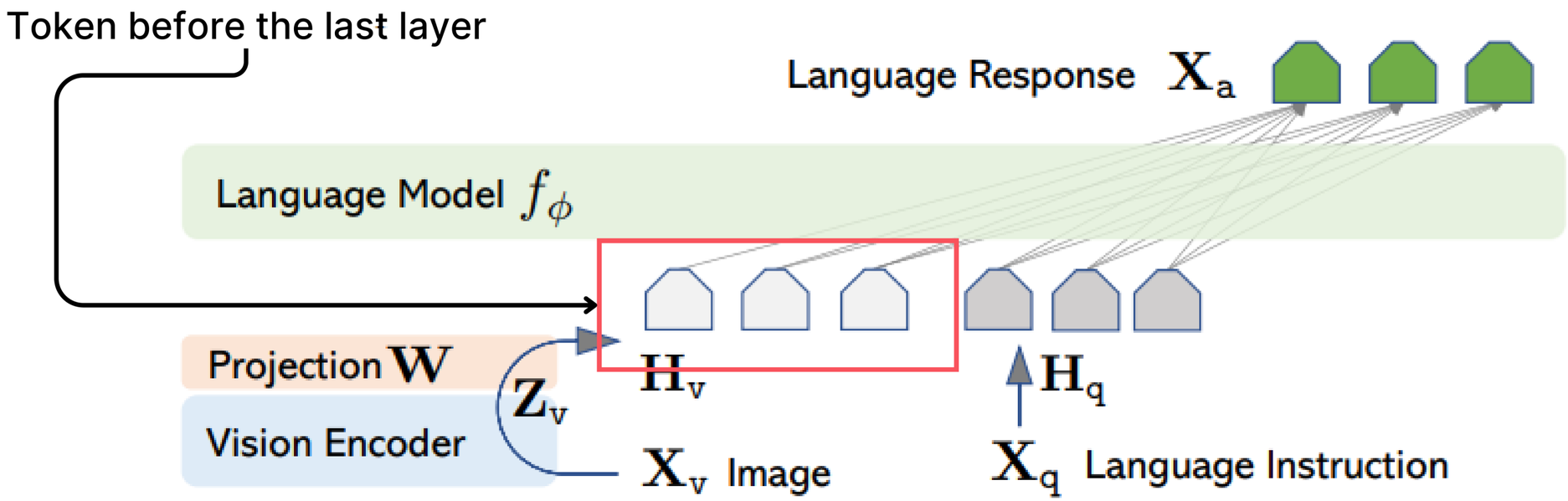
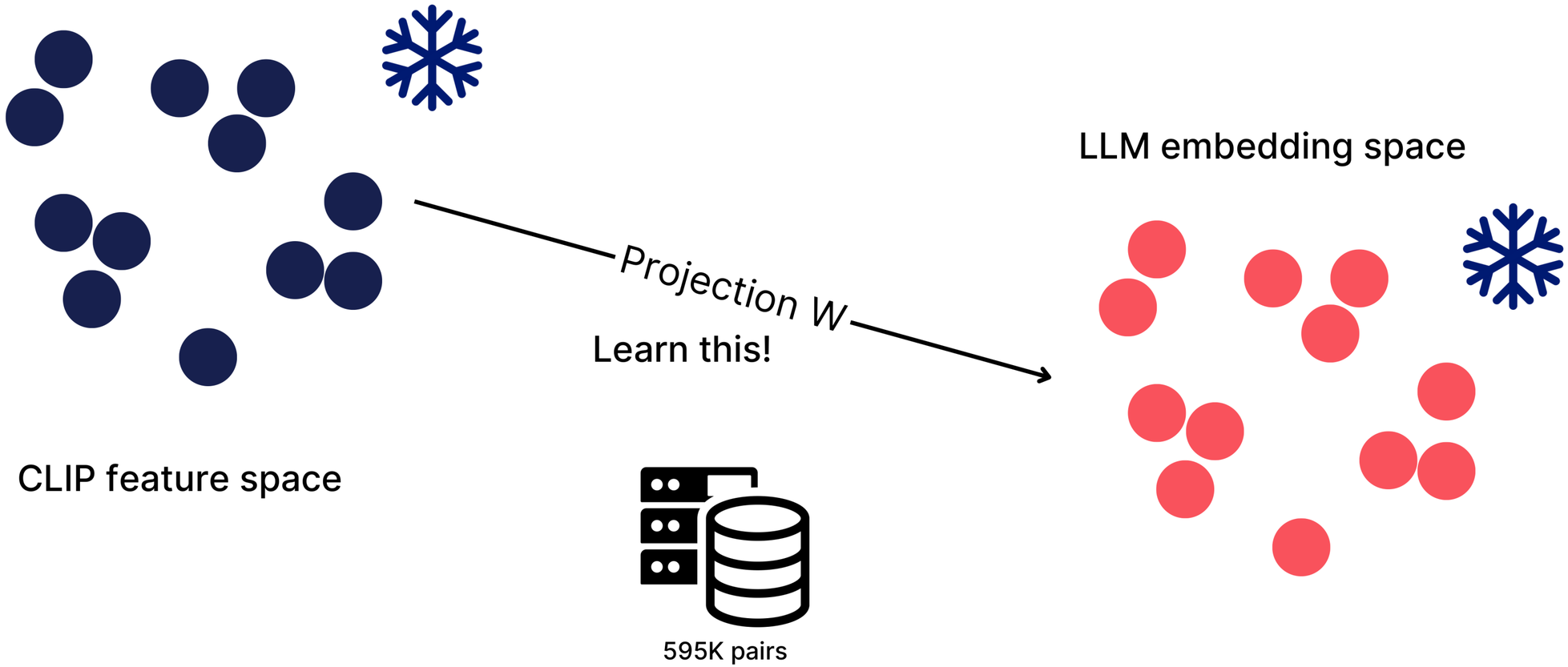


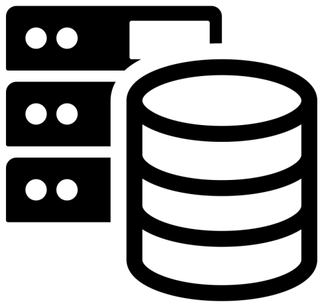
Figure 1: LLaVA network architecture.

Stage 1

Pre-training for Feature Alignment



Stage 2



158K samples Visual-Instruction following data
+
Science QA

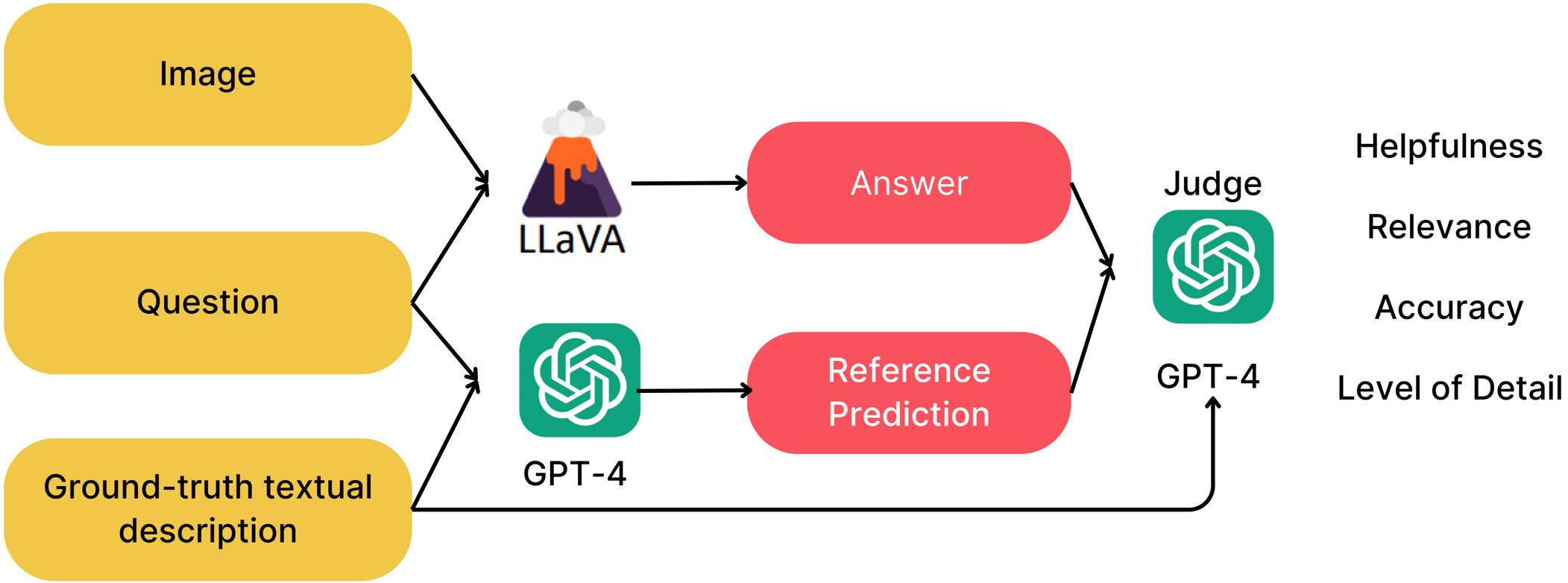
$$\mathbf{X}_{\texttt{system-message}} \texttt{<STOP>}$$
$$\texttt{Human} : \mathbf{X}^1_{\texttt{instruct}} \texttt{<STOP> Assistant:} \mathbf{X}^1_{\texttt{a}} \texttt{<STOP>}$$
$$\texttt{Human} : \mathbf{X}^2_{\texttt{instruct}} \texttt{<STOP> Assistant:} \mathbf{X}^2_{\texttt{a}} \texttt{<STOP>}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

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

Next-word prediction

How to evaluate this model?

- Leverage GPT-4 to measure the quality of generated responses.



Benchmarks

LLaVA-Bench (COCO)

- Randomly select 30 images from COCO-Val-2014
- Generate 3 types of questions. (Conversation, detailed description, complex reasoning)

LLaVA-Bench (In-the-wild)

- A diverse set of 24 images with 60 questions
- Indoor, outdoor scenes, paintings, sketches ...

Results

LLaVA-Bench (COCO)
LLaVA-Bench (In-the-Wild)

	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)

	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

Results

ScienceQA - Dataset

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**

Results

Ablations

No stage-1

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)	-

LLaVA-1.5

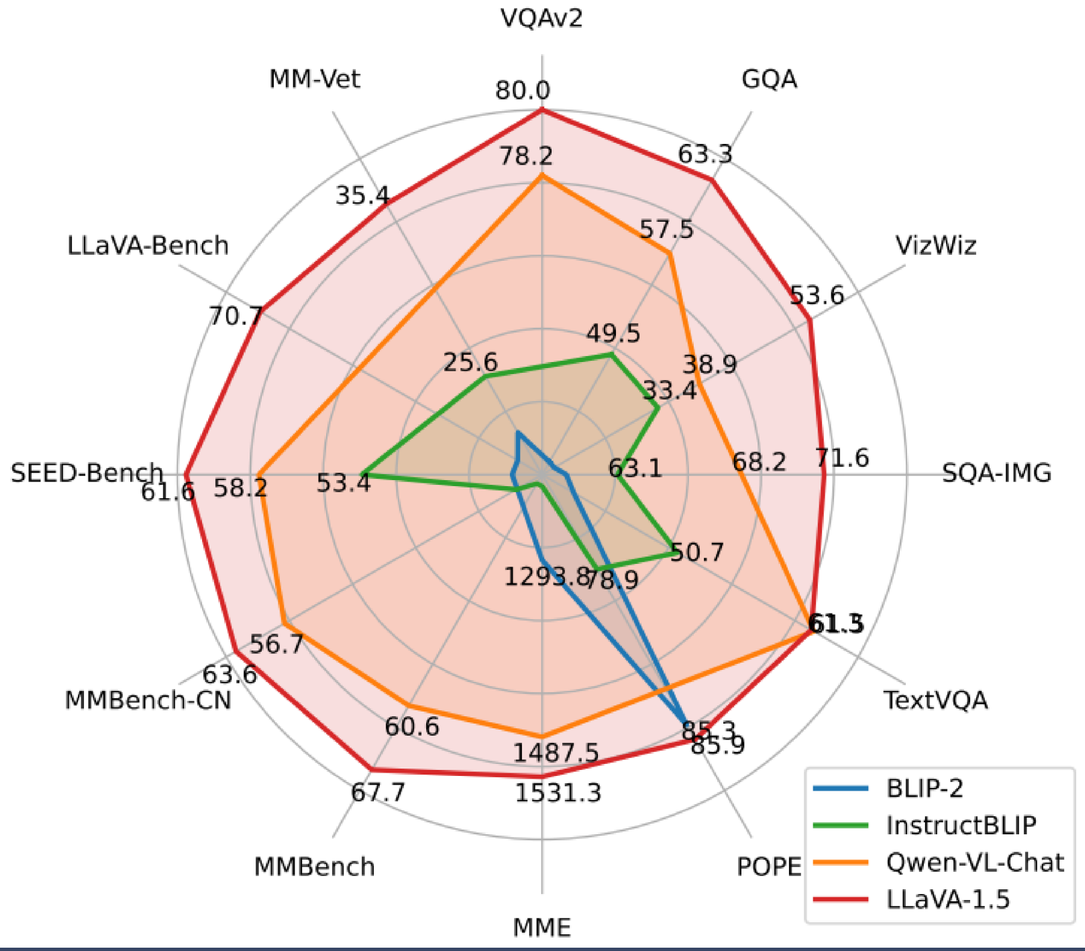
What is an electric circuit?

Scaling for more training data

Scale up image resolution

Vision language connector:
2 Layer MLP

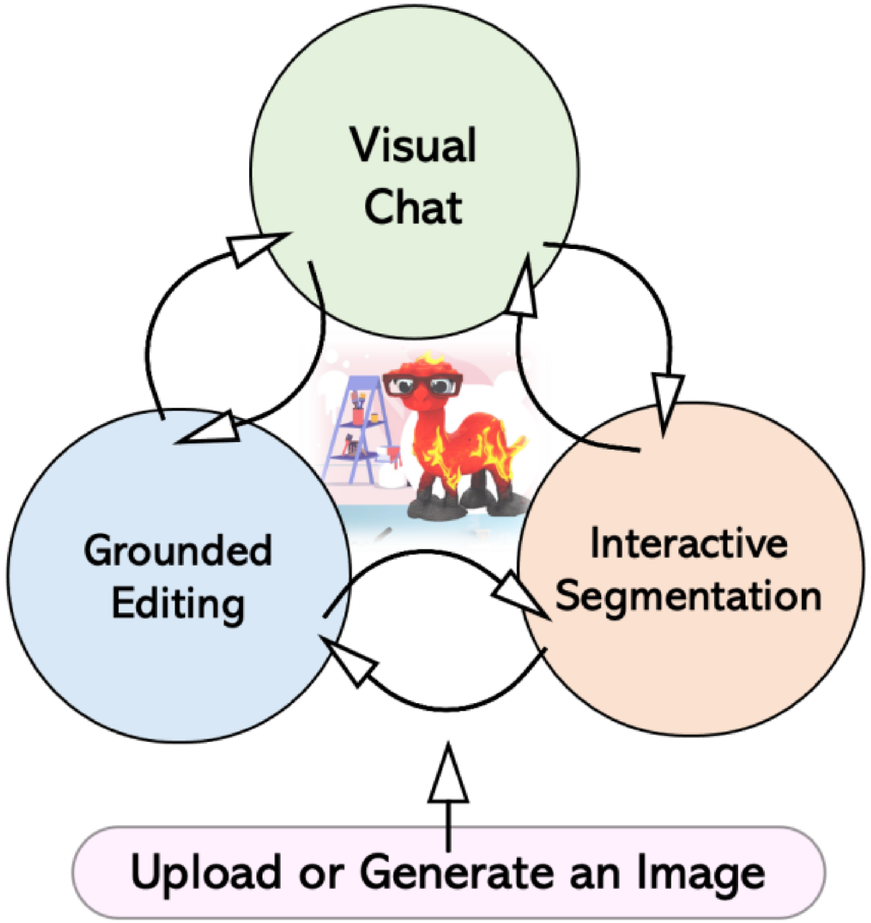
LLaVA-1.5


VQAv2
GQA
VizWiz
SQA-IMG
TextVQA
POPE
MME
MMBench
MMBench-CN
SEED-Bench
LLaVA-Bench
MM-Vet
80.0
78.2
63.3
57.5
53.6
38.9
71.6
68.2
61.3
85.9
85.3
1531.3
1487.5
67.7
60.6
63.6
56.7
61.6
58.2
70.7
35.4
25.6
49.5
33.4
53.4
63.1
50.7
1293.8
978.9
BLIP-2
InstructBLIP
Qwen-VL-Chat
LLaVA-1.5

LLaVA-Plus

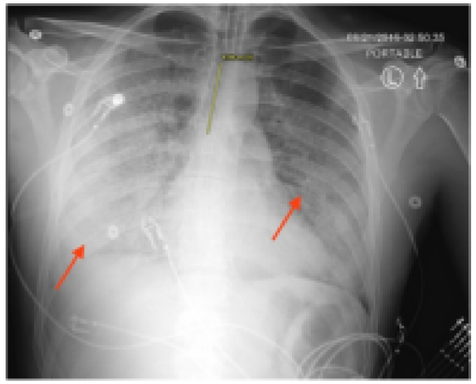

Visual Generation
Visual Interaction
Input Image
Conditional Gen.
Editing
Point to multi-level Seg
Box to Seg
Stroke to Seg
2024 16-MONTH CALENDAR
The image depicts a wooden pier extending into a large body of water, with a bench placed in the middle of the pier. The scene is set against a backdrop of trees and mountains, creating a tranquil and picturesque environment.
Lake
bridge
The text "2024 16-MONTH CALENDAR" is detected, suggesting that this calendar is a 16-month calendar for the year 2024.
The lake in question is Lost Lake, which is located in Whistler, British Columbia, Canada.
floor-wood
This image is paw-some! Who says cats can't be outdoorsy? 🌟🌲 #naturelover
LLaVA
Object Seg and Det
Semantic Seg
OCR
Search
Social Media Post
Visual Understanding
External Knowledge
Composition

LLaVA-Interactive



LLaVA - Med

Visual input example, Biomedical image from [6]



Caption: Chest x-ray (CXR) on Day 2 of admission post-intubation (yellow line showing the level of the endotracheal tube). Rapidly worsening ground-glass opacities are seen throughout the lungs with relative subpleural sparing (red arrows)

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7398722/

User	What is shown in this image?
LLaVA-Med	The image is a chest X-ray (CXR) that shows bilateral patchy infiltrates, which are areas of increased opacity in the lungs. These infiltrates can be indicative of various lung conditions, such as infections, inflammation, or other lung diseases.
User	Are there any devices implanted in this patient?
LLaVA-Med	Yes, it appears that the patient has several devices implanted. These devices may include endotracheal tubes, central venous catheters, and/or pulmonary artery catheters. These devices are typically used for monitoring and managing various medical conditions, as well as for providing support during surgical procedures.

DEMO