From Multimodal LLM to Human-level AI

Modality, *Instruction*, *Reasoning*, *Efficiency* and **Beyond**

https://mllm2024.github.io/CVPR2024/

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Hao Fei

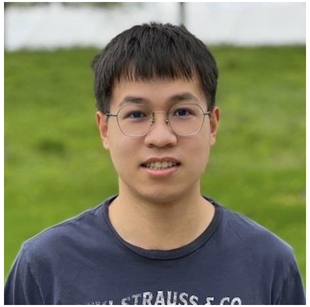
National University of Singapore

Yuan Yao

National University of Singapore

Ao Zhang

National University of Singapore

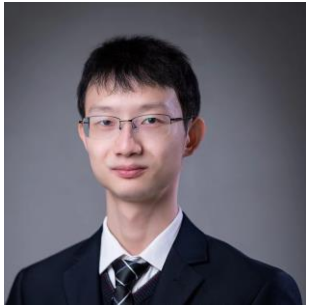


Haotian Liu

University of Wisconsin-Madison

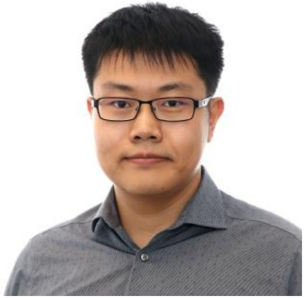
Fuxiao Liu

University of Maryland, College Park



Zhuosheng Zhang

Shanghai Jiao Tong University



Hanwang Zhang

Nanyang Technological University

Shuicheng Yan

Kunlun 2050 Research, Skywork AI

Part-IV

Multimodal Instruction Tuning

Haotian Liu

Ph.D.

University of Wisconsin, Madison

[*https://hliu.cc*](https://hliu.cc)

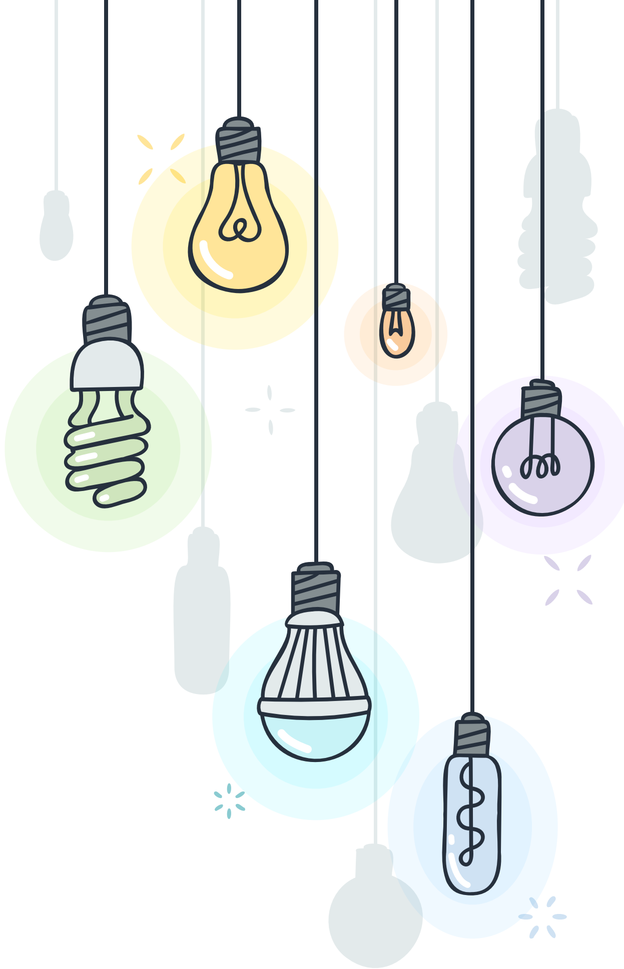
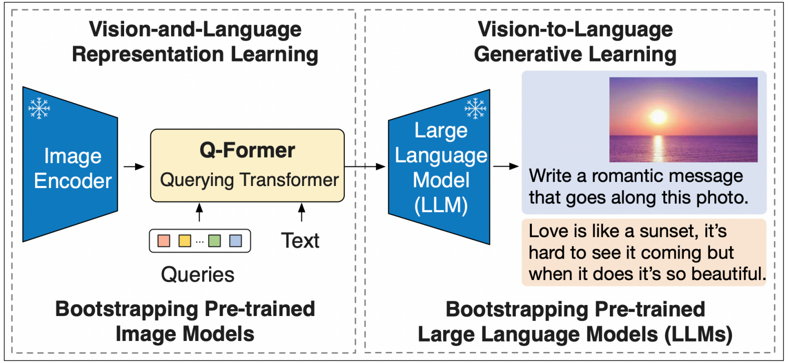
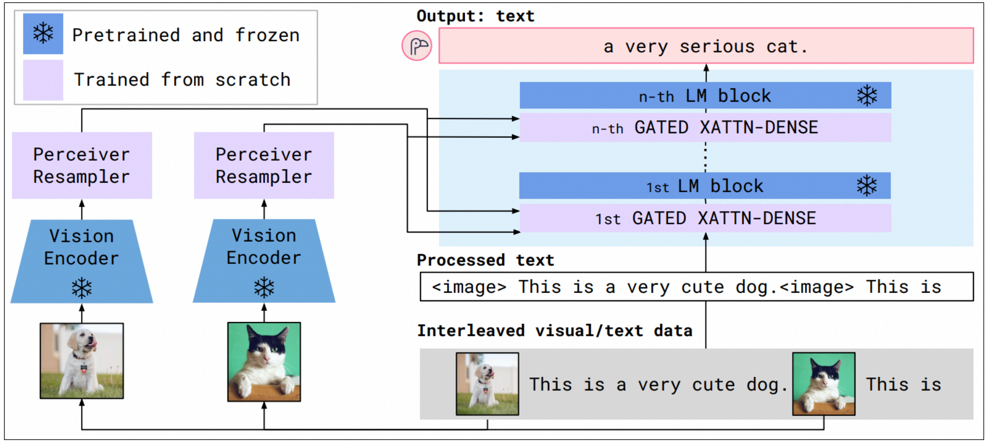


Table of Content

Why Multimodal Instruction Tuning?





Pretrained models aligns multiple modalities, can understand basic information from different modalities, and sometimes perform simple question-answering.

Cannot follow complex instructions, and often require task-specific fine-tuning for it to perform well on downstream tasks.

[Wang et al. 2022] GIT: A Generative Image-to-text Transformer for Vision and Language
[Li et al. 2023] Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models
[Alayrac et al. 2022] Flamingo: a visual language model for few-shot learning

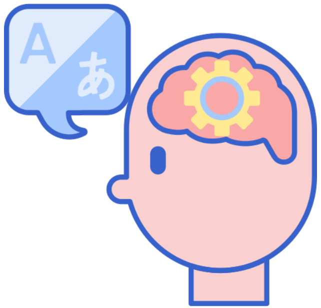
Why Multimodal Instruction Tuning?

- From Single-Purpose to General-Purpose

  ☞ Traditional vision models are task-specific, which requires training and using multiple models for different tasks and restrict the potential synergies from diverse tasks;

  ☞ These vision models typically have a pre-defined and fixed interface, leading to limited interactivity and adaptability in following users' task instructions.

  ☞ Multimodal Instruction Tuning allows multimodal models to generalize to unseen tasks by following new instructions, thus boosting zero-shot performance.



Instruction Tuning is NOT multitask learning

- Multitask learning (with task tokens)

Training

INPUT: <image><tok_task_1=short_cap>
OUTPUT: <generated short descriptions>

INPUT: <image><tok_task_2=yes_no>
OUTPUT: yes/no

Testing

Only with <tok_task_1>, <tok_task_2>…

Does not work with **<new_task=long_cap>**

- Instruction tuning (with natural language task instructions)

Training

INPUT: <image>Describe this image briefly.
OUTPUT: <generated short descriptions>

INPUT: <image>Is this xxx?
OUTPUT: yes/no

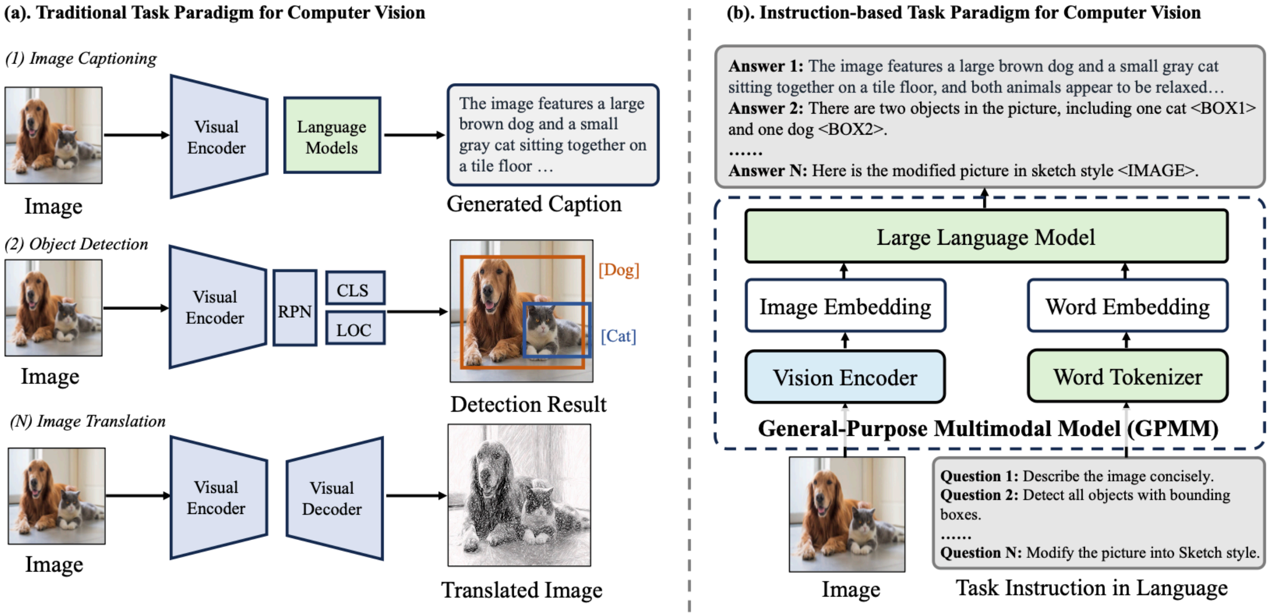
Testing

INPUT: <image>Describe this image in detail.
OUTPUT: <long descriptions>

Generalizes to new instructions zero-shot.

Why Multimodal Instruction Tuning?

- From Single-Purpose to General-Purpose



[1] Visual Instruction Tuning towards General-Purpose Multimodal Model: A Survey. 2023
[2] A Survey on Multimodal Large Language Models. 2024

1

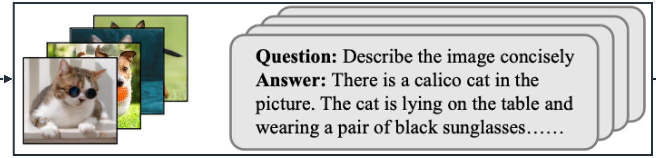
Multimodal Instruction Tuning Framework

MLLM Instruction Tuning Framework

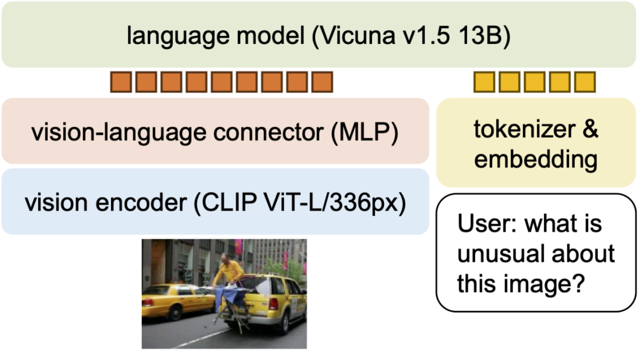
Data Construction

- Data Adaptation
- Self-Instruction
- Data Mixture

Visual Instruction-following data



Visual Instruction Tuning Framework
Example: LLaVA-1.5



Popular MLLMs: *LLaVA*, *MiniGPT4*, *LLaVA-NeXT*, *ViP-LLaVA*, *LLaVA-UHD*, *MiniCPM*, *Qwen-VL*, *CogAgent*, *InternVL*, *mPLUG-OWL*, Monkey, MiniGemini, LLaVA-HR, SPHINX, DeepSeek-VL, MoAI

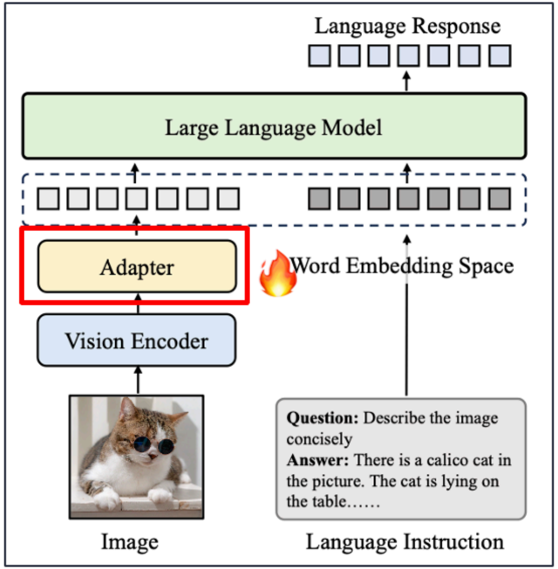
Training Paradigms

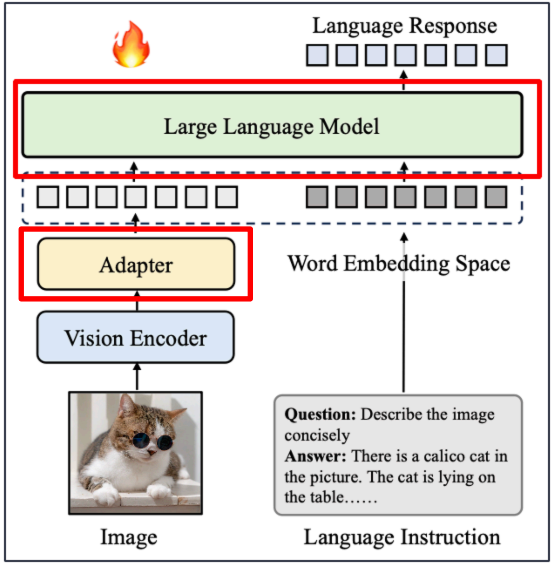
Stage1: Pretraining Stage

- Align different modalities, provide world knowledge

Stage2: Instruction Tuning Stage

- Teach models to better understand the instructions from users and fulfill the demanded tasks.

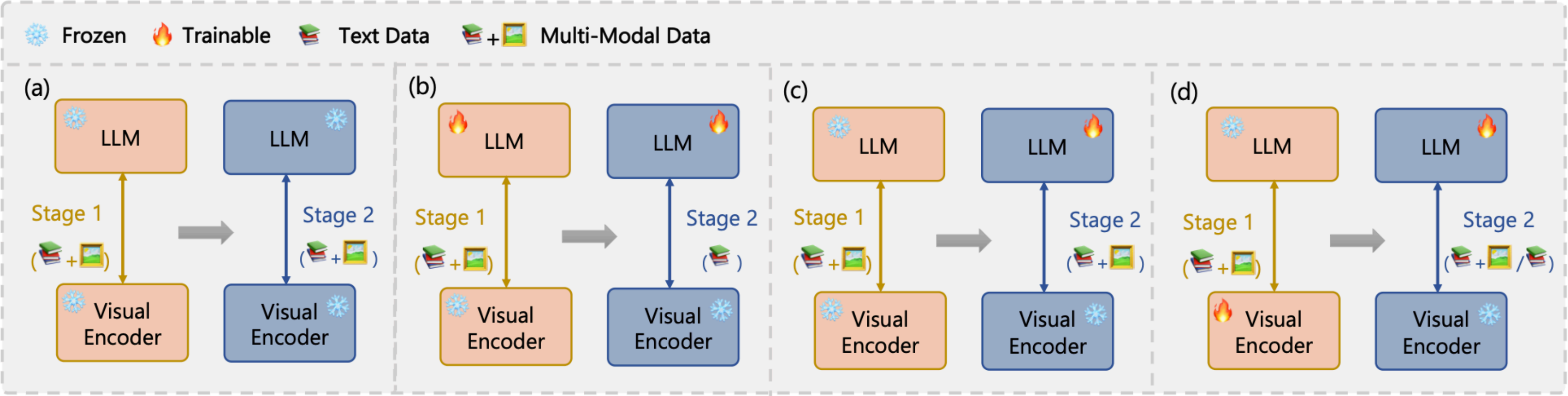




[1] MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. NAACL 2024.
[2] Visual Instruction Tuning. NeurIPS 2023.

Training Paradigms

☞ Training paradigms of popular multimodal large language models.



MiniGPT4 Kosmos-1 LLaVA mPLUG-Owl

[1] mPLUG-Owl: Language Models with Multimodality. 2023.
[2] Visual Instruction Tuning. NeurIPS 2023.
[3] MINIGPT-4: ENHANCING VISION-LANGUAGE UNDERSTANDING WITH ADVANCED LARGE LANGUAGE MODELS. 2023.
[4] Language Is Not All You Need: Aligning Perception with Language Models. 2023.

1.5 Another Perspective of Multimodal Instruction Tuning

How can we create such multimodal models that follow human's intent?

How can we create such multimodal models that follow human's intent efficiently?

How can we create such multimodal models that follow human's intent efficiently?

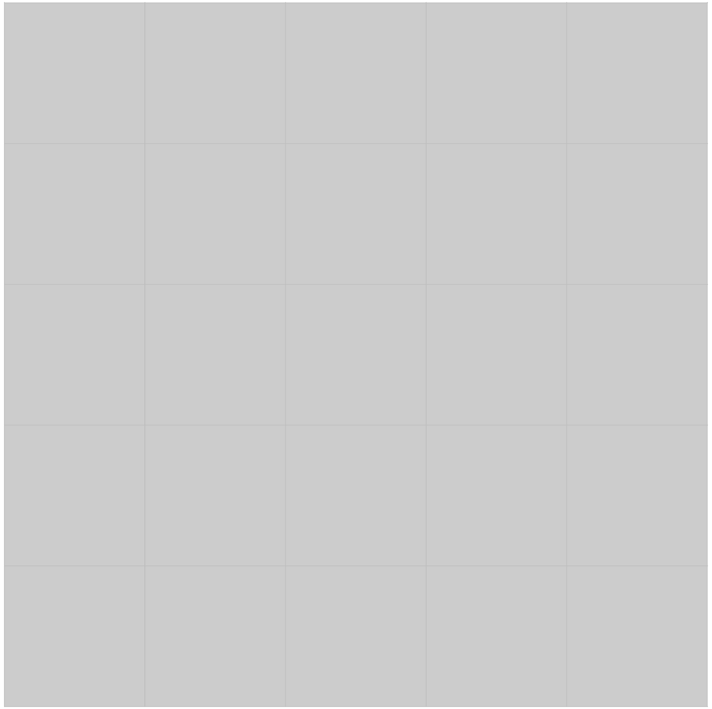
How can we create such multimodal models that follow human’s intent efficiently?

How can we make an instruction-following LLM multimodal efficiently?

LLM “learns” a foreign language efficiently.

- LLaMA is almost trained on English tokens solely.
- LLaMA learns foreign languages with 52K conversations
    - E.g. Chinese, Japanese, etc.
    - ~1 hour training

Multimodal learning as a translation problem



Multimodal learning as a translation problem

		llama lava		
	llama lava glasses	llama lava glasses	Glasses	
		llama lava	llama lava	
		llama lava	llama lava	llama lava
		llama feet	llama feet	

Q: What’s in the image?
A: A llama that’s made of lava.

Q: What’s special of this image?
A: The llama is wearing glasses.

Multimodal learning as a translation problem

Output

The color of the shirt the man’s wearing is yellow.

Instruction

What’s the color of the shirt that the man is wearing?

Image

Language Decoder

Cross-modal Connector

Visual Encoder

Multimodal Tokenizer
(i.e. translator)

LLM “learns” a visual foreign language efficiently.

Some questions are still hard

		llama lava		
	llama lava glasses	**llama lava glasses**	**Glasses**	
		llama lava	**llama lava**	
		llama lava	**llama lava**	**llama lava**
		llama feet	**llama feet**	

Q: Is the llama facing left or right?

A: Hmm…

Still struggles to follow complex visual instructions

Output

The unusual aspect of this image is …

Instruction

What is **unusual** about this image?

Language Decoder

Cross-modal Connector

Image

Visual Encoder

What do we need?

Tuning the model for following multimodal instructions

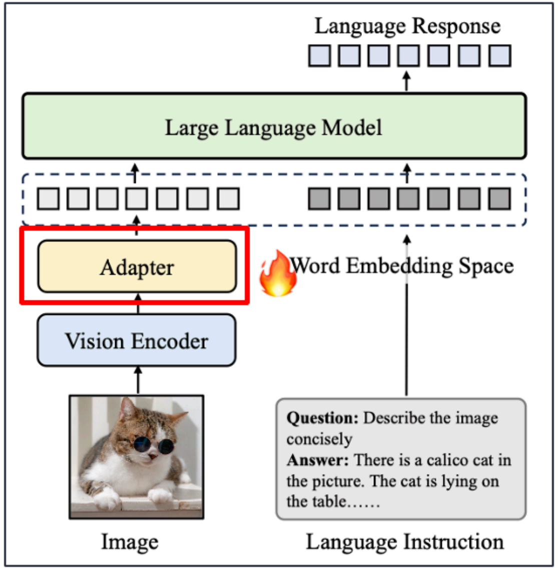
2

Multimodal Instruction Tuning Data Generation

Pretraining Data

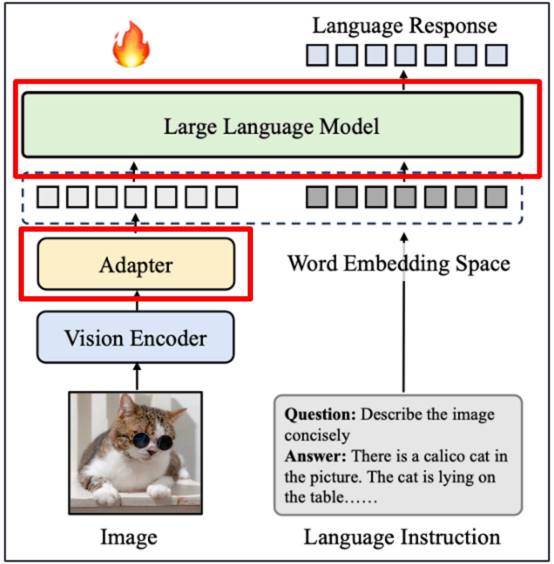
Stage1: Pretraining Stage

- Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

- Teach models to better understand the instructions from users and fulfill the demanded tasks.



[1] MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. NAACL 2024.
[2] Visual Instruction Tuning. NeurIPS 2023.

Pretraining Data (Paired)

- **Coarse-gained Image-text**

  Data volume is large, the captions are shorts and noisy.

- **Fine-gained Image-Text**

  High quality, longer and more accurate descriptions, fine-gained alignment between different modalities.

Dataset	Samples	Taxonomies
CC-3M	3.3M	Coarse-grained Image-Text
CC-12M	12.4M	Coarse-grained Image-Text
LAION-5B	5.9B	Coarse-grained Image-Text
SBU-Captions	1M	Coarse-grained Image-Text
ShareGPT4V-PT	1.2M	Fine-grained Image-Text
LVIS-Instruct4V	111k	Fine-grained Image-Text
ALLaVA	709k	Fine-grained Image-Text
MSR-VTT	200k	Video-Text
WavCaps	24k	Audio-Text
MMC-Instruction	600k	Chart-Text

[1] A Survey on Multimodal Large Language Models. 2024

Pretraining Data

ShareGPT4V

COCO: Young children standing on a platform waiting for a train to arrive.
Adults and children watching a train slowly leave.
A family near a railroad track watching the train pass.
People waiting on a platform as a train pulls up.
A train station with a green chain on the tracks and children waiting for it to go by.

Coarse-gained Image-text

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure.

There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, suggesting that someone may have set it down while waiting.

Coarse-gained Image-text

ShareGPT4V: The image captures a moment at a train station. A green train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.

A red signal stands on the tracks, its vibrant color contrasting with the green of the train. Signs are prominently displayed around the platform. One warns "Beware of Trains", a cautionary reminder of the powerful machines that frequent these tracks. Another sign advises "Please Use The Footbridge To Cross The Line. If You Require Assistance Please Ask The Station Staff", guiding passengers to safely navigate the station.

The sky above is a clear blue, providing a serene backdrop to the bustling activity of the station. In the distance, trees can be seen, adding a touch of nature to this man-made setting. The image is a snapshot of everyday life at a train station, capturing both its routine operations and its inherent charm.

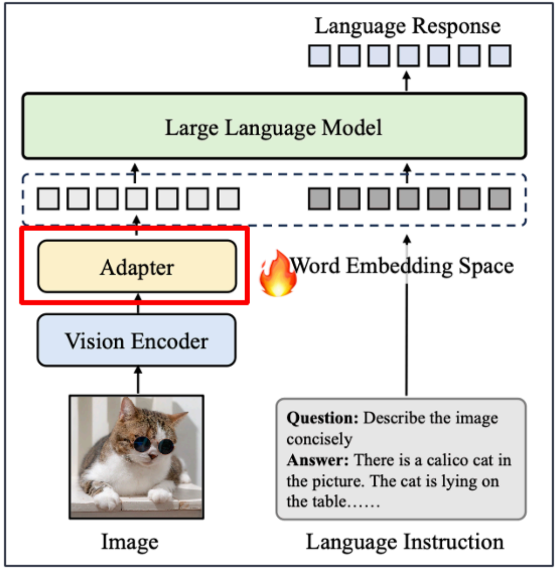
Fine-gained Image-text

[1] ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. 2023.

Instruction Data Generation

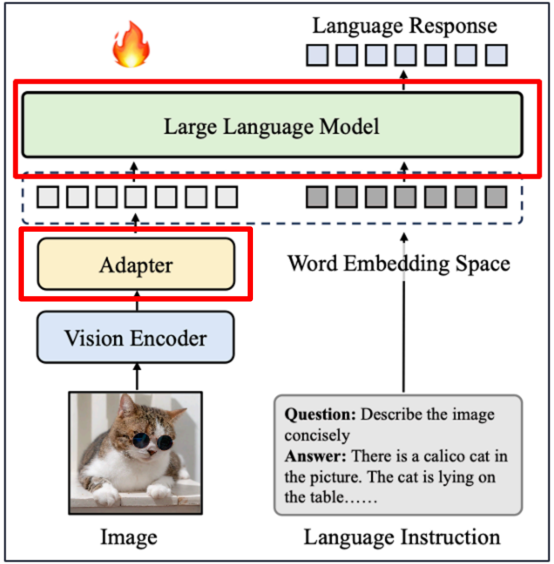
Stage1: Pretraining Stage

- Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

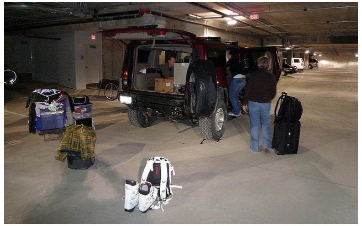
- Teach models to better understand the instructions from users and fulfill the demanded tasks.



[1] MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. NAACL 2024.
[2] Visual Instruction Tuning. NeurIPS 2023.

Instruction Data Generation

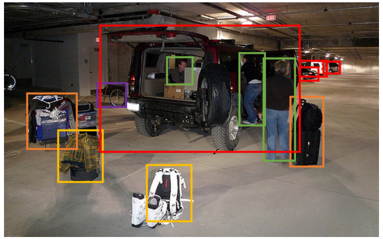
Image



Context (caption)

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

Context (bbox)



→

person: [0.68, 0.24, 0.77, 0.69], person: [0.63, 0.22, 0.68, 0.51], person: [0.44, 0.23, 0.48, 0.34], backpack: [0.38, 0.69, 0.48, 0.91], ….

[1] Visual Instruction Tuning. NeurIPS 2023.

Instruction Data Generation

❖ **Self Instruction**

First, Translate images into dense captions and bounding boxes. Second, prompt text-only GPT-4.

Prompt:
Give an image with following information: bounding box, positions that are the object left-top corner coordinates(X, Y), object sizes(Width, Height). Highly overlapping bounding boxes may refer to the same object.

Bounding boxes, dense Captions →

bounding box:
elephant heard on rocks X: 73 Y: 80 Width: 418 Height: 418
woman wearing long dress X: 176 Y: 298 Width: 35 Height: 83
group of green chairs X: 153 Y: 326 Width: 95 Height: 126
an orange bucket on the ground X: 91 Y: 341 Width: 38 Height: 36
a group of white umbrellas X: 99 Y: 82 Width: 112 Height: 28
a man in an orange shirt X: 204 Y: 265 Width: 31 Height: 47
a woman wearing a yellow dress X: 169 Y: 298 Width: 47 Height: 76
. . .

Task Descriptions →

Task: *image captioning, Image Sentiment Analysis, Image Quality Assessment, Object Interaction Analysis, Object Attribute Detection, Muli-choice VQA ...*

Generation Requirement →

Come up with 20 diverse instructions for all the tasks above with different language styles and accurate answers. The instructions should contain interrogative sentence and declarative sentences. The answers should be less than 30 words. Each task should have less than 3 instructions.

GPT4 OUTPUT Example:

Output from GPT4 →

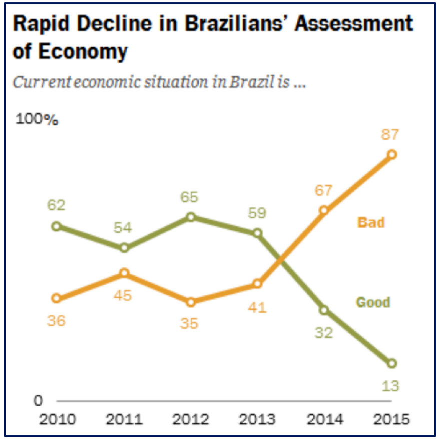
Instruction:
Craft a brief narrative about the baby elephant and adult elephant.
Answer:
A baby elephant is depicted behind an adult elephant, possibly seeking protection.

[1] Aligning large multi-modal model with robust instruction tuning. ICLR 2024.

Instruction Data Generation

- **Existing Data**

  The answers of existing VQA and caption datasets are usually concise, directly using these datasets for instruction tuning may limit the output length of MLLMs.



Question:

Which year has the most divergent opinions about Brazil's economy?

Answer:

2015

[1] ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. ACL 2022 Findings.

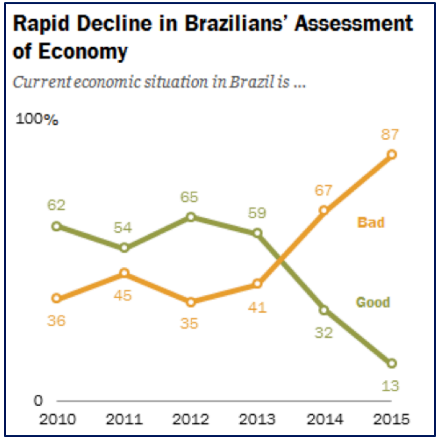
Instruction Data Generation

- **Existing Data**

  The answers of existing VQA and caption datasets are usually concise, directly using these datasets for instruction tuning may limit the output length of MLLMs.

- **Data Adaptation**

  Declares short and brief for short-answer data in the instruction.



Question:

Which year has the most divergent opinions about Brazil's economy? A short answer to the question is:

Answer:

2015

[1] ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. ACL 2022 Findings.

Instruction Data Generation

User

Can you tell me what I can cook with these?

With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, …

before

Directly training with VQA-v2 data (short answer) with our conversation data. The model refuses to provide natural answers.

[1] Improved Baselines with Visual Instruction Tuning. CVPR 2024.

Instruction Data Generation

Visual input example, Different Format Prompts:	
	
Normal prompt	What is the color of the shirt that the man is wearing?
Response	The man is wearing a yellow shirt.
Ambiguous prompt	Q: What is the color of the shirt that the man is wearing? A:
Response	The man is wearing a yellow shirt.

→ Yellow

→ Yellow

[1] *Improved Baselines with Visual Instruction Tuning. CVPR 2024.*

Instruction Data Generation

Visual input example, Different Format Prompts:	
	
Normal prompt	What is the color of the shirt that the man is wearing?
Response	The man is wearing a yellow shirt.
Ambiguous prompt	Q: What is the color of the shirt that the man is wearing? A:
Response	The man is wearing a yellow shirt.
Formatting prompt	What is the color of the shirt that the man is wearing? **Answer the question using a single word or phrase.**
Response	Yellow.

[1] Improved Baselines with Visual Instruction Tuning. CVPR 2024.

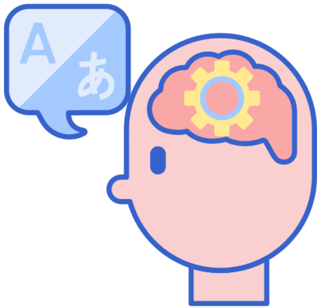
Existing Instruction Tuning Dataset

Dataset	Size	Modalities	Constructions
LLaVA-Instruct-158k	158k	Image, Text	ChatGPT-generated
LRV-Instruction	400k	Image, Text	GPT4-generated
MMC-Instruction	600k	Chart, Text	GPT4-generated/adapted
Clotho-Detail	3.9k	Text, Audio	GPT4-generated
MACAW-LLM	119k	Image, Video, Text	GPT-3.5-turbo-generated
MIMIC-IT	2.8M	Image, Video, Text	ChatGPT-generated
StableLLaVA	126k	Image, Text	StableDiffusion & ChatGPT-generated
LAMM	196k	Image, PointCloud, Text	GPT4-generated
VIGC-LLaVA	1.8M	Image, Text	Model-generated
X-LLM	10k	Image, Video, Text	ChatGPT-generated

Summary

- How we teach multimodal models:

  - Pretraining:
    A dictionary to teach LLM to understand (vocabularies from) a new modality
  - Instruction tuning (short answer VQA):
    Small puzzles to effectively/efficiently injects new domain knowledge
  - Instruction tuning (natural conversation VQA):
    Real-world applications to practice the skills



Any questions?