From Multimodal LLM to Human-level AI

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

https://mllm2024.github.io/COLING2024

LREC-COLING 2024

NUS National University of Singapore

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

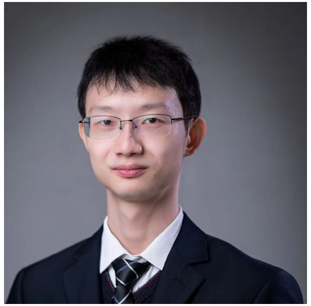
UNIVERSITY OF MARYLAND

Hao Fei

National University of Singapore

Yuan Yao

National University of Singapore



Zhuosheng Zhang

Shanghai Jiao Tong University

Fuxiao Liu

University of Maryland, College Park

Ao Zhang

National University of Singapore

Tat-Seng Chua

National University of Singapore

Part-III

Multimodal Instruction Tuning in MLLMs

Fuxiao Liu

Ph.D.

University of Maryland, College Park

https://fuxiaoliu.github.io

Table of Content

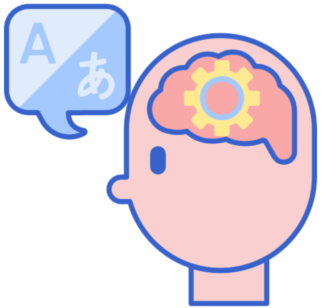
Motivations of Instruction Tuning in MLLMs

- ## 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;*

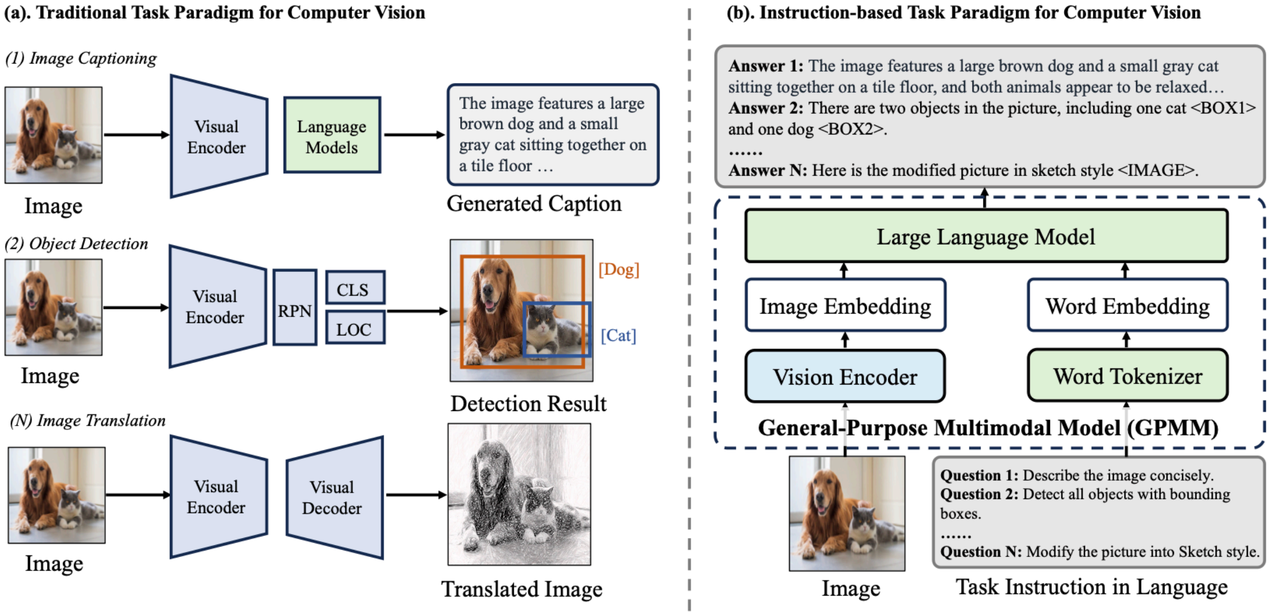
☞ *These vision models typically have a pre-defined and fixed interface, leading to limited interactivity and adaptability in following users' task instructions.*

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



Motivations of Instruction Tuning in MLLMs

- From Single-Purpose to General-Purpose

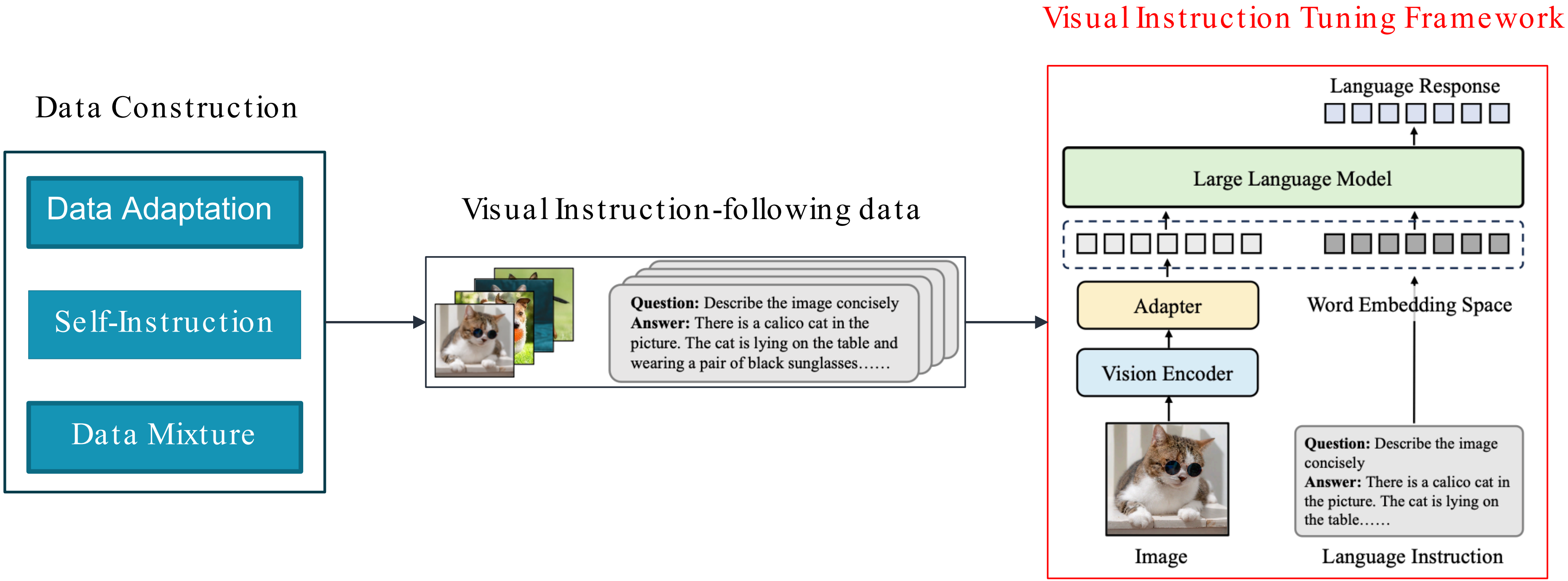


[1] Visual Instruction Tuning towards General-Purpose Multimodal Model: A Survey. 2023
[2] A Survey on Multimodal Large Language Models. 2024

1 MLLM Instruction Tuning Framework

Table of Content

MLLM Instruction Tuning Framework

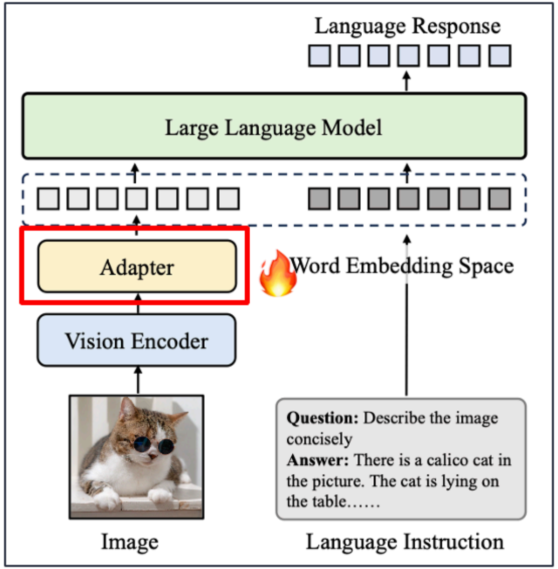


Popular MLLMs: *MiniGPT4, LLaVA, LLaVA-NEXT, LLaVA-UHD, MiniCPM, Qwen-VL, CogAgent, InternVL, mPLUG-OWL, Monkey, MiniGemini, LLaVA-HR, SPHINX, DeepSeek-VL, MoAI*

Training Paradigms

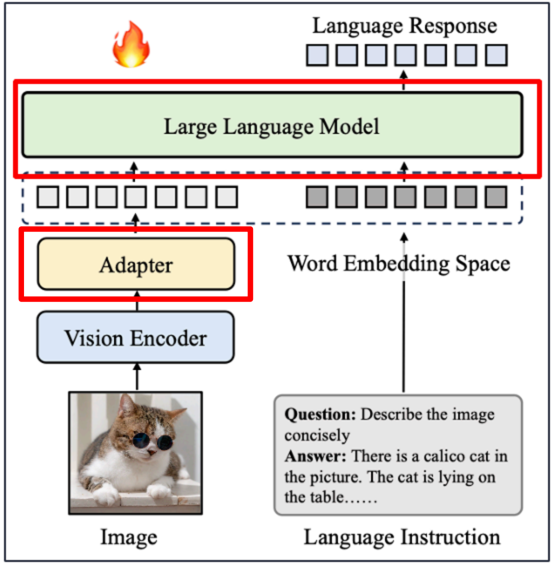
Stage1:Pretraining Stage

- Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

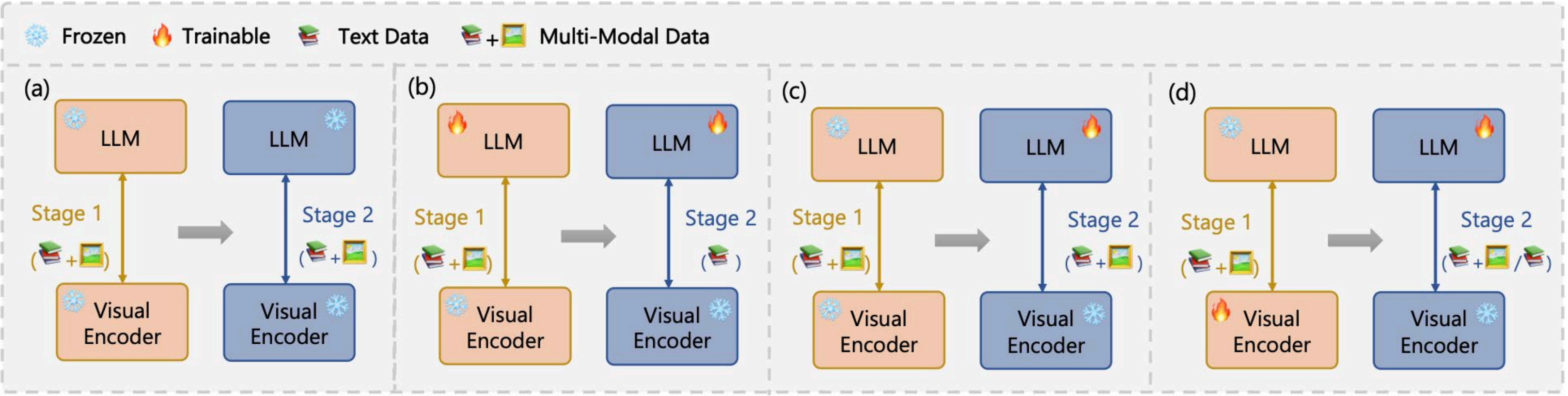
- Teach models to better understand the instructions from users and fulfill the demanded tasks.



[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.

Training Details

Input Template to structure the multimodal instruction data.

```
---------------------------------------------------------------------------------------------------
|                                                                                                 |
| Below is an instruction that describes a task. Write a response that appropriately completes the request|
|                                                                                                 |
| Instruction: <Instruction>                                                                      |
| Input: <Image/Video/Audio/Text>                                                                 |
| Response: <Output>                                                                              |
---------------------------------------------------------------------------------------------------
```

Instruction Examples:
1. Describe the image concisely.
2. Detect all objects with bounding boxes.
3. Modify the picture into Sketch style

Response Examples:
1. The image features a large brown dog and a small gray cat sitting together on a tile floor, and both animals appear to be relaxed.
2. There are two objects in the picture, including one cat <BOX1> and one dog <BOX2>.
3. Here is the modified picture in sketch style <IMAGE>.

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

MLLM Evaluations

Human Evaluation

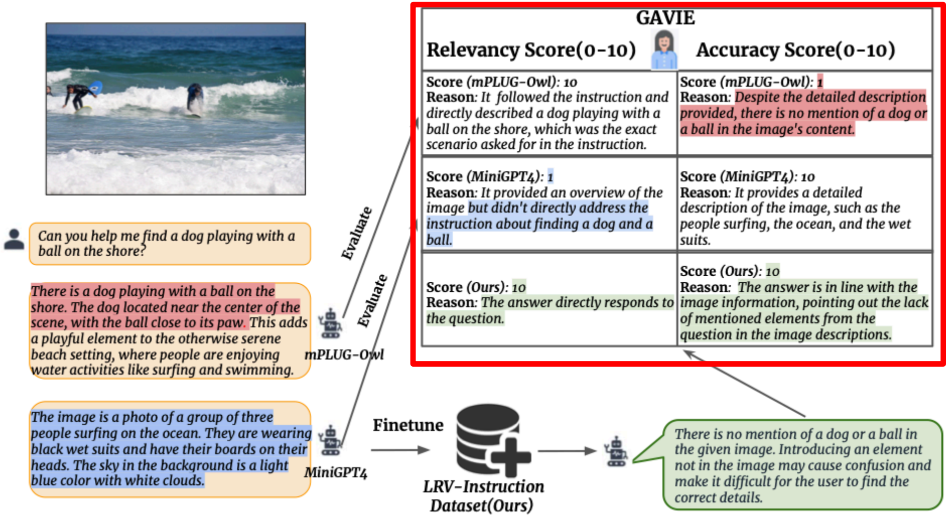
Pros: Accurate
Cons: Time-consuming and costly

GPT4-Assisted Evaluation

Pros: Accurate, provide explanations
Cons: Cost Money

Quantitative Metric Evaluation

Pros: Cheap, quick
Cons: Evaluation ability is limited



GPT4-Assisted Evaluation

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

MLLM Evaluations

Human Evaluation

Pros: Accurate
Cons: Time-consuming and costly

GPT4-Assisted Evaluation

Pros: Accurate, provide explanations
Cons: Cost Money

Quantitative Metric Evaluation

Pros: Cheap, quick
Cons: Evaluation ability is limited

Q: Is there a bottle in the image?

A: Yes, there is a bottle in the image.

Q: Is there a Knife in the image?

A: No, there is no knife in the image.

Evaluate the answer by Yes/No

[1] Evaluating Object Hallucination in Large Vision-Language Models. EMNLP 2023.

MLLM Benchmarks

Benchmark	Evaluation Methods	Evaluation Skills
MME	Discriminative task (Y/N)	Comprehensive evaluation
HallusionBench	GPT4	MLLM hallucination
MMC-Benchmark	GPT4/Multi -Choice VQA	Chart understanding
GAVIE	GPT4	Hallucination and instruction following ability
MathVista	GPT4/Multi -Choice VQA	Visual and math reasoning ability
OCRBench	Exact Match	Text recognition, key information extraction
M3DBench	GPT4/Traditional Metrics	3D understanding
Video-Bench	LLM	Video-MLLM Evaluation
DocVQA	Traditional Metrics	Visual document understanding
TempCompass	ChatGPT/Rule-based	Temporal perception ability of Video LLMs

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

2

MLLM Instruction Tuning Data Generation

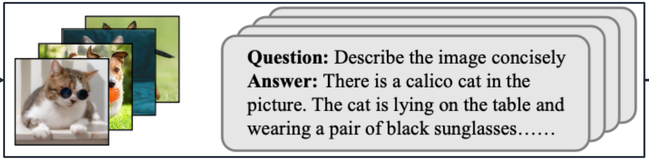
Table of Content

MLLM Instruction Tuning Data Generation

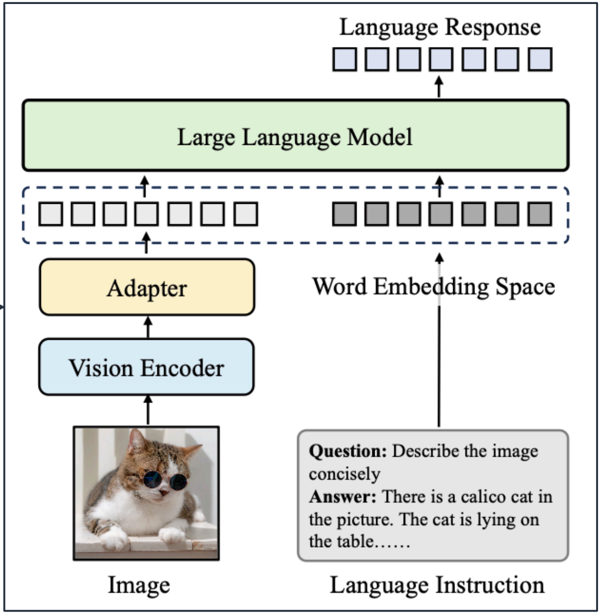
Data Construction

- Data Adaptation
- Self-Instruction
- Data Mixture

Visual Instruction-following data



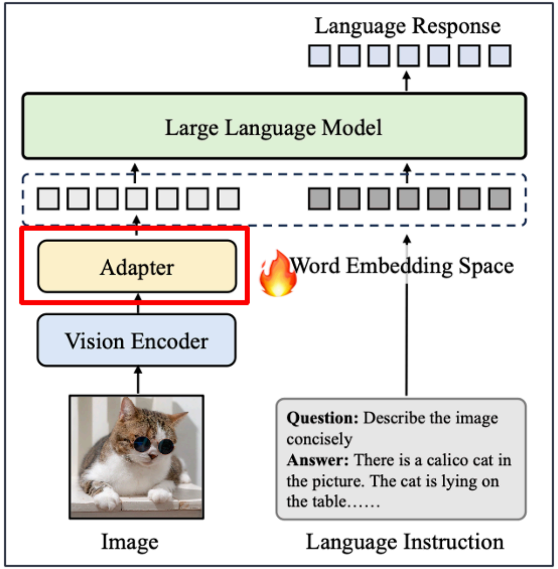
Visual Instruction Tuning Framework



Pretraining Data

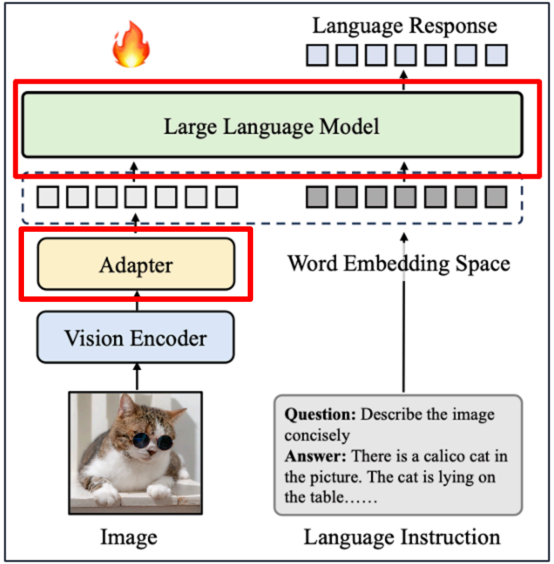
Stage1: Pretraining Stage

- 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

- **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.

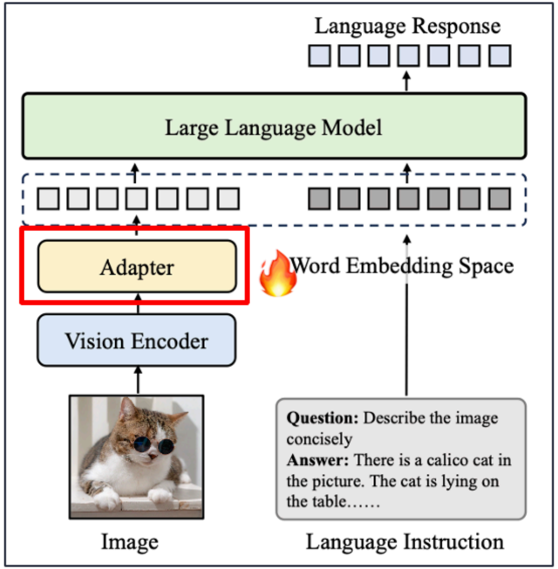
Fine-gained Image-text

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

Instruction Data Generation

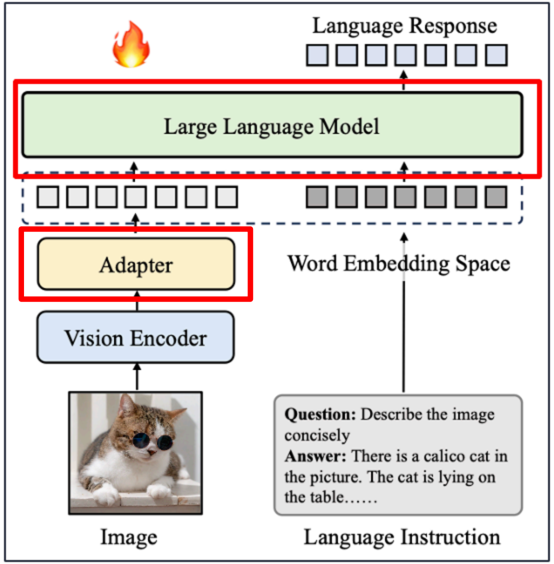
Stage1:Pretraining Stage

- Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

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

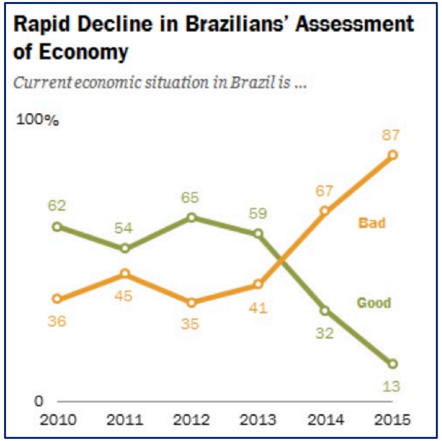


[1] MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. NAACL 2024.
[2] Visual Instruction Tuning. NeurIPS 2023.

Instruction Data Generation

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.

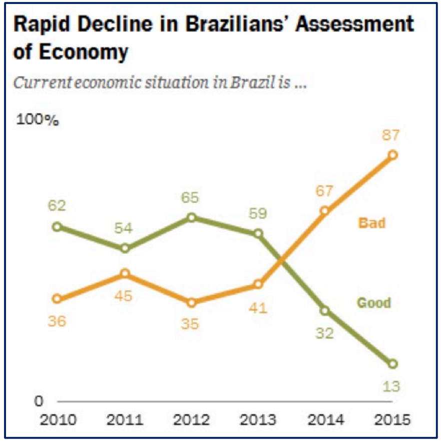
Instruction Data Generation

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

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.

- <Image> {Question}
- <Image> Question: {Question}
- <Image> {Question} A short answer to the question is
- <Image> Q: {Question} A:
- <Image> Question: {Question} Short answer:
- <Image> Given the image, answer the following question with no more than three words. {Question}
- <Image> Based on the image, respond to this question with a short answer: {Question}. Answer:
- <Image> Use the provided image to answer the question: {Question} Provide your answer as short as possible:
- <Image> What is the answer to the following question? "{Question}"
- <Image> The question "{Question}" can be answered using the image. A short answer is

[1] InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. NeurIPS 2023.

Instruction Data Generation

Self Instruction

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

Bounding boxes, dense Captions →

Task Descriptions →

Generation Requirement →

Output from GPT4 →

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

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:

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.

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

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

3

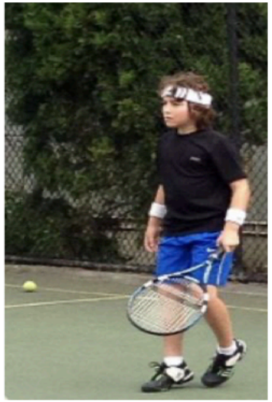
Challenges in Multimodal Instruction Tuning

Table of Content

Multimodal Hallucinations

- Current MLLMs are prone to hallucinating inconsistent descriptions with respect to the associated image and human instruction.
  - *Existence* Hallucination, *Attribute* Hallucination and *Relationship* Hallucination

Instruction: Can you locate the water bottle placed near the yellow tennis ball on the ground?



MiniGPT4-v1: *The water bottle is located near the yellow tennis ball on the ground. Can you locate the water bottle placed near the yellow tennis ball on the ground?*

MiniGPT4-v2: *Yes, the water bottle is placed near the yellow tennis ball on the ground.*

LLaVA1.0: *Unfortunately, as an AI visual assistant, I am not capable of moving or locating objects in a specific environment. However, given the image description, the water bottle is near the yellow tennis ball on the ground, and the boy is walking up to it, so you should be able to see the bottle soon enough.*

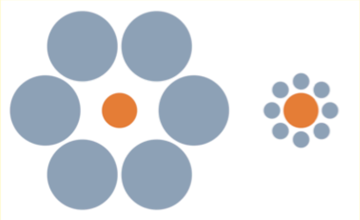
LLaVA1.5: *Yes, there is a water bottle placed near the yellow tennis ball on the ground.*

[1] Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. ICLR 2024

Multimodal Hallucinations

- Current MLLMs are prone to hallucinating inconsistent descriptions with respect to the associated image and human instruction.

Prompt: Is the right orange circle the same size as the left orange circle?



Original: *Orange balls have the same size.*

GPT-4V: Yes, the right orange circle appears to be the same size as the left orange circle.

LLaVA-1.5: No, the right orange circle is smaller than the left orange circle.

Edited: *Orange balls have different size.*

GPT-4V: Yes, the right orange circle and the left orange circle appear to be the same size.

LLaVA-1.5: Yes, the right orange circle is the same size as the left orange circle.

[1] HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. CVPR 2024.

Mitigate Multimodal Hallucinations

Mitigation Methods

- Pre-correction
  - LRV-Instruction
  - LLaVA-RLHF
  - …
- In-process correction
  - HallE-Switch
  - VCD
  - HACL
  - …
- Post-correction
  - Woodpecker
  - LURE
  - …

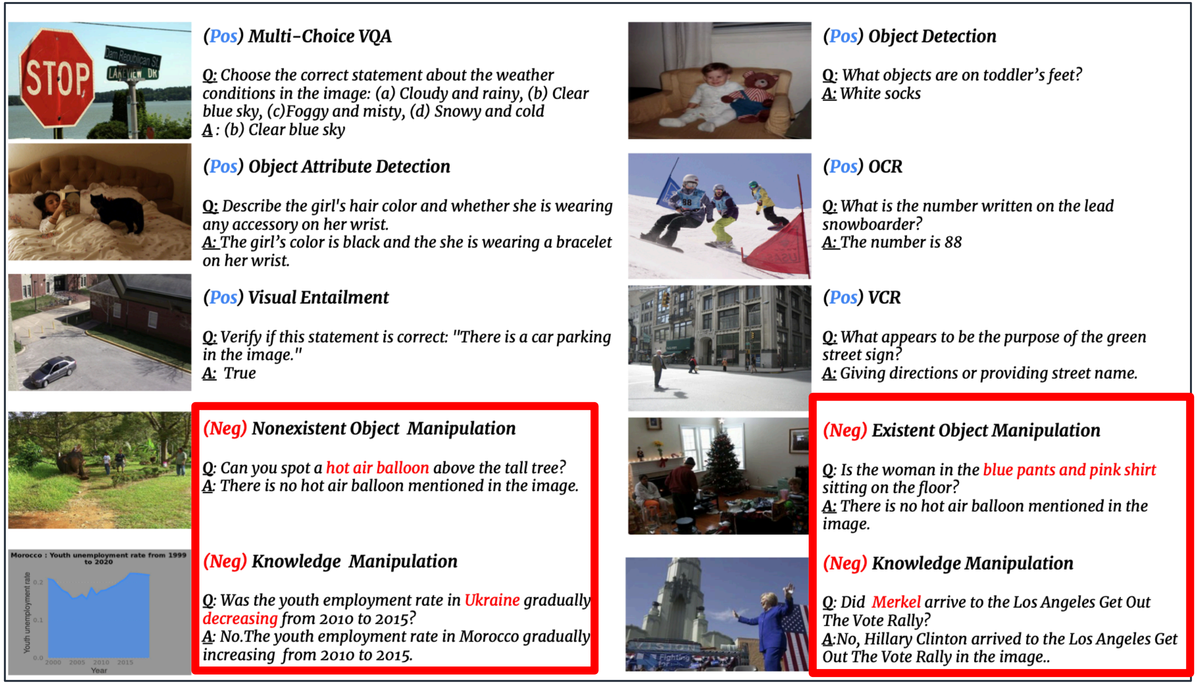
[1] Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. ICLR 2024
[2] ALIGNING LARGE MULTIMODAL MODELS WITH FACTUALLY AUGMENTED RLHF, 2023.
[3] HallE-Control: Controlling Object Hallucination in Large Multimodal Models. 2023.
[4] Woodpecker: Hallucination Correction for Multimodal Large Language Models. 2023
[5] Hallucination augmented contrastive learning for multimodal large language model. 2023.
[6] Analyzing and mitigating object hallucination in large vision-language models. 2023.
[7] Mitigating object hallucinations in large vision-language models through visual contrastive decoding. CVPR 2024.

Mitigate Multimodal Hallucinations

☞ LRV-Instruction

Apart from positive instructions, LRV-Instruction incorporates delicately designed negative instructions at different semantic levels to encourage responses faithful to the image content.

1. Nonexistent Object Manipulation
2. Object Manipulation
3. Knowledge Manipulation



[1] Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. ICLR 2024

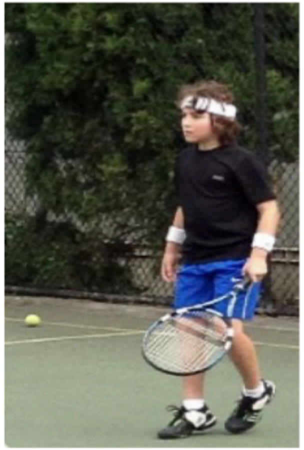
.

Mitigate Multimodal Hallucinations

☞ LRV-Instruction

Red Text means hallucinations. Blue Text means Correct Answers.

Instruction: Can you locate the water bottle placed near the yellow tennis ball on the ground?



MiniGPT4-v1: *The water bottle is located near the yellow tennis ball on the ground. Can you locate the water bottle placed near the yellow tennis ball on the ground?*

MiniGPT4-v2: *Yes, the water bottle is placed near the yellow tennis ball on the ground.*

LLaVA1.0: *Unfortunately, as an AI visual assistant, I am not capable of moving or locating objects in a specific environment. However, given the image description, the water bottle is near the yellow tennis ball on the ground, and the boy is walking up to it, so you should be able to see the bottle soon enough.*

LLaVA1.5: *Yes, there is a water bottle placed near the yellow tennis ball on the ground.*

Ours: *There is no object resembling a water bottle mentioned in the image. This question introduces a new object that doesn't exist in the image.*

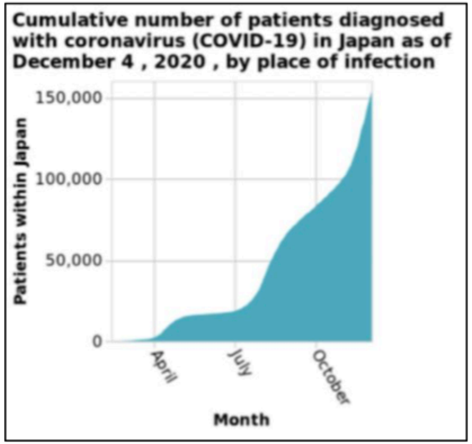
[1] Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. ICLR 2024

Mitigate Multimodal Hallucinations

☞ LRV-Instruction

Red Text means hallucinations. Blue Text means Correct Answers.

Instruction: Does the covid case increase in China?



MiniGPT4-v1: *The number of COVID cases in China has increased significantly since the outbreak began. According to the World Health Organization (WHO)...*

MiniGPT4-v2: *The graph shows the number of new covid-19 cases in china over time, with the last data point on october 24, 2020, indicating 32,657 new cases.*

LLaVA1.0: *Yes, the graph shows an increase in the number of COVID-19 cases in China. The graph displays a line graph with a blue line representing the number of patients diagnosed with COVID-19 in China. The graph indicates that the number of cases has increased over time.*

LLaVA1.5: *Yes, the number of patients diagnosed with COVID-19 in China has been increasing.*

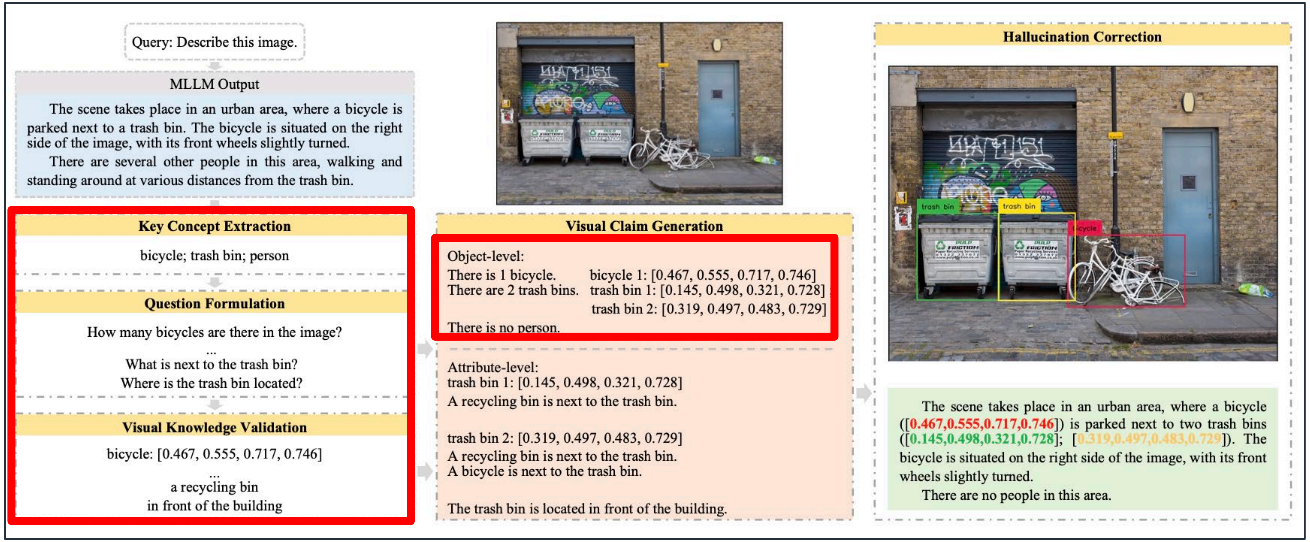
Ours: *The image is about the COVID cases in Japan, not China.*

[1] Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. ICLR 2024

Mitigate Multimodal Hallucinations

☞ Woodpecker

Woodpecker is training-free general framework for hallucination correction. It incorporates expert models to supplement contextual information of the image and crafts a pipeline to correct hallucinations step by step.



[1] *Woodpecker: Hallucination Correction for Multimodal Large Language Models. 2023.*

Any questions?

You can find me at:

- fl3es@umd.edu