

A Survey on Multimodal Large Language Models

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, Enhong Chen

Link: <https://arxiv.org/abs/2306.13549>

Dinh-Thang Duong
Nguyen-Thuan Duong

Abstract

◆ Content

Multimodal Large Language Model (MLLM) recently has been a new rising research hotspot, which uses powerful Large Language Models (LLMs) as a brain to perform multimodal tasks. The surprising emergent capabilities of MLLM, such as writing stories based on images and OCR-free math reasoning, are rare in traditional methods, suggesting a potential path to artificial general intelligence. In this paper, we aim to trace and summarize the recent progress of MLLM. First of all, we present the formulation of MLLM and delineate its related concepts. Then, we discuss the key techniques and applications, including Multimodal Instruction Tuning (M-IT), Multimodal In-Context Learning (M-ICL), Multimodal Chain of Thought (M-CoT), and LLM-Aided Visual Reasoning (LAVR). Finally, we discuss existing challenges and point out promising research directions. In light of the fact that the era of MLLM has only just begun, we will keep updating this survey and hope it can inspire more research. An associated GitHub link collecting the latest papers is available at <https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models>.

Introduction

❖ Introduction

 You
write a python code that print "hello world"

 ChatGPT
Certainly! Here's a simple Python code that prints "Hello, World!" to the console:

```
python
print("Hello, World!")
```

You can run this code in a Python interpreter, and it will display "Hello, World!" as the output.

Copy code

3



Textual Input



GPT3.5:
Unimodal LLM

Textual Output

Introduction

❖ Introduction

AI You



How many people are there in this image?

ChatGPT

There are four people in this image.

Save Share Print

A screenshot of a ChatGPT interface showing the iconic Abbey Road cover. The image is labeled 'You' and has an 'AI' icon. A question 'How many people are there in this image?' is asked, with an orange arrow pointing from the text to the image. The AI response 'There are four people in this image.' is shown below, with another orange arrow pointing from the text to the image. The interface includes standard sharing and saving options at the bottom.

GPT4:
Multimodal LLM

Visual Input

Textual Input

Textual Output

Introduction

❖ What is LLMs?

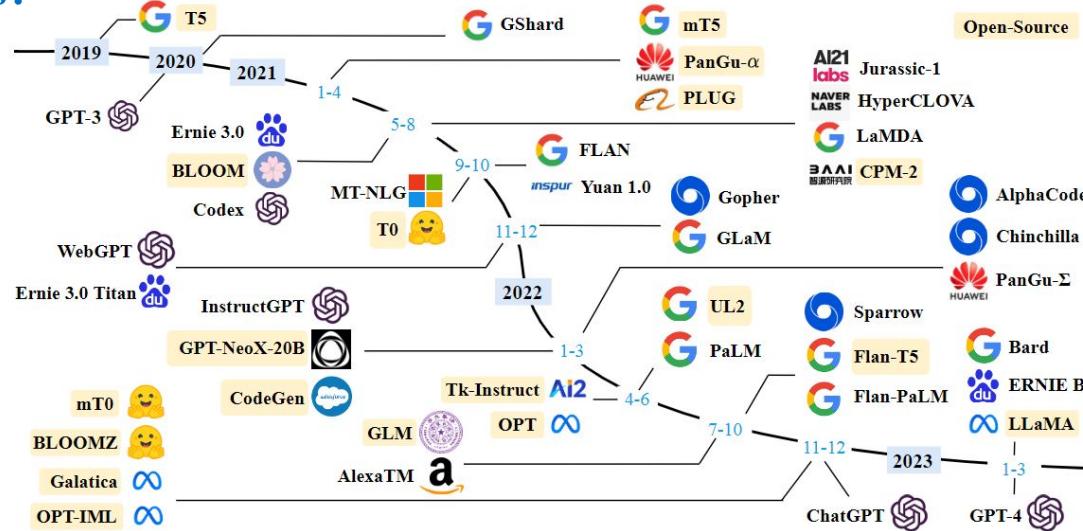


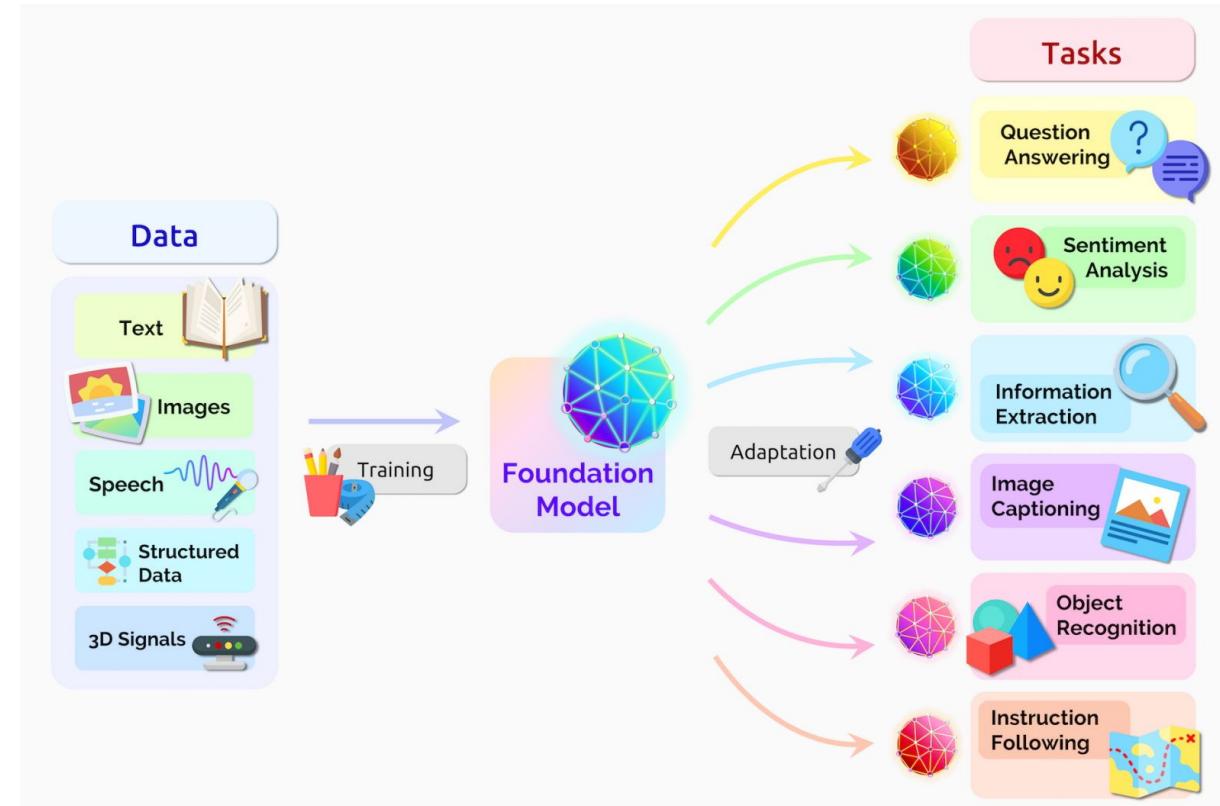
Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

LLMs (Large Language Models): Are language models that were trained on a very large corpus of text. This made them capable of performing various NLP tasks with high precision.

Introduction

❖ What is LLMs?

LLMs are often pretrained on a vast majority of data and designed to be adaptable to a wide variety of tasks (Foundation models).

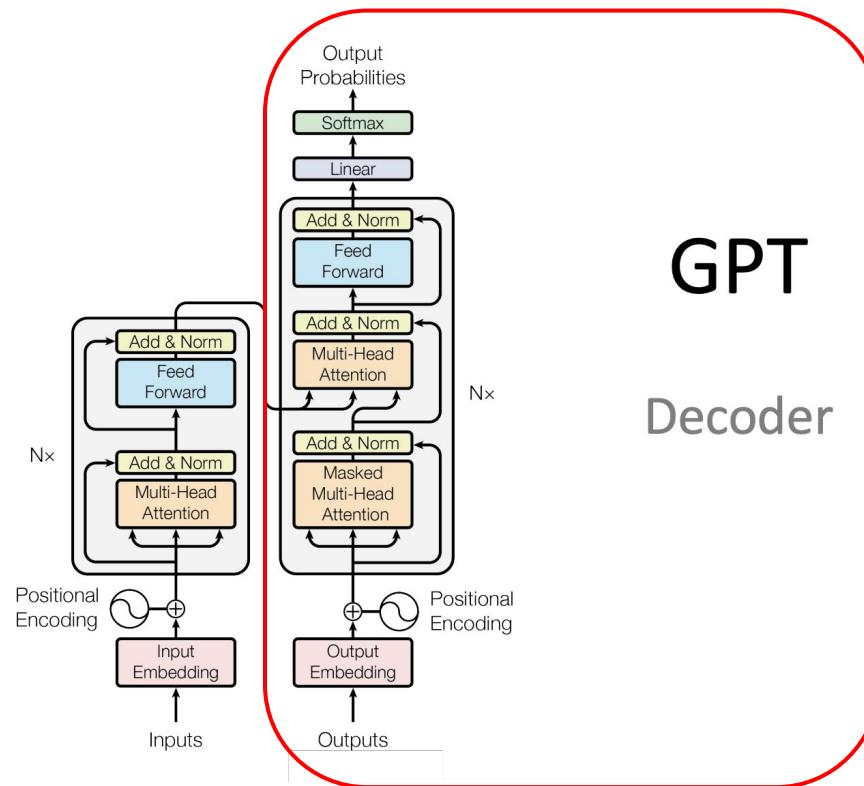


Introduction

❖ What is LLMs?

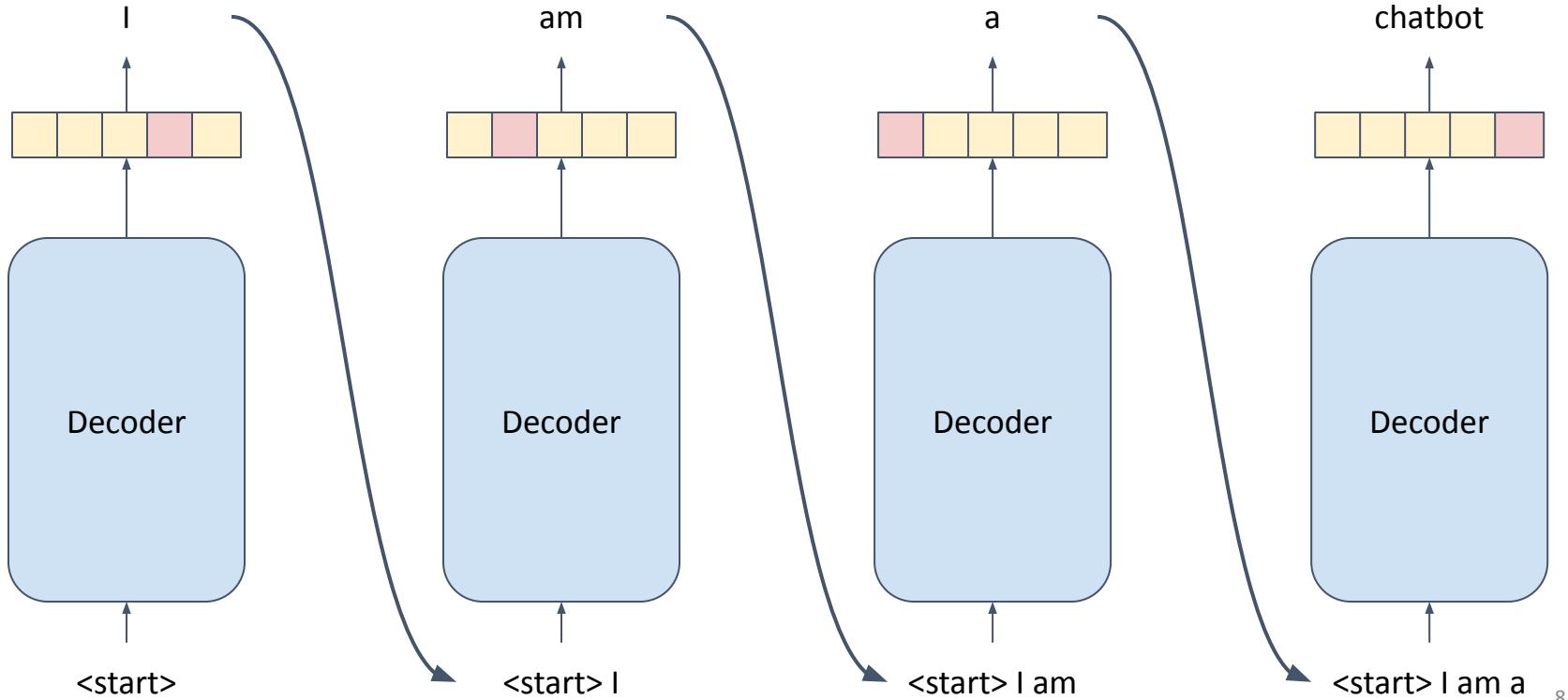
BERT

Encoder



Introduction

❖ What is LLMs?



Introduction

❖ What is LLMs?

Prompt:

Write a python function that receive an image and plot it using matplotlib library.



You

Write a python function that receive an image and plot it using matplotlib library.



ChatGPT

To create a Python function that receives an image and plots it using the Matplotlib library, you can use the following code as an example. You'll need to have Matplotlib and a library like PIL (Pillow) installed to work with images:

python

Copy code

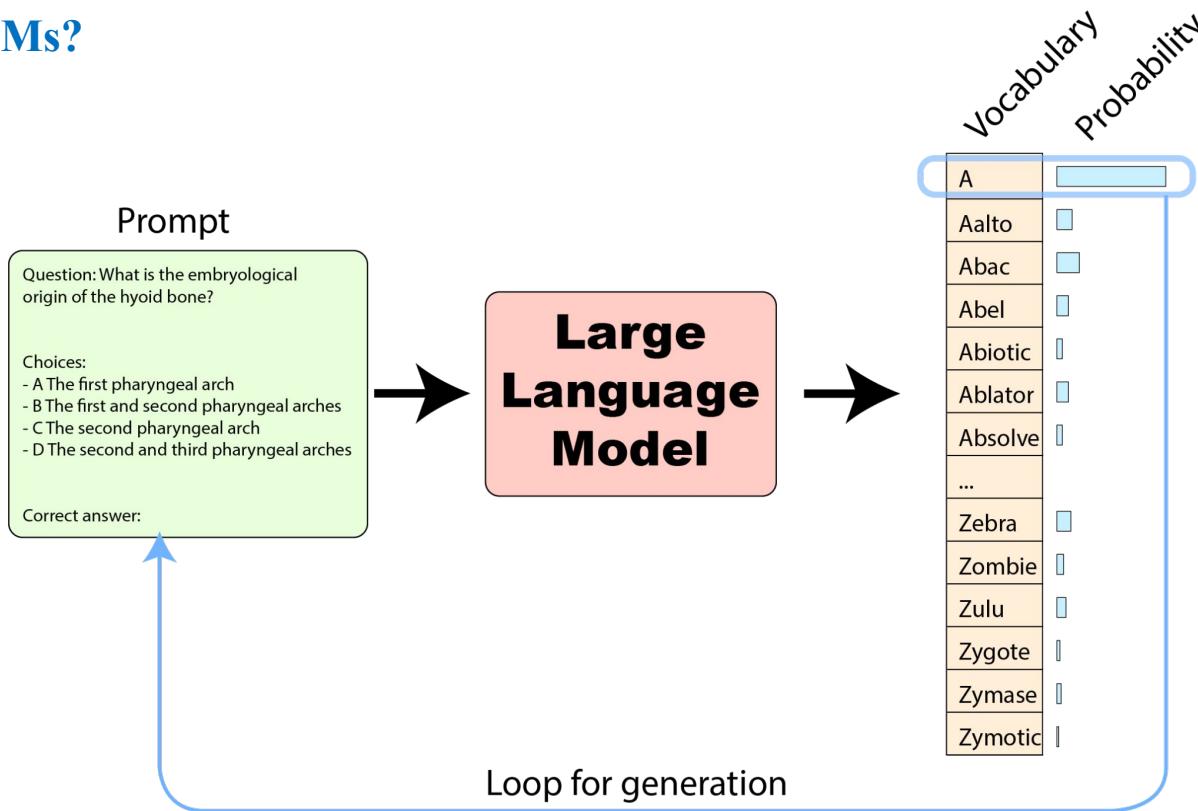
```
import matplotlib.pyplot as plt
from PIL import Image

def plot_image(image_path):
    try:
        # Open the image using PIL (Pillow)
        img = Image.open(image_path)

        # Create a Matplotlib figure and axis
        fig, ax = plt.subplots()
```

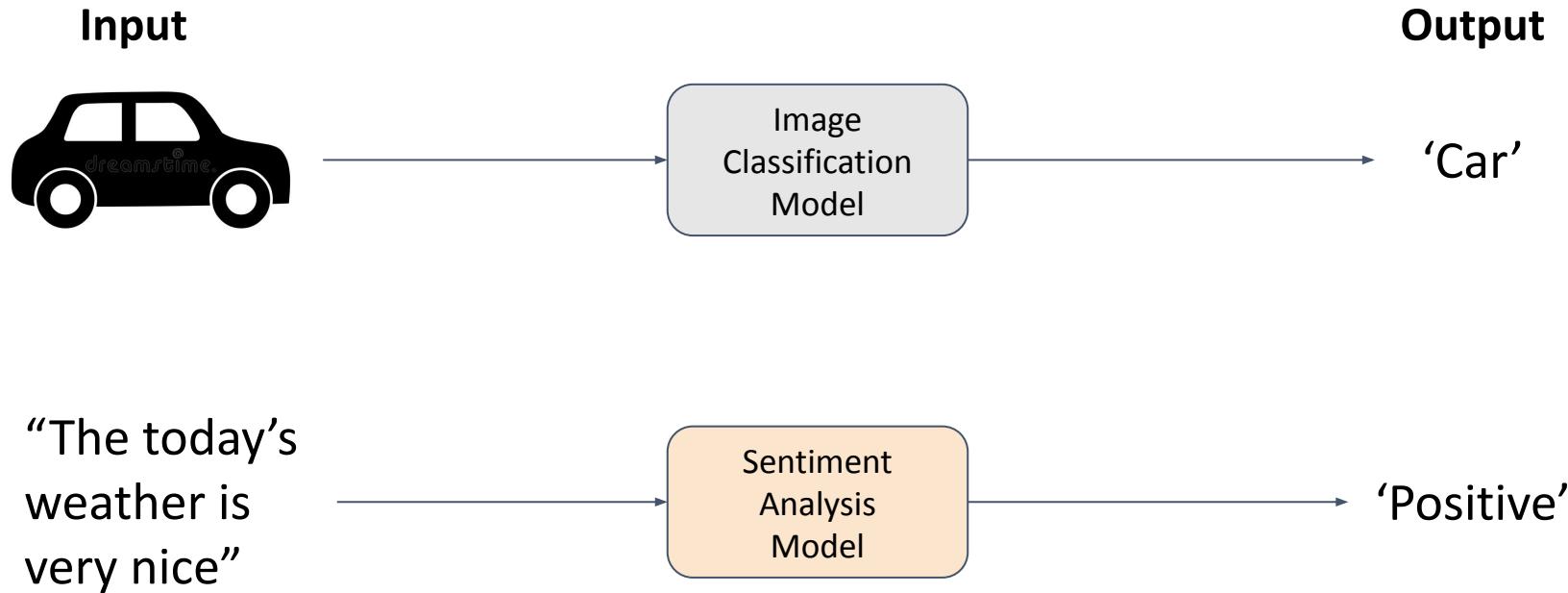
Introduction

❖ What is LLMs?



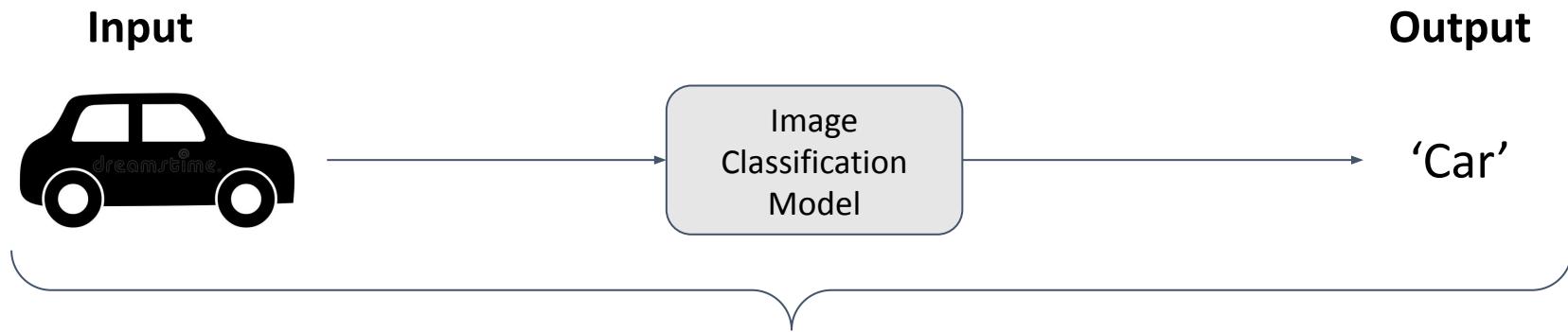
Introduction

❖ What is multimodality?



Introduction

❖ What is multimodality?



Involve processing and integrating only one type of data => Unimodal

Introduction

❖ What is multimodality?

Answer: No



Answer: Yes



complementary scenes

Tuple: <girl, walking, bike>

Question: Is the girl walking the bike?

Input:

- Image (Context)
- Text (Question)

Output:

- Text (Answer)

Involve processing and integrating multi type of data => **Multimodal**

Introduction

❖ What is M-LLMs



AI

You



How many people are there in this image?



ChatGPT

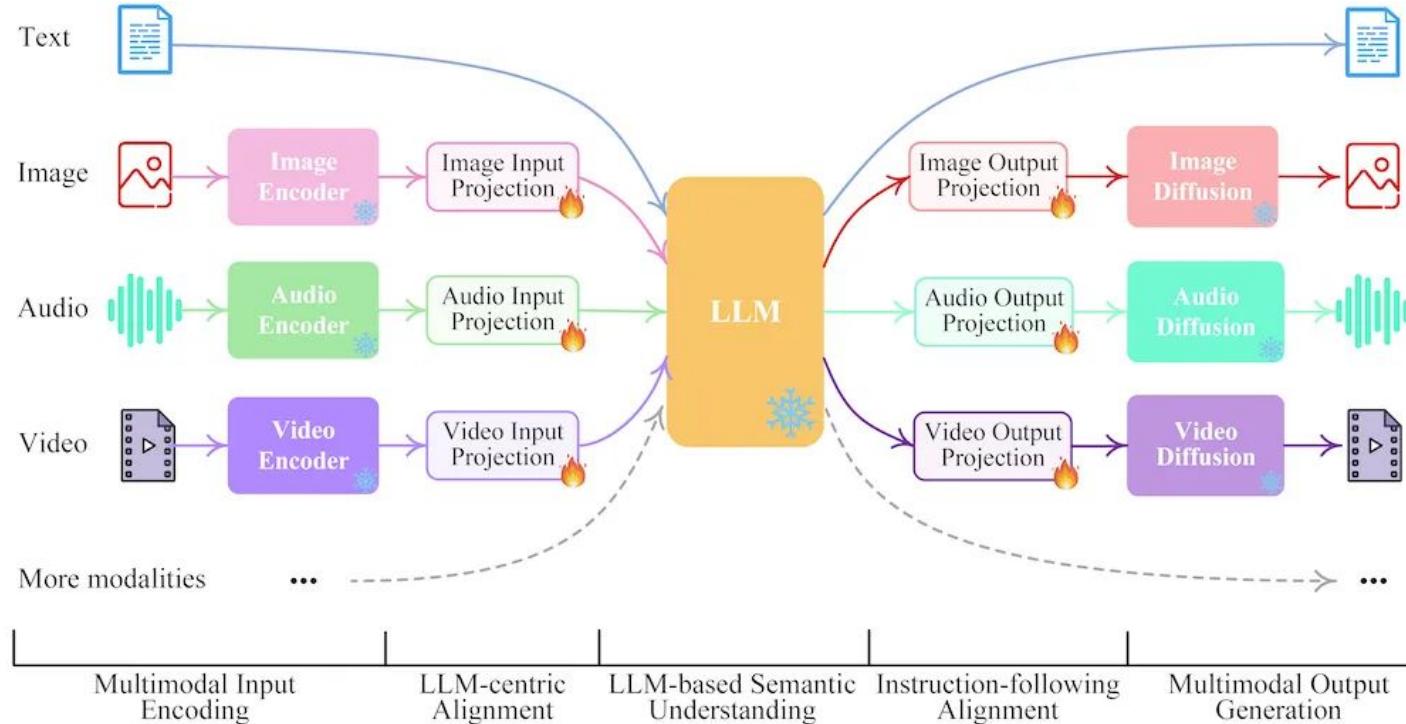
There are four people in this image.



Multimodal Large Language Models (M-LLMs): uses powerful Large Language Model as a brain to perform multimodal tasks, which may involve processing multi types of data.

Introduction

❖ What is M-LLMs



Introduction

M-LLMs Applications

Prompt:

Localize each person in the image using bounding box. What is the image size of the input image?



GPT-4V:

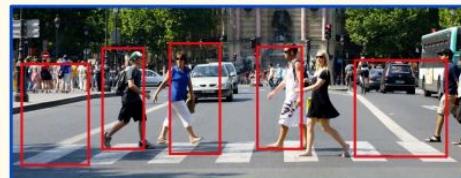
The image size of the input image is **1000 x 333** pixels.

Bounding boxes for each person:

- Person 1 (leftmost): $(x_1, y_1) = (22, 108), (x_2, y_2) = (170, 300)$
- Person 2: $(x_1, y_1) = (200, 80), (x_2, y_2) = (290, 270)$
- Person 3: $(x_1, y_1) = (350, 70), (x_2, y_2) = (460, 280)$
- Person 4: $(x_1, y_1) = (540, 80), (x_2, y_2) = (640, 270)$
- Person 5 (rightmost): $(x_1, y_1) = (760, 100), (x_2, y_2) = (920, 290)$

Note: These bounding box coordinates are approximate and may not be exact.

Visualization of Detected Bbox:



Object Localization

Prompt:

Localize each person in the image using bounding box. The image size is (width, height) = (998,786)



GPT-4V:

There is only one person in the image, so the bounding box for this person would be:

- Top left: (254, 118)
- Bottom right: (731, 693)

Explanation: The top left point of the bounding box is at $(x, y) = (254, 118)$, which is the top left corner of the person in the image. The bottom right point of the bounding box is at $(x, y) = (731, 693)$, which is the bottom right corner of the person in the image. This bounding box encompasses the entire person in the image.

Visualization of Detected Bbox:



Prompt:

Localize each car in the image using bounding box. What is the image size of the input image?



GPT-4V:

The image size of the input image is **500 x 281** pixels.

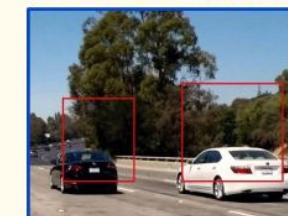
Bounding box for the black car:

- Top left corner: (63, 114)
- Bottom right corner: (196, 220)

Bounding box for the white car:

- Top left corner: (287, 96)
- Bottom right corner: (474, 220)

Visualization of Detected Bbox:



Introduction

❖ M-LLMs Applications

Prompt:

Count the number of oranges in the image.



GPT-4V:

There are 3 oranges in the image.

Prompt:

Count the number of people in the image.



GPT-4V:

There are 3 people in the image.

Object Counting

Introduction

❖ M-LLMs Applications

Prompt:

Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{"Surname": xxx, "Given Name": xxx, "USCIS #": xxx, "Category": xxx, "Country of Birth": xxx, "Date of Birth": xxx, "SEX": xxx, "Card Expires": xxx, "Resident Since": xxx}
```



GPT-4V:

```
{"Surname": "STEVENS", "Given Name": "ICHIKA", "USCIS #": "000-000-000", "Category": "IR1", "Country of Birth": "Japan", "Date of Birth": "18 AUG 1988", "SEX": "F", "Card Expires": "07/04/2027", "Resident Since": "07/05/2017"}
```

OCR-Free eKYC

Introduction

❖ M-LLMs Applications

Prompt:

Who is the person in the image and what is the person doing?



GPT-4V:

The person in the image is the current President of the United States, Joe Biden. He is standing at a podium, likely delivering a speech, at the G7 Summit in Hiroshima, Japan in 2023.

Prompt:

Who is the person in the image and what is the person doing?



GPT-4V:

The person in the image is Jensen Huang, the CEO and co-founder of NVIDIA. He is holding and presenting a product from NVIDIA, likely a graphics processing unit (GPU).

Knowledge-based
Visual Question Answering

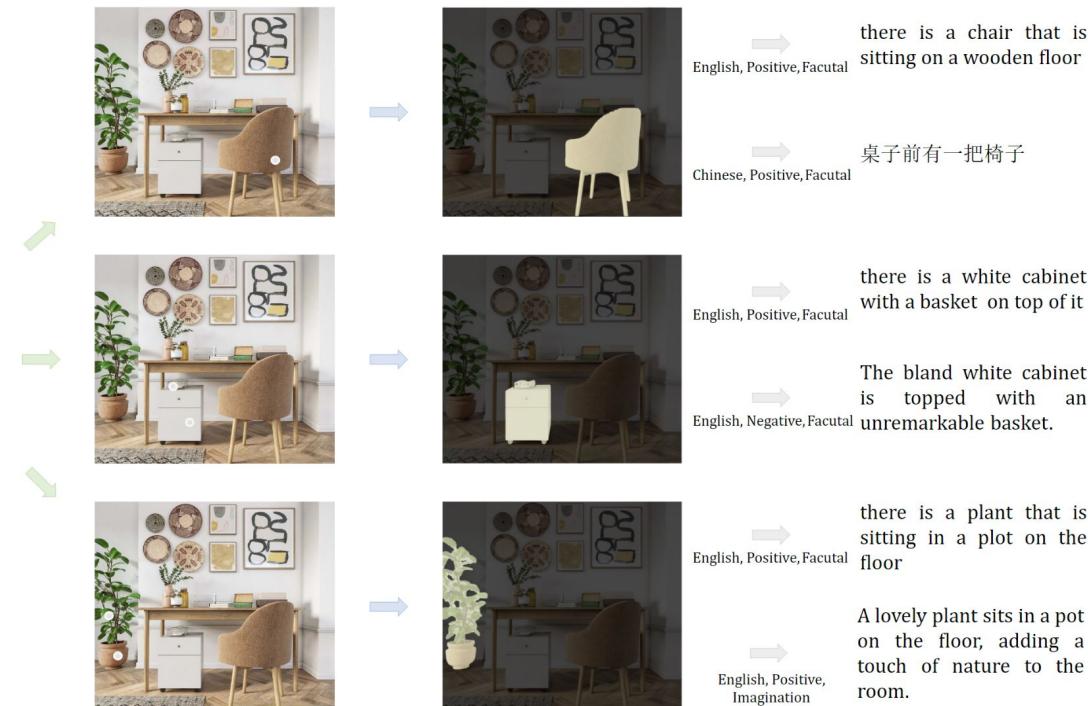
Introduction

❖ M-LLMs Applications

Visual Control
Language Control



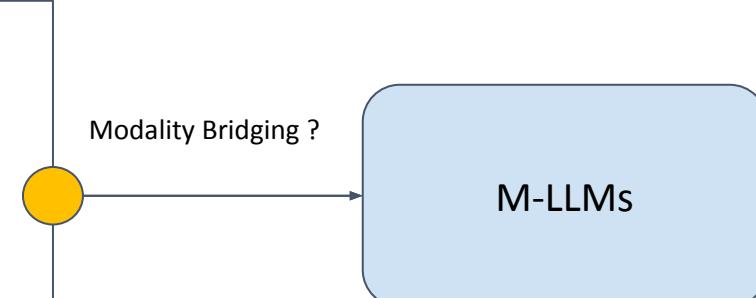
Click Prompt



Caption Anything

Introduction

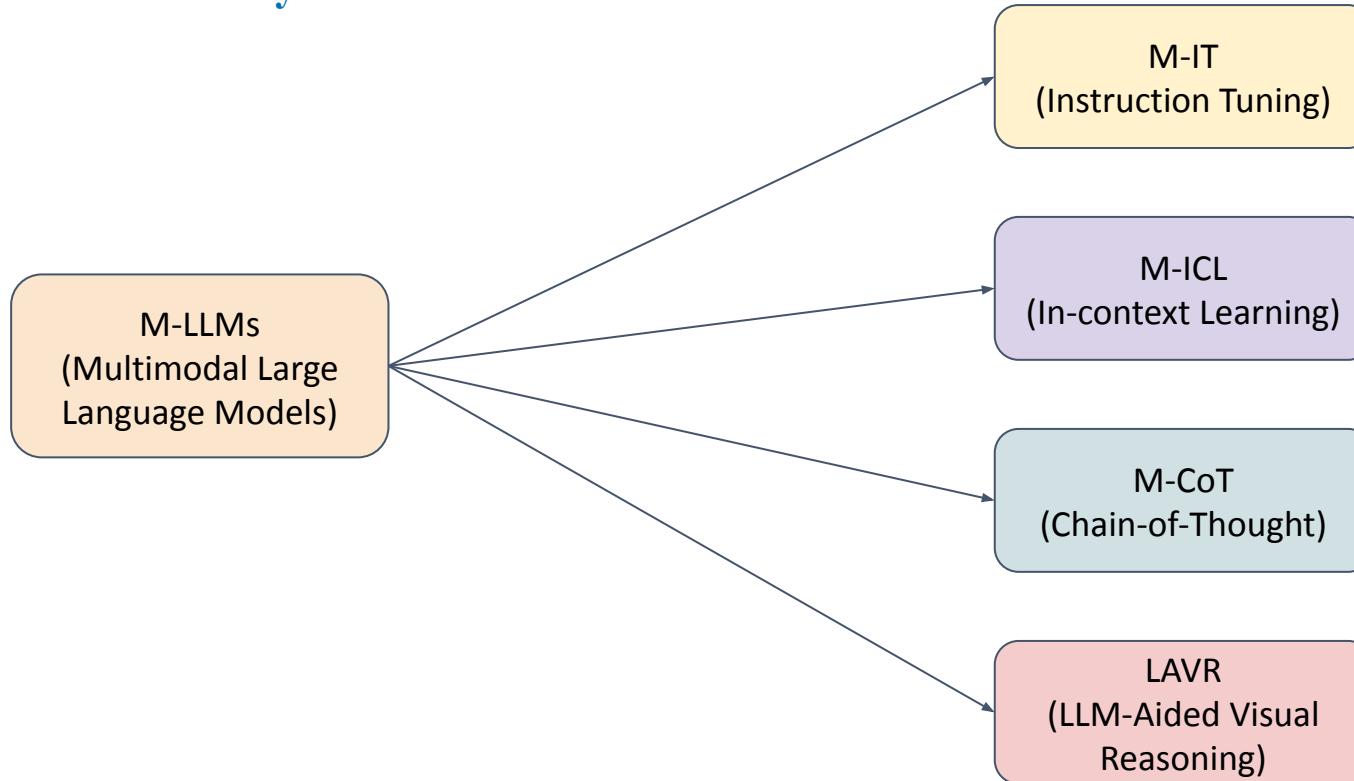
❖ M-LLMs Challenges



Instruction: Complete the below task
Input: Describe the image

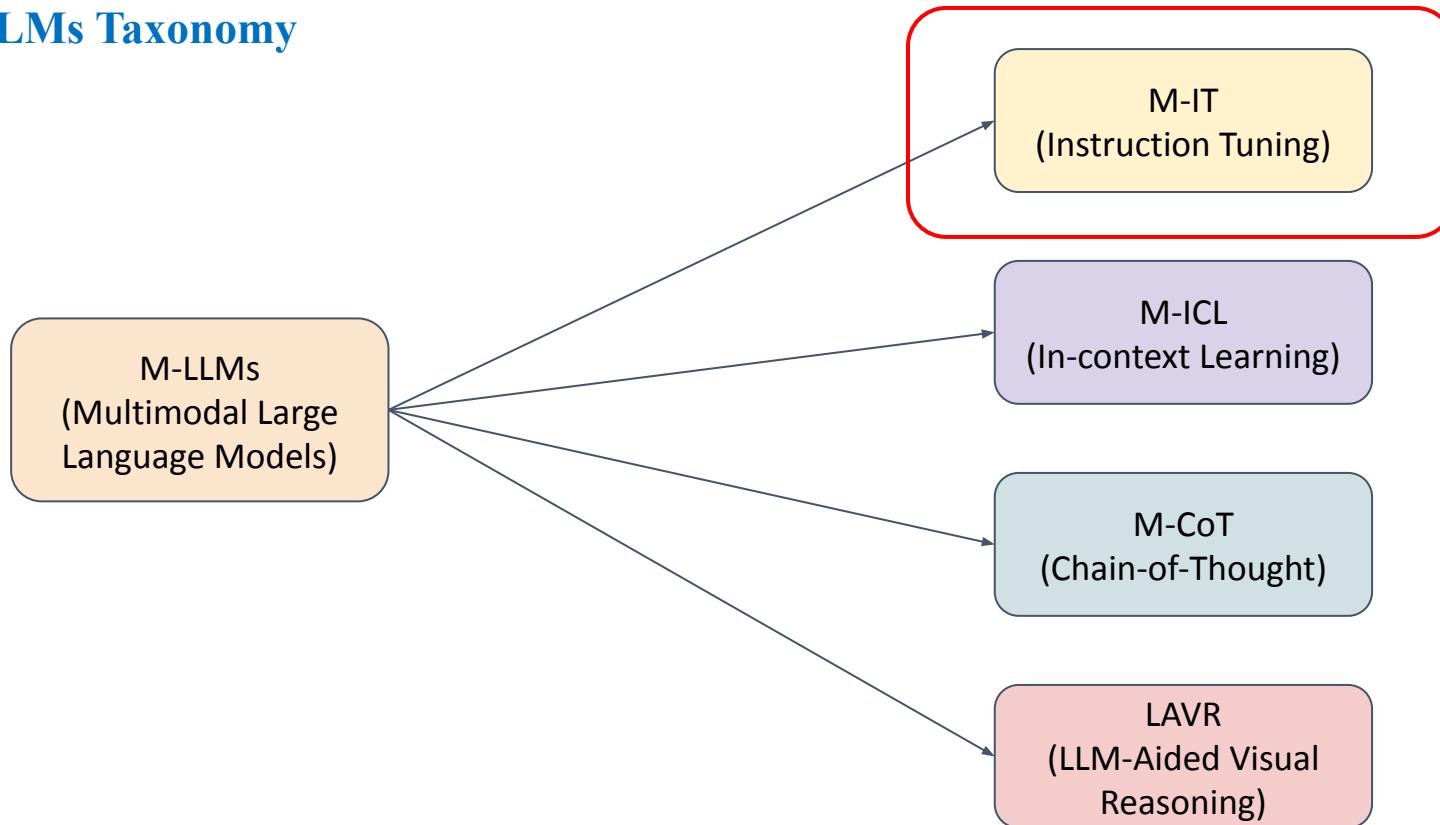
Methods

❖ M-LLMs Taxonomy



Methods

❖ M-LLMs Taxonomy



Methods

❖ M-IT: Introduction

prompt

string · lengths



Could you suggest recipes for classic southern dishes?

What temperature should I dial down my water heater to?

What is the boiling point for a sweet potato?

How can I clean rusty canning jars without a special cleaner? 1

How long will my leftovers keep refrigerated?

What role does feminism have in current society?

How can I reduce Carbon Monoxide emission levels?

Outline the Imhotep theory

Common prompting: General or standard questions or statement used to initiate a response or interaction.

Methods

❖ M-IT: Instruction-following data

instruction string	input string	output string
Detect hate speech in the following sentence:	We should hate all those immigrants.	Yes
Find the total price for the items in the given list, given...	List: laptop, cellphone, computer mouse	30
Given a product description, identify the main product...	The Samsung Galaxy Fold has three AMOLED screens, one o...	The Samsung Galaxy Fold boasts a unique triple-screen design and an impressive long...
Output a way to reduce the effects of a given issue.	Air Pollution	One way to reduce the effects of air pollution is to reduce the amount of...
Give three proverbs related to the input.	Love	1. Love conquers all. 2. Love makes the world go round. 3. Love knows no bounds.
Summarize an article about the importance of being eco...		Reducing consumption of resources and increasing respect and regard for our...

Instruction
Prompting: Specific, directive statements or questions that guide the LLMs to perform a particular action.

Methods

❖ M-IT: Instruction-following data

Positive Neutral Negative

Instruction: Given a sentence, classify its sentiment into one of three classes: positive, negative, neutral

Input: What is the weather like today?

Output: Neutral

Input: I feel bad

Output:



You

Instruction: You will be given a sentence, please classify its sentiment into one of three classes: positive, negative, neutral

Input: This is a very bad food.

Output: Negative

Input: Today's weather is not bad, don't you think?

Output:



ChatGPT

Neutral



You

Input: This restaurant is well-known by its customer service



ChatGPT

Positive



Methods

❖ M-IT

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>

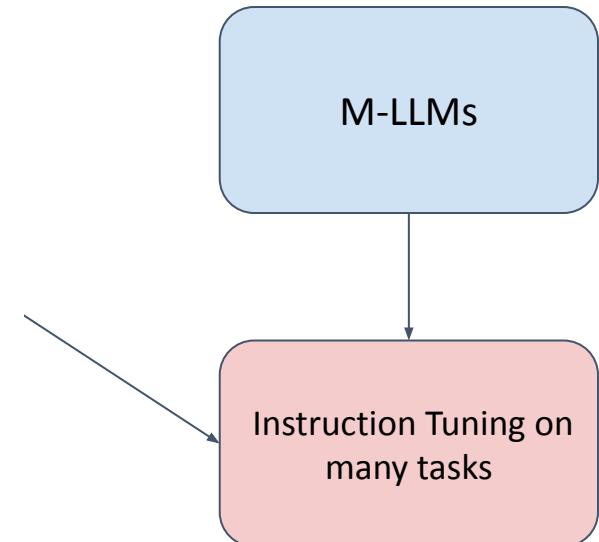
Response type 2: detailed description

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>



Methods

❖ M-IT: Instruction-following data with image

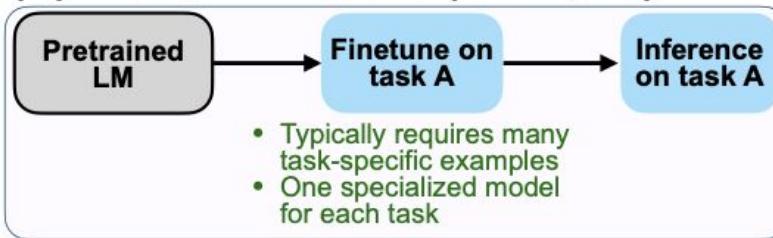
- <Image> {Question}
- <Image> Question: {Question}
- <Image> {Question} A short answer to the question is
- <Image> Q: {Question} A:
- <Image> Question: {Question} Short answer:
- <Image> Given the image, answer the following question with no more than three words. {Question}
- <Image> Based on the 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

Table 2. Instruction templates for VQA datasets, cited from [23]. <Image> and {Question} are the image and the question in the original VQA datasets, respectively.

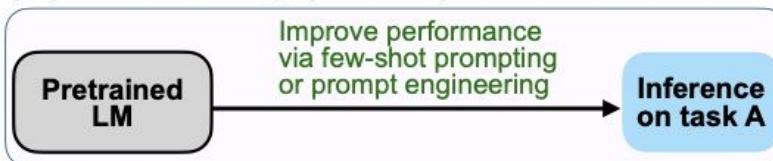
Methods

❖ M-IT: Introduction

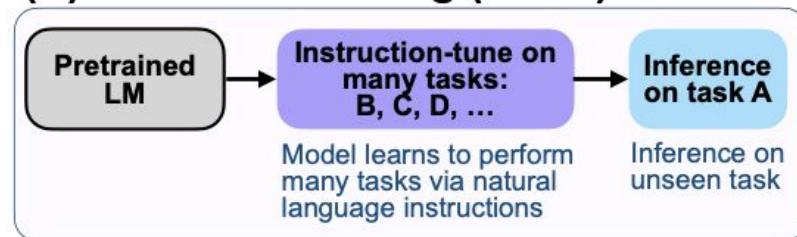
(A) Pretrain–finetune (BERT, T5)



(B) Prompting (GPT-3)



(C) Instruction tuning (FLAN)



Methods

❖ M-IT template for multimodal instruction data

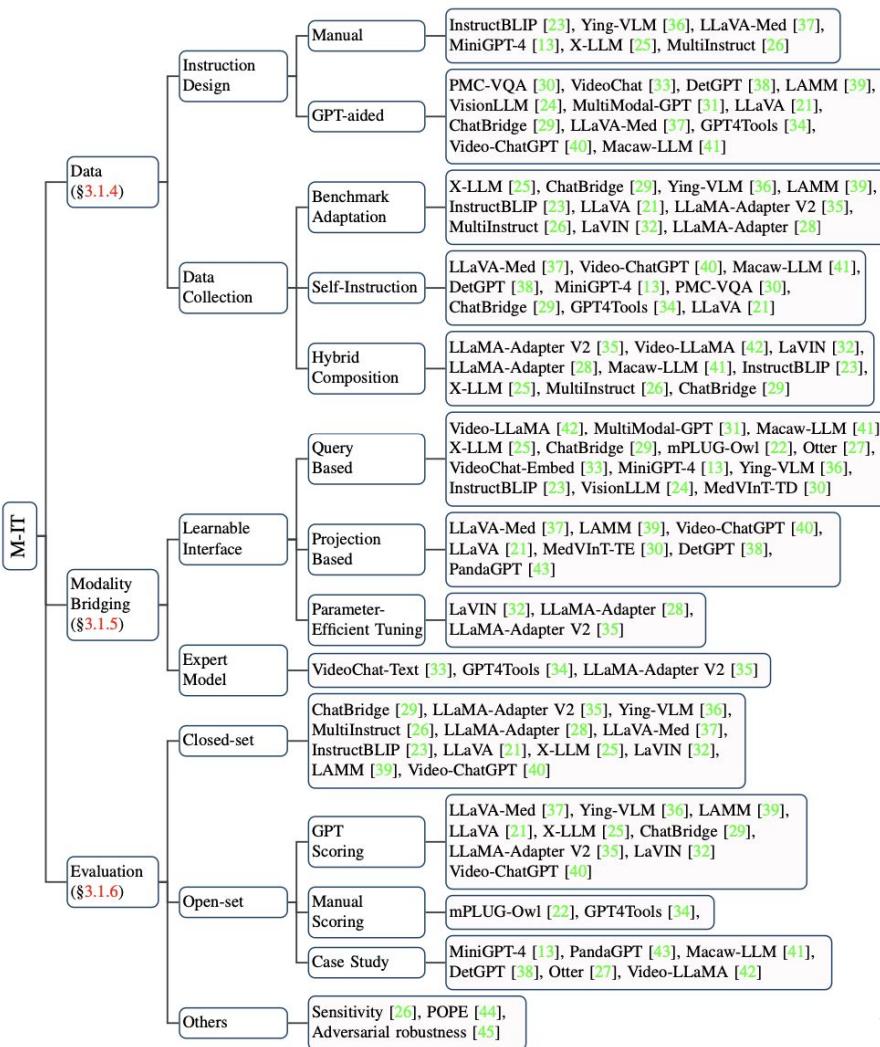
<BOS> Below is an instruction that describes a task.
Write a response that appropriately completes the request

Instruction: **<instruction>**
Input: **{<image>, <text>}**
Response: **<output><EOS>**

- ❖ **{instruction}**: A textual description of the task.
- ❖ **{<image>, <text>}**: Input of the data sample.
- ❖ **{output}**: Output if the data sample.
- ❖ **<BOS>**: The start input token.
- ❖ **<EOS>**: The end input token.

Methods

❖ M-IT Taxonomy



Methods

❖ M-IT Demo: LLaVA

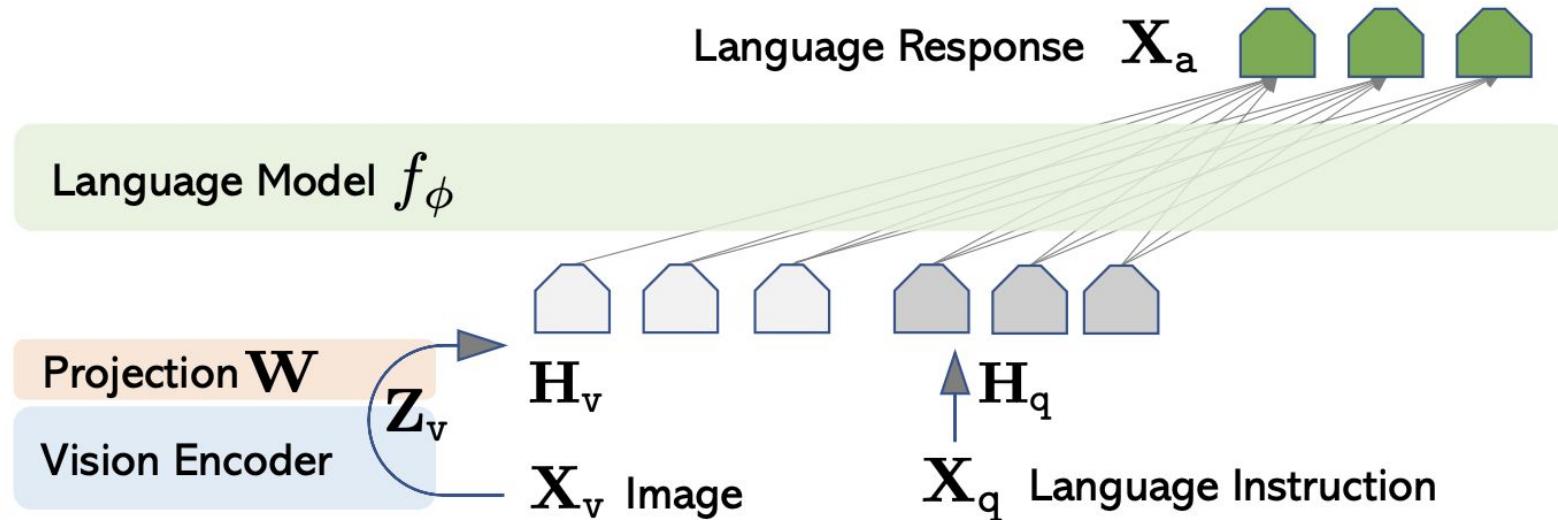


Figure 1: LLaVA network architecture.

Methods

❖ M-IT Demo: LLaVA

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>

Methods

❖ M-IT Demo: LLaVA

Visual input example, Extreme Ironing:



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

User
LLaVA

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as at home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

[Start a new conversation, and clear the history]

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User
LLaVA

Visual input example, Chicken Nugget Map:

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



User
LLaVA

Can you explain this meme in detail?

The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is.." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.

User
GPT-4 [36]

Can you explain this meme?

This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Methods

❖ M-IT Demo: LLaVA 1.5

 **LLaVA: Large Language and Vision Assistant**

[Project Page] [Code] [Model] | 📁 [LLaVA] [LLaVA-v1.5]

llava-v1.5-7b

Image

Drop Image Here
- or -
Click to Upload

⋮ Examples

 What is unusual about this image?

 What are the things I should be cautious about when I visit here?

LLaVA Chatbot



which language does the signboard in this image use?

The signboard in this image uses Vietnamese.

Enter text and press ENTER

Send

Parameters

◀

Upvote

Downvote

Flag

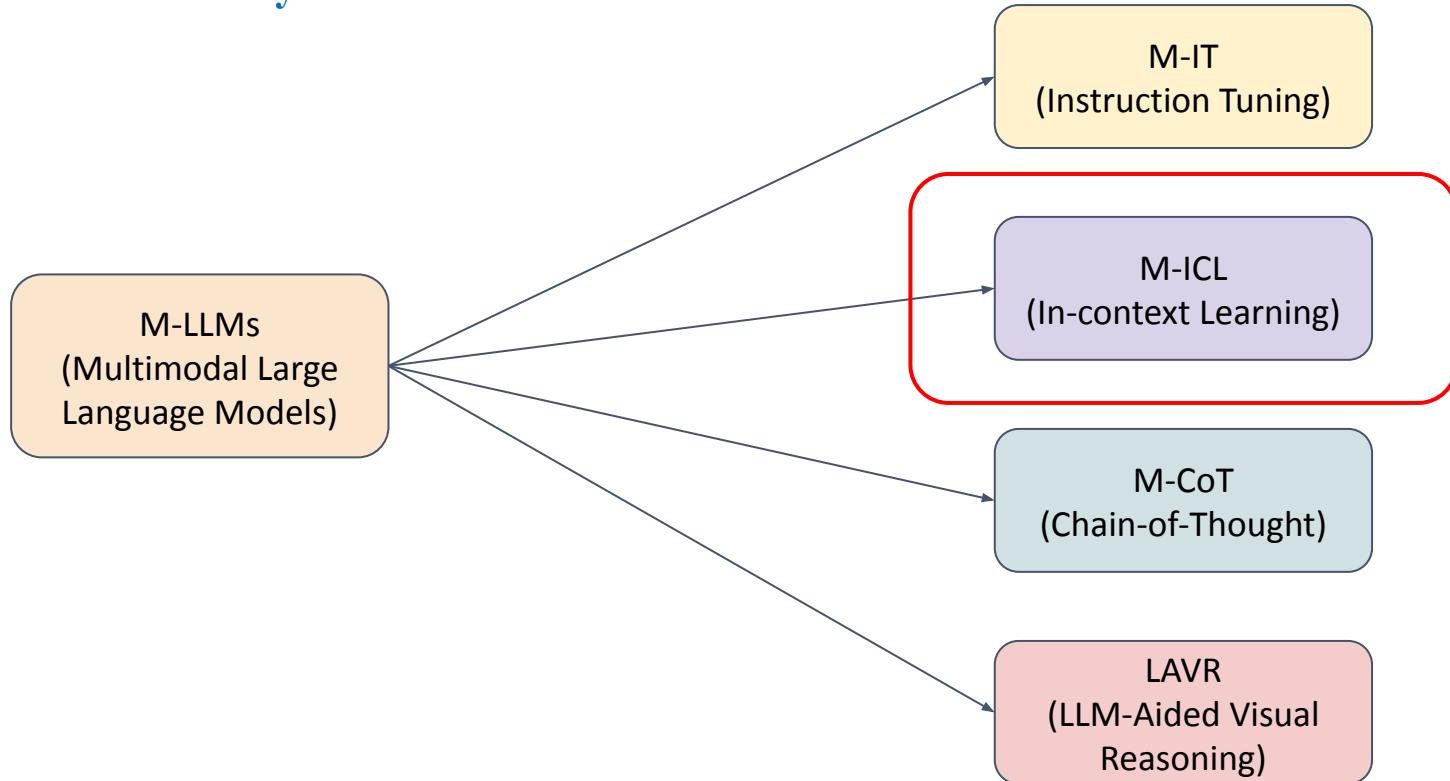
Regenerate

Clear

35

Methods

❖ M-LLMs Taxonomy



Methods

❖ M-ICL: Introduction

```
### Instruction: Classify the sentiment of the text  
below in terms of positive, negative or neutral.  
### Text: "This food is delicious!"  
### Response:
```



Zero-shot Prompting

Methods

❖ M-ICL: Introduction

Instruction: Classify the sentiment of the text below in terms of positive, negative or neutral.

Text: "This food is delicious!"

Response: "Positive"

Text: "It's raining outside. I bet we cannot go swimming today."

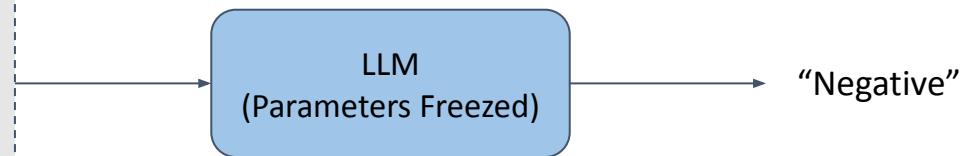
Response: "Negative"

Text: "Despite the bad service, this restaurant serves very tasty dishes."

Response: "Neutral"

Text: "The new employee is getting so mean, that is not a proper manner as a beginner."

Response:

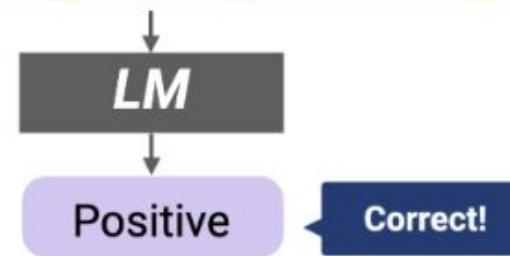


Methods

❖ M-ICL: Introduction

1,999 ×

Circulation revenue has increased by 5% in Finland. \n Positive
Panostaja did not disclose the purchase price. \n Neutral
Paying off the national debt will be extremely painful. \n Negative
The company anticipated its operating profit to improve. \n _____



In-context learning: A prompt engineering technique that give a list of input-output pairs for a specific task. At the end of the prompt, we give the test sample for the final result.

Methods

❖ M-ICL query structure example

<BOS> Below are some examples and an instruction that describes a task. Write a response that appropriately completes the request

Instruction: {instruction}

Image: <image>

Response: {response}

Image: <image>

Response: {response}

Image: <image>

Response: <EOS>

- ❖ **{instruction}**: A textual description of the task.
- ❖ **<image>**: A place-holder to represent an image.
- ❖ **{response}**: A response (answer) to the task related to **{instruction}** and **<image>**.
- ❖ **<BOS>**: The start input token.
- ❖ **<EOS>**: The end input token.

Methods

❖ M-ICL Demo: OpenFlamingo

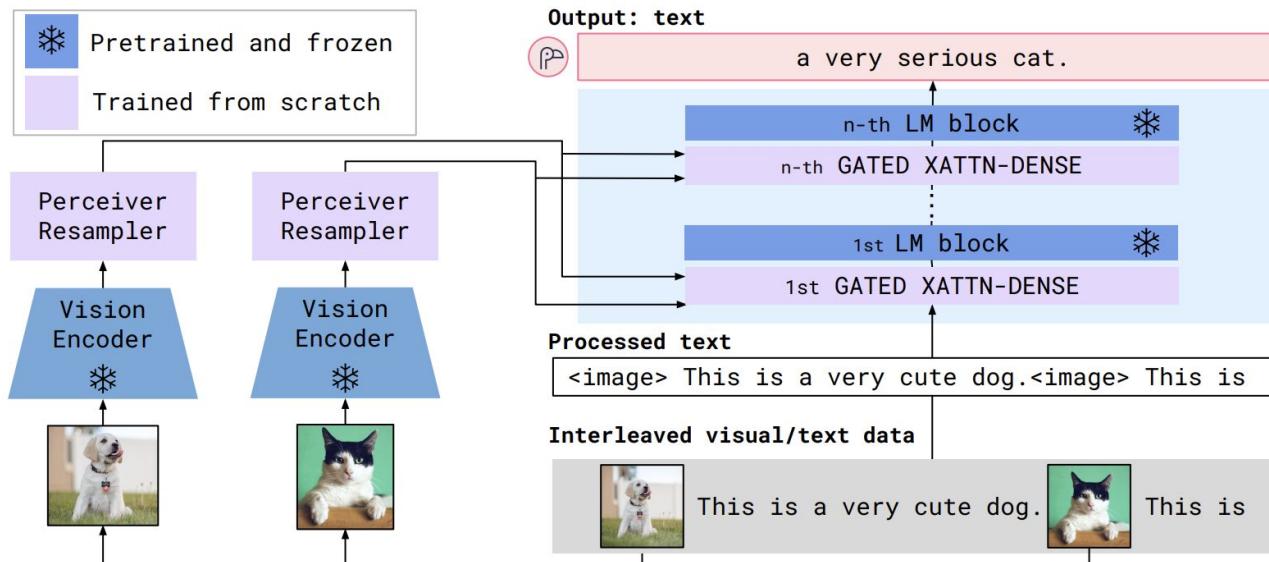


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

Methods

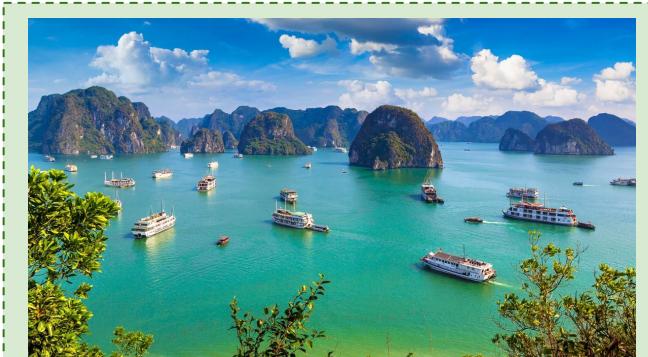
❖ M-ICL Demo: OpenFlamingo



This is an image of **shopping mall**.



This is an image of **Ancient Town**.



This is an image of
???

Methods

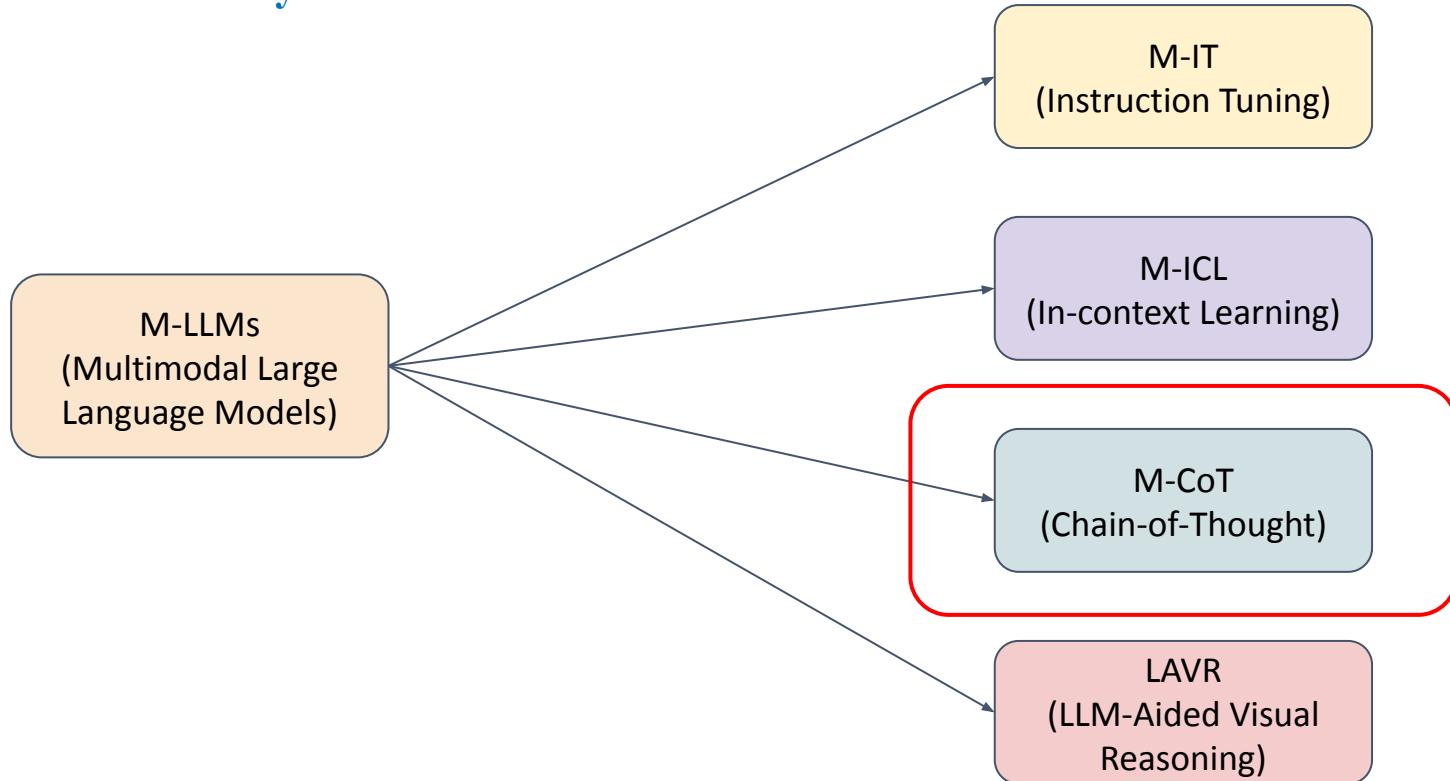
❖ M-ICL Demo: OpenFlamingo

```
● ● ●
1 vision_x = [image_processor(demo_image_one).unsqueeze(0),
2             image_processor(demo_image_two).unsqueeze(0),
3             image_processor(query_image).unsqueeze(0)]
4 vision_x = torch.cat(vision_x, dim=0)
5 vision_x = vision_x.unsqueeze(1).unsqueeze(0)
6
7
8 tokenizer.padding_side = "left" # For generation padding tokens should be on the left
9 lang_x = tokenizer(
10     ["<image>An image of shopping mall.<|endofchunk|><image>An image of an Ancient Town.<|endofchunk|><image>An image of"],
11     return_tensors="pt",
12 )
13
14
15 generated_text = model.generate(
16     vision_x=vision_x,
17     lang_x=lang_x["input_ids"],
18     attention_mask=lang_x["attention_mask"],
19     max_new_tokens=20,
20     num_beams=3,
21 )
22
23 print("Generated text: ", tokenizer.decode(generated_text[0]))
```

```
Flamingo model initialized with 1046992944 trainable parameters
Setting `pad_token_id` to `eos_token_id`:50277 for open-end generation.
Generated text: <image>An image of shopping mall.<|endofchunk|><image>An image of an Ancient Town.
<|endofchunk|><image>An image of Halong Bay.<|endofchunk|>
```

Methods

❖ M-LLMs Taxonomy



Methods

❖ M-CoT: Introduction

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27.

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

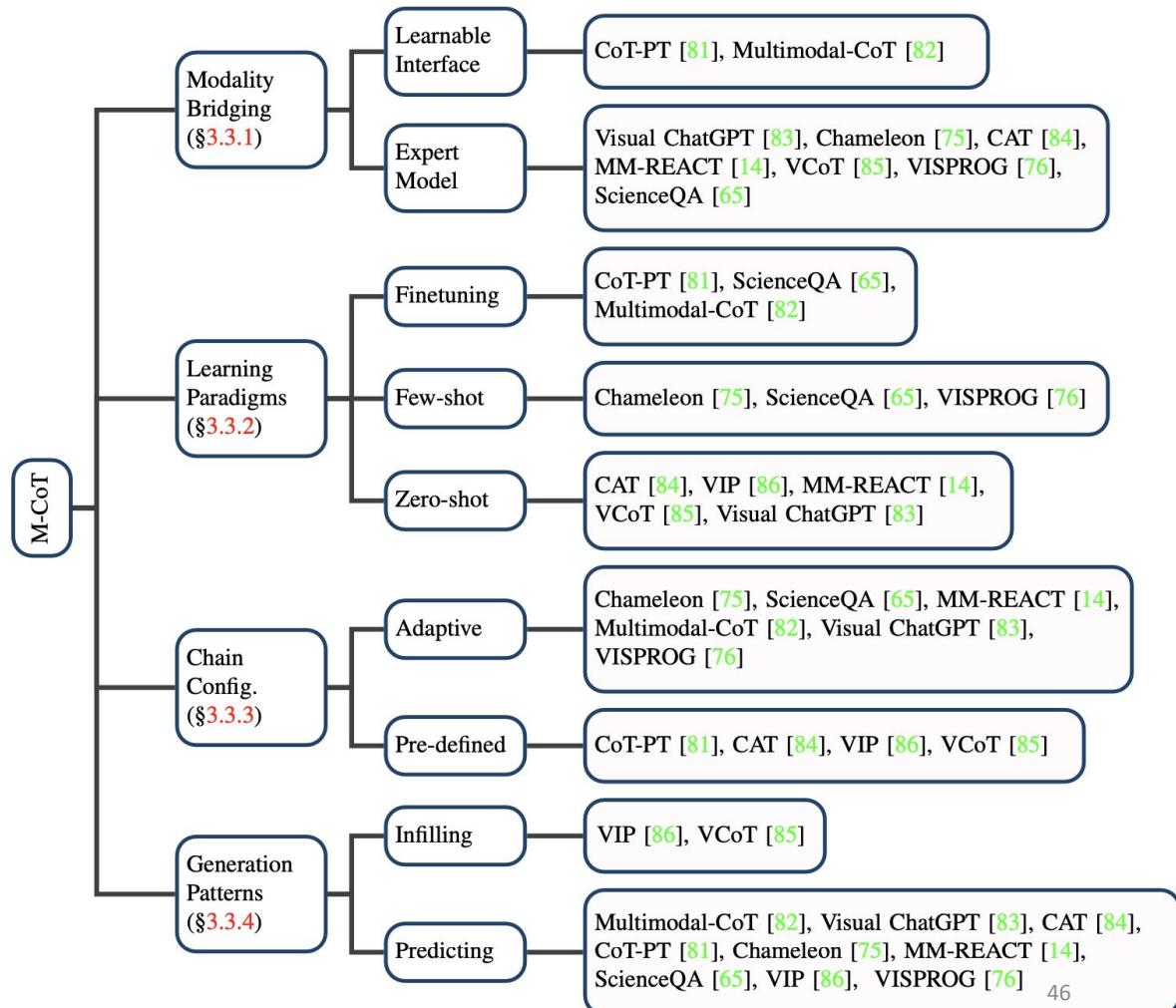
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.

Methods

❖ M-CoT Taxonomy



Methods

❖ M-CoT Demo: MM-CoT

Multimodal Chain-of-Thought Reasoning in Language Models

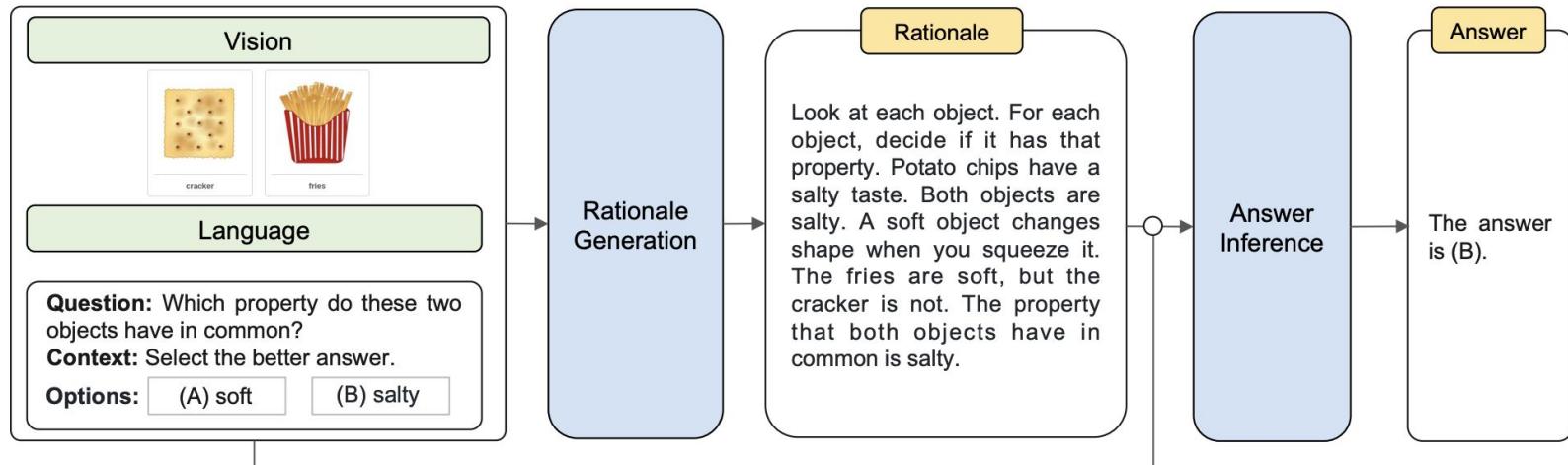


Figure 4. Overview of our Multimodal-CoT framework. Multimodal-CoT consists of two stages: (i) rationale generation and (ii) answer inference. Both stages share the same model architecture but differ in the input and output. In the first stage, we feed the model with language and vision inputs to generate rationales. In the second stage, we append the original language input with the rationale generated from the first stage. Then, we feed the updated language input with the original vision input to the model to infer the answer.

Methods

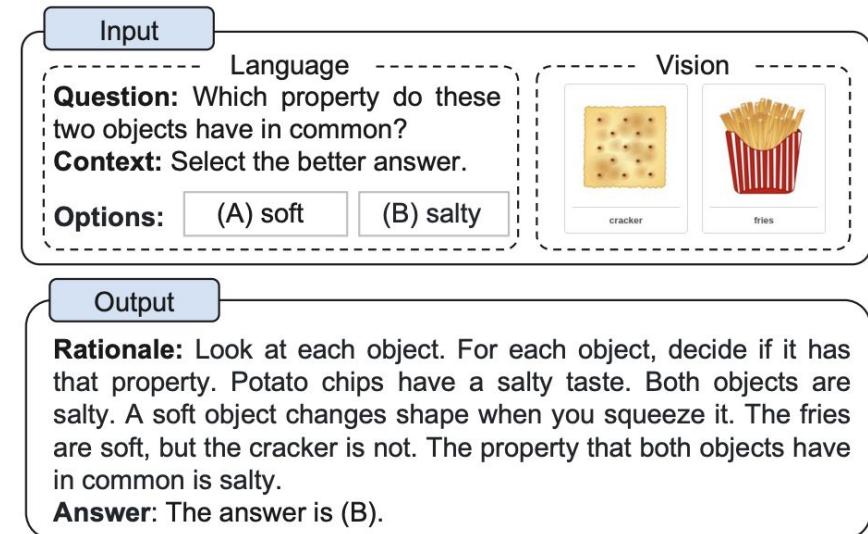
❖ M-CoT Demo: MM-CoT

Algorithm 1 Multimodal-CoT

Input: Language input X_{language}^1 , vision input X_{vision}

Output: Generated rationale R , inferred answer A

- 1: Construct the input $X = \{X_{\text{language}}, X_{\text{vision}}\}$
- 2: Generate rationale $R = F(X)$ using the model $F(\cdot)$
- 3: Append the rationale R to the original language input $X_{\text{language}}^2 = X_{\text{language}}^1 \circ R$.
- 4: Construct new input $X' = \{X_{\text{language}}^2, X_{\text{vision}}\}$
- 5: Infer the answer A by conditioning on the new input, $A = F(X')$.
- 6: **procedure** $F(X)$
- 7: Encode the language and vision inputs H_{language} and H_{vision} , respectively
- 8: Build the interaction between language and vision features by attention $H_{\text{vision}}^{\text{attn}}$
- 9: Fuse H_{language} and $H_{\text{vision}}^{\text{attn}}$ by a gated fusion mechanism to have H_{fuse}
- 10: Feed H_{fuse} to the decoder to obtain the target prediction Y
- 11: **return** Y
- 12: **end procedure**



Methods

❖ M-CoT Demo: MM-CoT



Input Image

```
37
38 question = "Between two objects in the image, which one is stronger?"
39 context = "Select the better answer."
40 choice = "(A) Lion (B) Human"
41 img_path = "./images/Human_vs_Lion.jpg"
42
43
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
(thuan_torch) server-ailab-12gb@serverailab12gb-System-Product-Name:/mnt/HDD/ThangDuongTeam/thuannd
● /mm-cot$ python predict.py
Solution: Look at the objects in the image.nFor each object, decide if it is stronger.nThe lion is
stronger than the other object.nThe other object is not strong. But the lion is stronger.
Answer: The answer is (A).
```

Methods

❖ M-CoT Demo: MM-CoT



Input Image

```
37
38 question = "Between two liquids in the image, which one is better for health?"
39 context = "Select the better answer."
40 choice = "(A) Left (B) Right"
41 img_path = "./images/water_vs_coca.jpg"
42
43
(thuan_torch) server-ailab-12gb@serverailab12gb-System-Product-Name:/mnt/HDD/ThangDuongTeam/thuann
● /mm-cot$ python predict.py
Solution: Look at the picture of the two liquids.nThe two liquids are mostly black, but they are no
t actually health.nSo, the right liquid is better for health.
Answer: The answer is (B).
```

Methods

❖ M-CoT Demo: MM-CoT



Input Image

```
37
38     question = "Between two liquids in the image, which one is better for health?"
39     context = "Select the better answer."
40     choice = "(A) Water (B) Coca"
41     img_path = "./images/water_vs_coca.jpg"
42
43
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS
(thuan_torch) server-ailab-12gb@serverailab12gb-System-Product-Name:/mnt/HDD/ThangDuongTeam/thuann
● /mm-cot$ python predict.py
Solution: Look at the picture of the two liquids. The two liquids are mostly black, but they are no
t actually health. So, the water is better for health.
Answer: The answer is (A).
```

Methods

❖ M-CoT Demo: MM-CoT



Input Image

```
38 question = "Between two foods on the table, which one is better for health?"  
39 context = "Select the better answer."  
40 choice = "(A) Left (B) Right"  
41 img_path = "./images/food.png"  
42  
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS  
  
(thuan_torch) server-ailab-12gb@serverailab12gb-System-Product-Name:/mnt/HDD/ThangDuongTeam/thuannd/mm-cot$ python predict.py  
Solution: Look at the picture of the table. It is shown that all food is unhealthy. But a diet of unhealthy food does not have all the health benefits of a healthy food. So, it is better for health.  
Answer: The answer is (B).
```

Methods

❖ M-CoT Demo: MM-CoT

$$\frac{x+3}{6} = \frac{2}{3}$$

```
37
38 question = "Solve for X"
39 context = "Select the better answer."
40 choice = "(A) -1 (B) 1"
41 img_path = "./images/math.png"
42
43
```

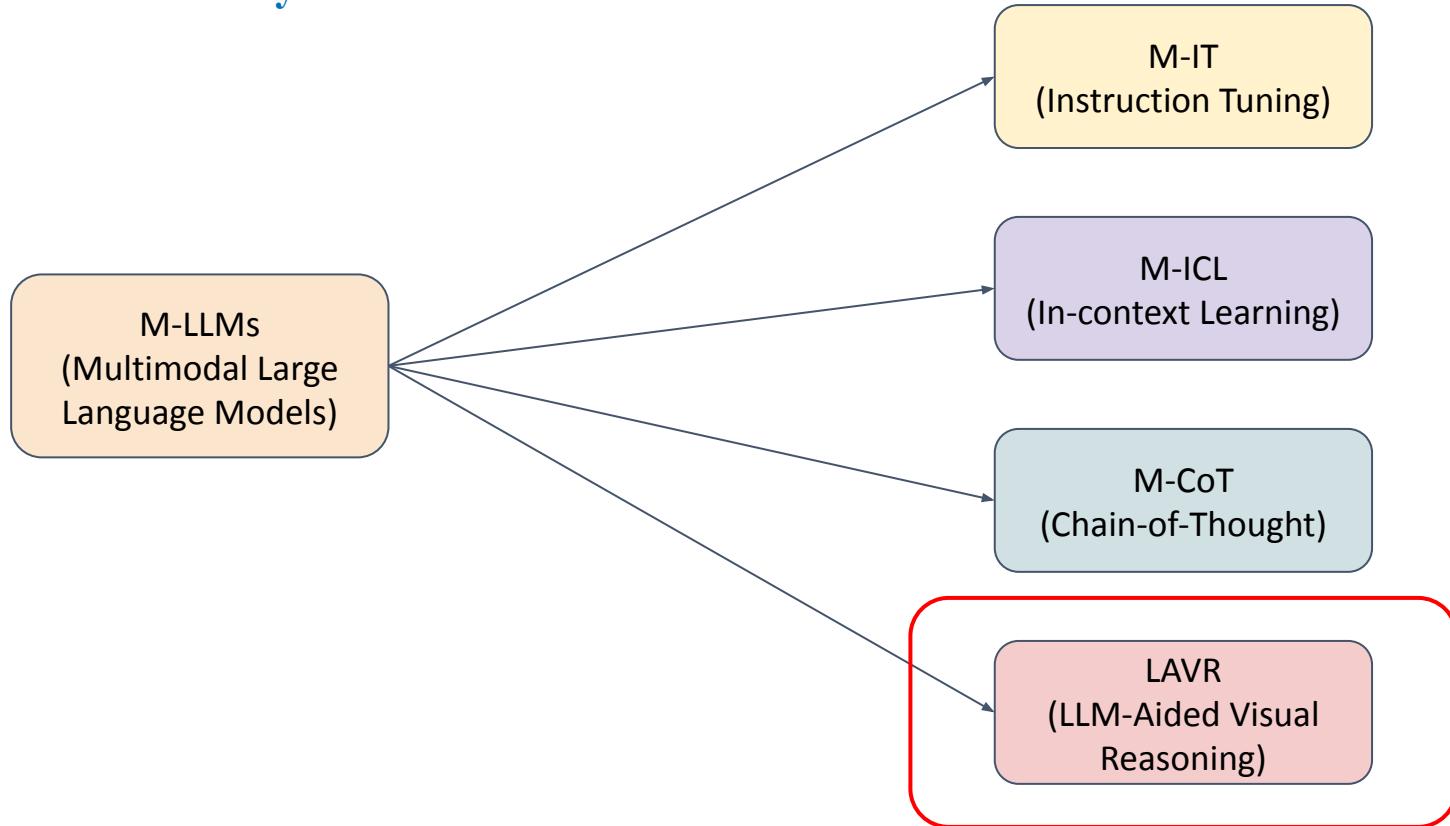
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

(thuan_torch) server-ailab-12gb@serverailab12gb-System-Product-Name:/mnt/HDD/ThangDuongTeam/thuannd
● /mm-cot\$ python predict.py
Solution: The better answer is -1.nThe better answer is -1.nThere is 1 left before the answer choic
e. So, 1 must be a right answer.
Answer: The answer is (A).

Input Image

Methods

❖ M-LLMs Taxonomy



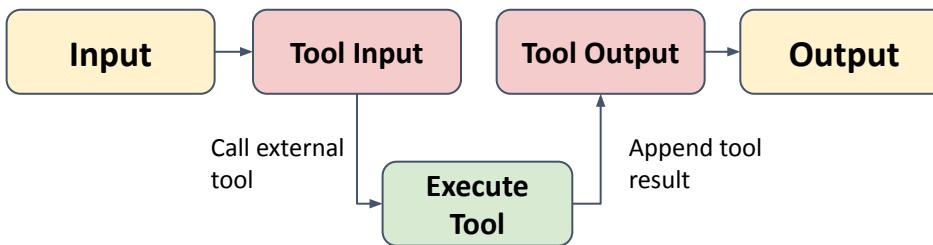
Methods

❖ LAVR: Introduction

Large Language Models (LLMs)



Tool-augmented LLMs



Tool-augmented LLMs: Refer to a cutting-edge approach where LLMs are equipped with the capability to interact with external tools or interfaces to extend their functionality beyond mere text generation

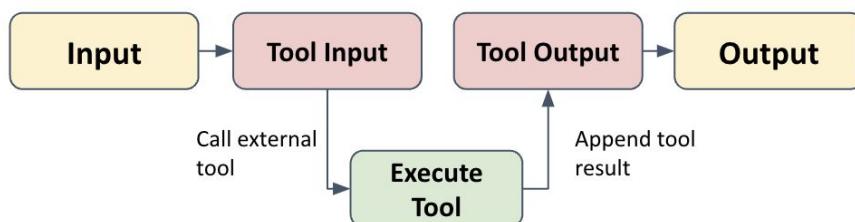
Methods

❖ LAVR: Introduction

Large Language Models (LLMs)



Tool-augmented LLMs



You search paper about multimodal instruction tuning

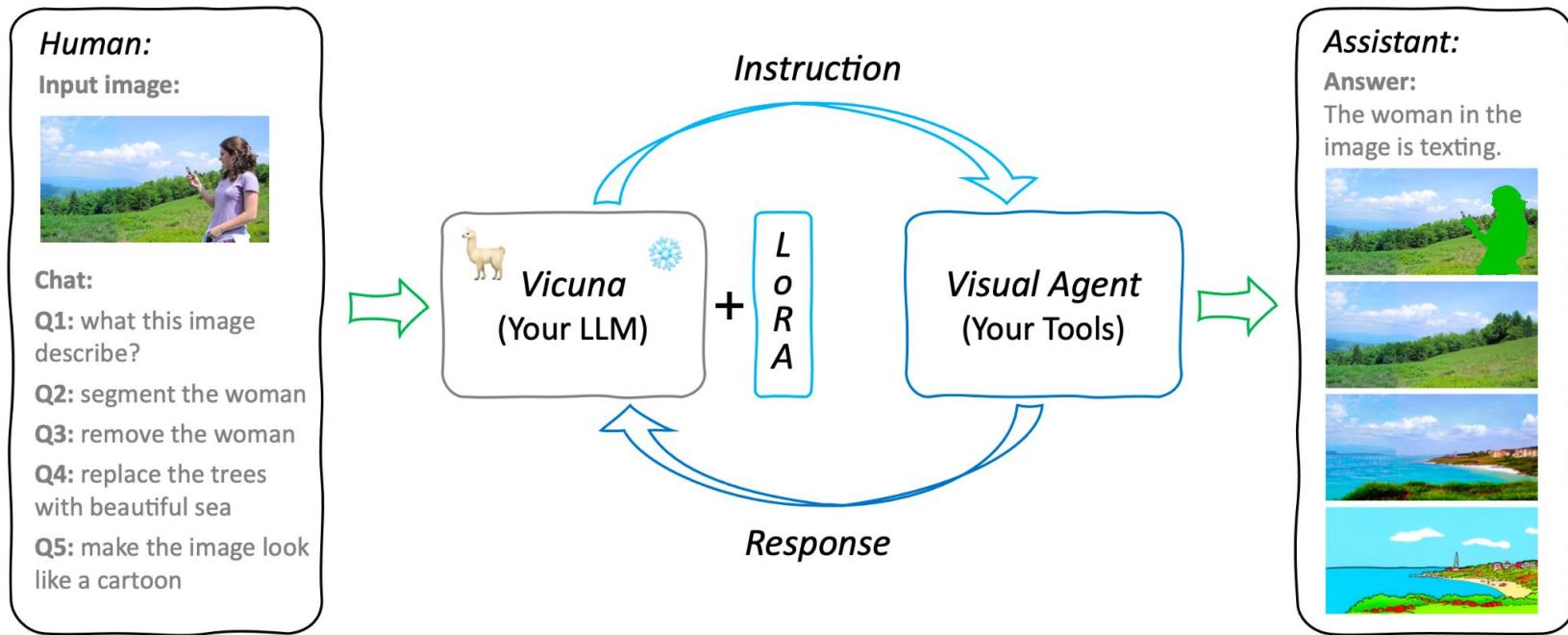
ChatGPT Used Paperpile

Here are some research papers related to multimodal instruction tuning:

1. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models
 - **Authors:** Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, Yang Liu
 - **Abstract:** This paper introduces Position-enhanced Visual Instruction Tuning (PVIT), enhancing Multimodal Large Language Models (MLLMs) with a region-level vision encoder for more detailed image comprehension. It includes data generation strategies for image-region-language instruction datasets and demonstrates the model's superiority through experiments and analysis.
 - **Year:** 2023
 - [Read more](#)

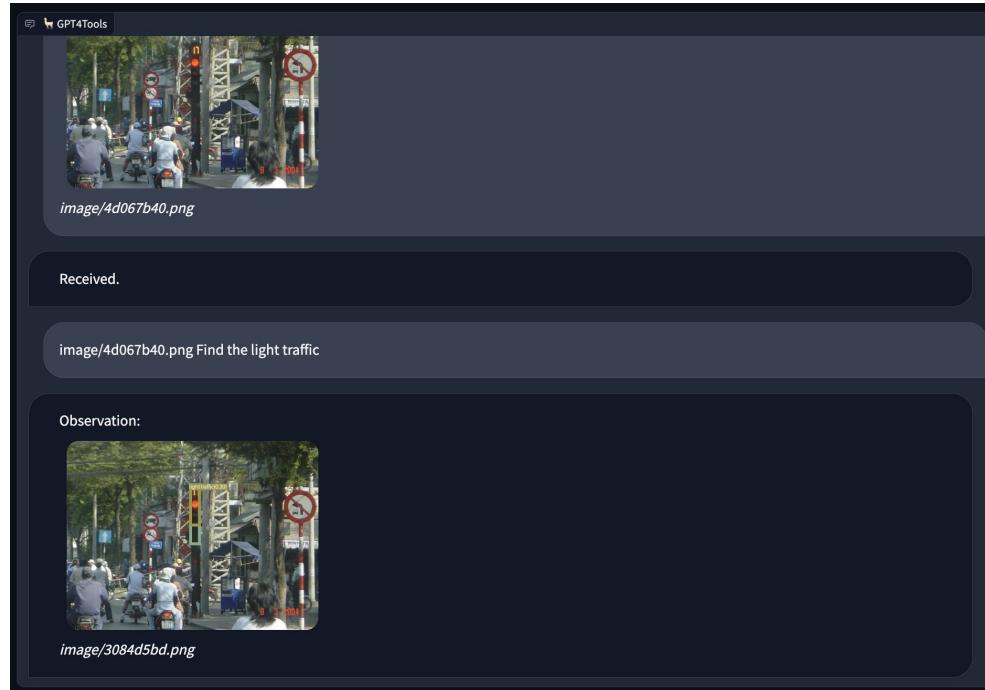
Methods

❖ LAVR: Introduction



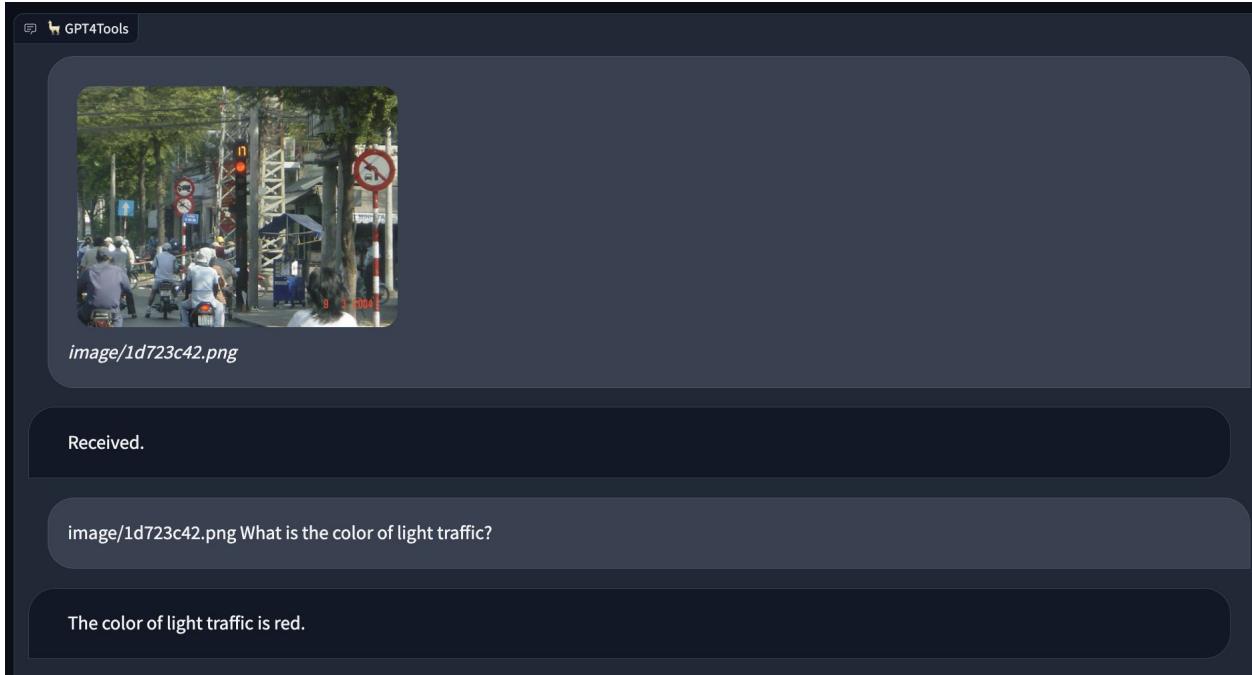
Methods

❖ LAVR: GPT4Tools Demo



Methods

❖ LAVR: GPT4Tools Demo



Methods

❖ LAVR: GPT4Tools Demo

GPT4Tools

image/1d723c42.png Segment the sign traffic.

Observation:



image/1f65af05.png

Methods

❖ LAVR: GPT4Tools Demo

The screenshot shows a dark-themed interface for GPT4Tools. At the top, there's a small thumbnail of a photo and the text "GPT4Tools". Below that is a larger thumbnail of two people eating at a table outdoors. Underneath the thumbnail is the file name "image/868cec6a.png".

The main area has a dark background with white text. It starts with a "Received." message, followed by a user input: "image/868cec6a.png image/7a3afe15.png This is an image about eating breakfast in Viet Nam. Tell me a story about this photo."

Three生成的文本响应如下：
1. In this photo, we see a group of people gathered around a table, enjoying their breakfast. The table is set with a variety of dishes, including bowls of rice, plates of fried eggs, and cups of tea. The people are all dressed in traditional Vietnamese clothing, and they are laughing and chatting with each other.
2. The scene is set in a small, open-air restaurant in Vietnam. The restaurant is located in a bustling city, and the people in the photo are all locals. They have come together to enjoy a meal together, and the atmosphere is warm and welcoming.
3. As they eat, the people in the photo share stories and laughter, and they take the time to appreciate the simple pleasures of life. They are grateful for the food on their plates, and they are grateful for the company of their friends and family.
4. This photo captures the essence of Vietnamese culture, which values the importance of community and family. It shows that, no matter where you are in the world, people come together to share meals and stories, and to enjoy the simple pleasures of life.

Conclusion

❖ Content

- ❖ **Performed a survey of the existing M-LLMs literature.**
- ❖ **Main directions:** M-IT, M-ICL, M-CoT and LAVR.
- ❖ **Applications:** Knowledge-based VQA, OCR-Free eKYC...
- ❖ **Limitation:**
 - Still limited in perception capabilities, leading to wrong visual information.
 - The reasoning chain of MLLMs are still fragile in some cases.
 - The instruction-following ability needs upgrading.
 - The object hallucination issues.
 - Parameter-efficient training is needed.

