

JUNE 18-22, 2023



Microsoft

# Large Multimodal Models

Towards Building and Surpassing Multimodal GPT-4

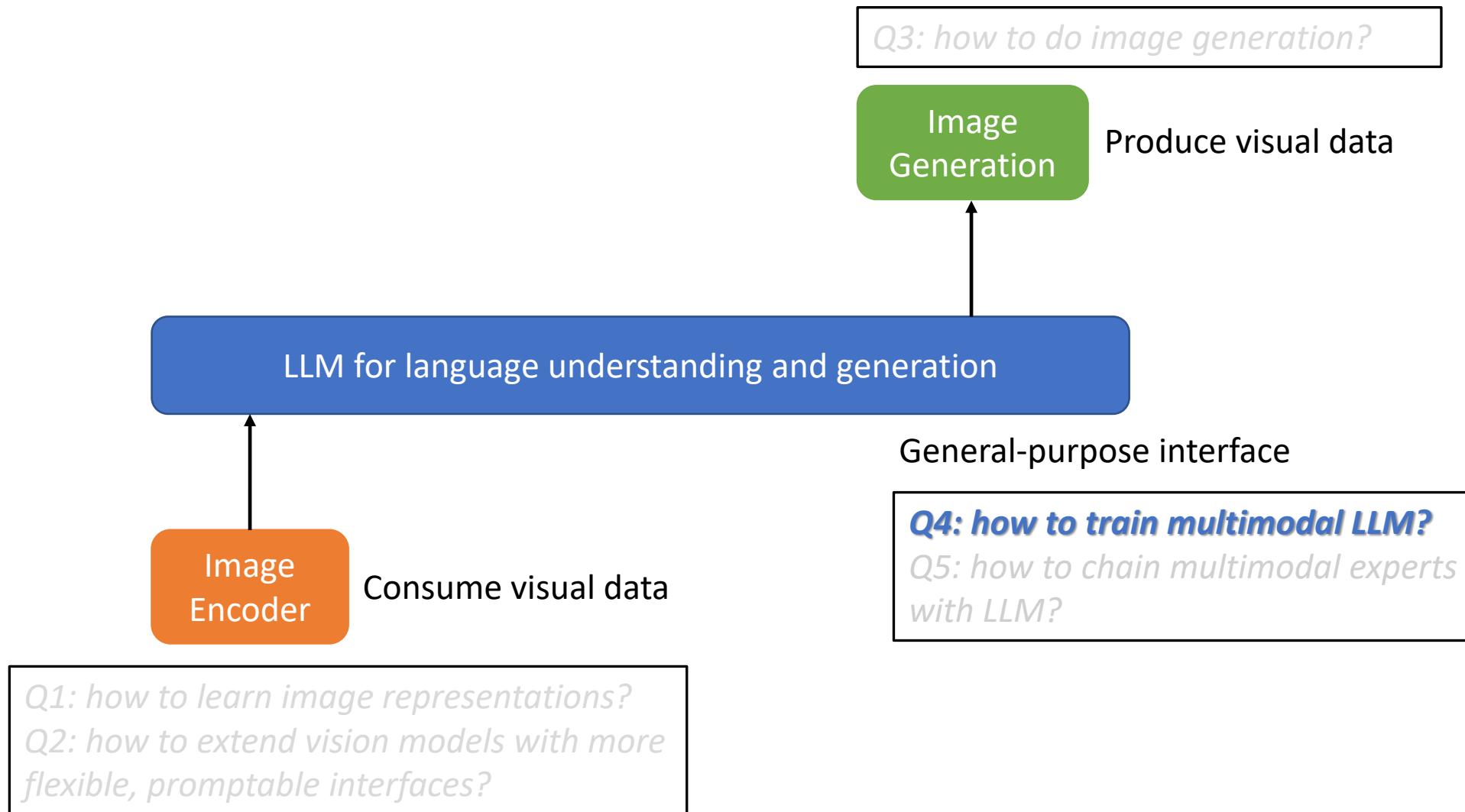
June 2023

Chunyuan Li

Deep Learning Team

Microsoft Research, Redmond

<https://chunyuan.li>



## Outline

### 1 Background

- Traditional Large Multimodal Models (LMM)
- Multimodal GPT4

### 2 Pre-requisite: Instruction Tuning in Large Language Models

### 3 Instructed Tuned Large Multimodal Models

- **Open-Source Prototypes: LLaVA / MiniGPT4**
- Emerging Extensions
- Benchmarking, Properties and Applications

1

Background:

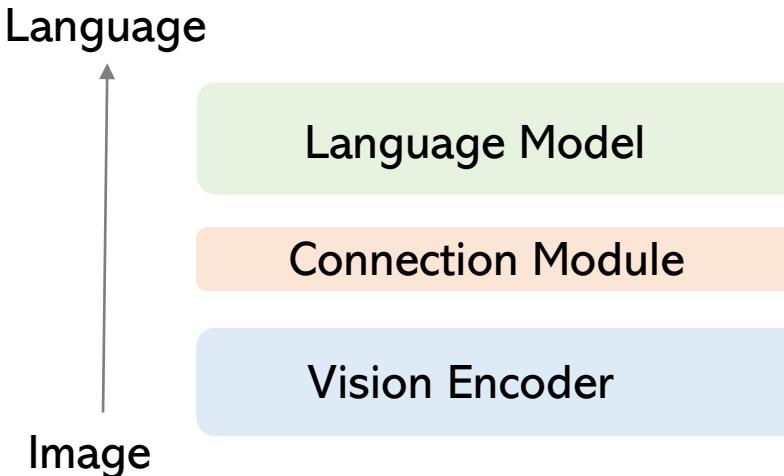
**Large Multimodal Models (LMM)  
& GPT4**

# Large Multimodal Models: Image-to-Text Generative Models

## ❑ Model Architectures

- (Pre-trained) Image Encoder and Language Models
- Trainable modules to connect to two modalities

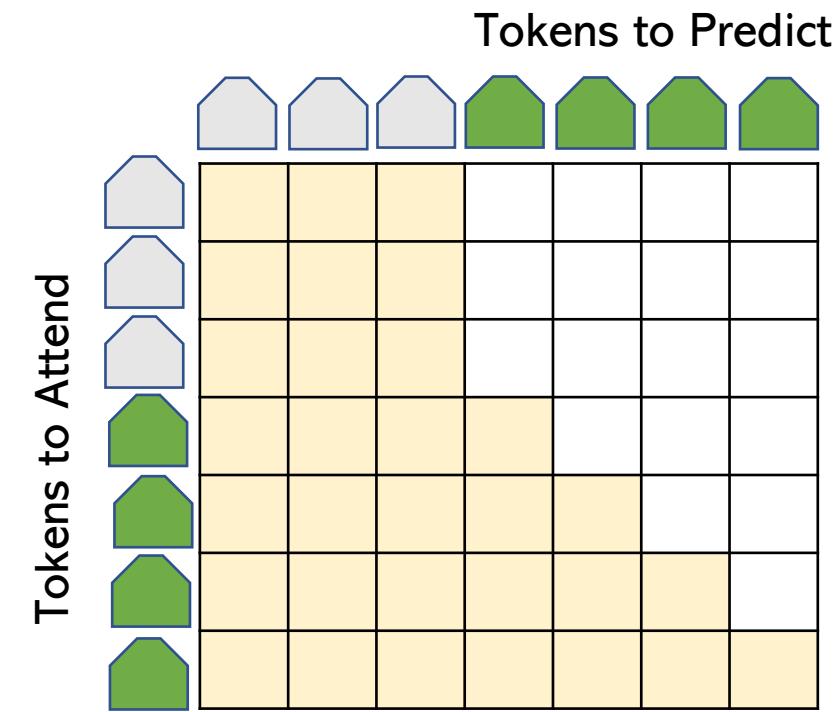
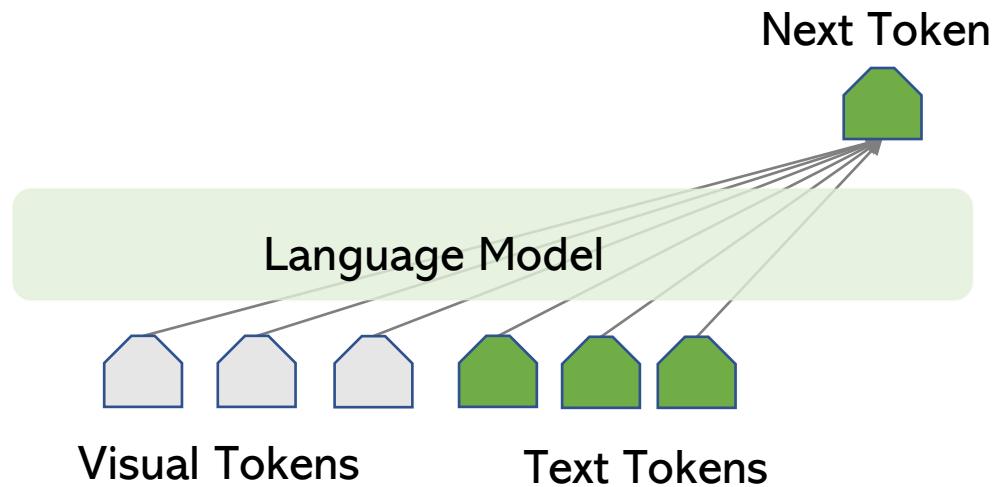
A dog lying on the grass next to a frisbee



# Large Multimodal Models: Image-to-Text Generative Models

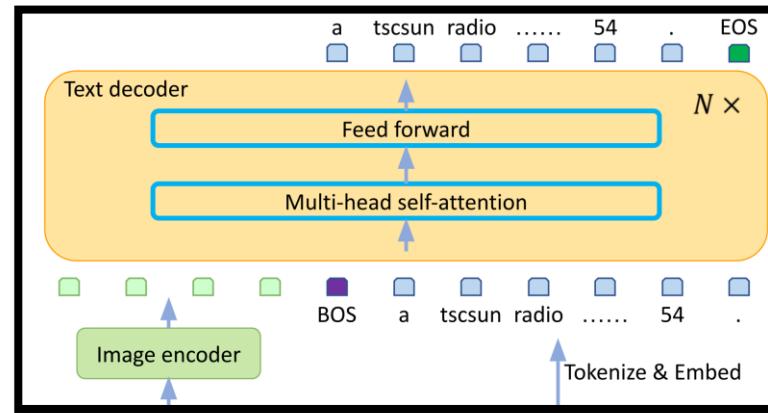
## Training Objective

- Cross-Attended Image-to-Text Generation
  - Autoregressive loss on language output

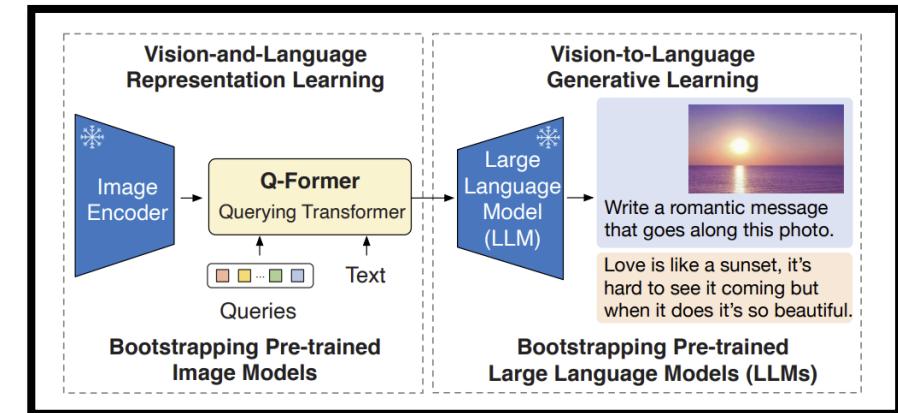


# Example 1: LMM with Image-Text Pairs

- GIT



- BLIP2



Language Model

Connection Module

Vision Encoder

From Scratch

Pre-trained: FLAN-T5/OPT

Q-Former: Lightweight  
Querying Transformer

Contrastive pre-trained:  
Florence/CLIP

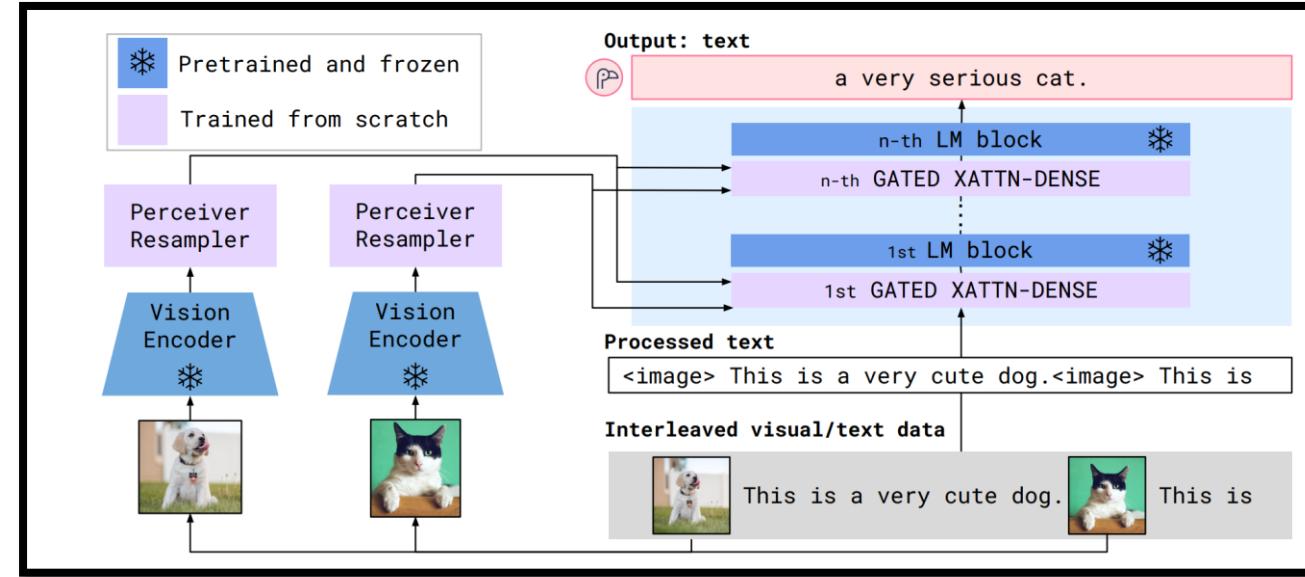
Contrastive pre-trained:  
EVA/CLIP

GIT: A Generative Image-to-text Transformer for Vision and Language

Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models

## Example 2: LMM with Interleaved Image-Text Data

- Flamingo:



Language Model

Connection Module

Vision Encoder

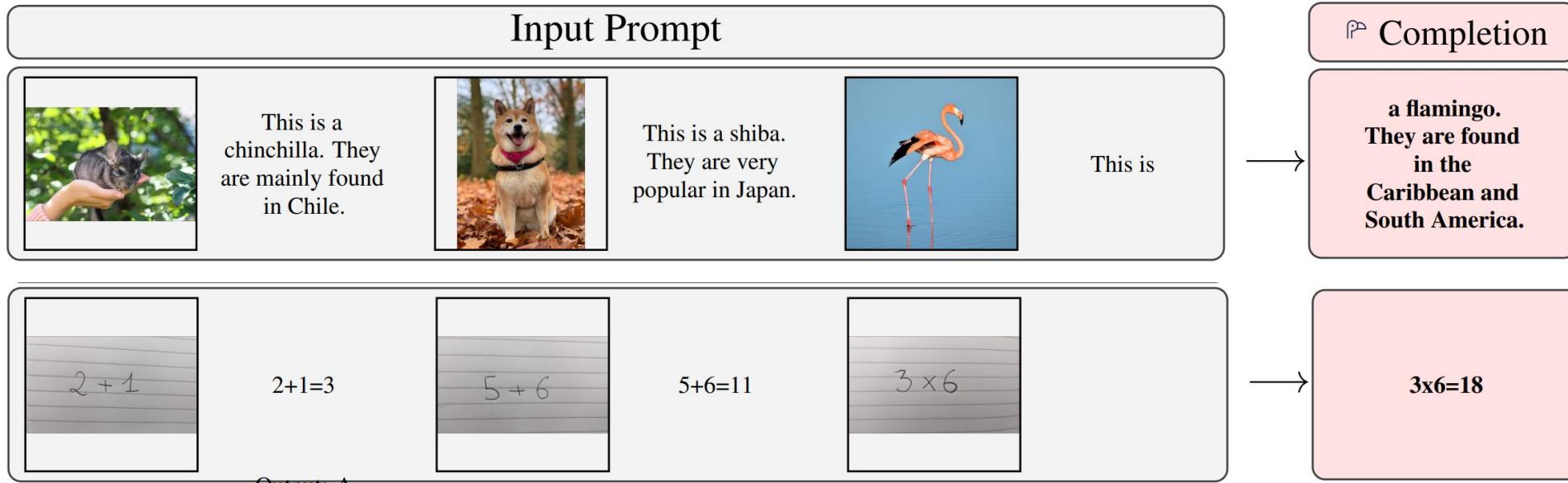
Pre-trained: 70B Chinchilla

Perceiver Resampler  
Gated Cross-attention + Dense

Pre-trained: Nonnormalizer-Free ResNet (NFNet)

# Example 2: LMM with Interleaved Image-Text Data

- Flamingo: Multimodal In-Context-Learning



- Model Details: Unknown
- Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild
- How can we build Multimodal GPT-4 like models?

---

**GPT-4 visual input example, Extreme Ironing:**


---

User      What is unusual about this image?



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

GPT-4      The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

---



---

**GPT-4 visual input example, Chicken Nugget Map:**


---

User      Can you explain this meme?

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



GPT-4      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.

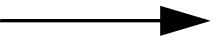
---

# Recap on Language Modeling: Large Language Models (LLM)



**What's new?**

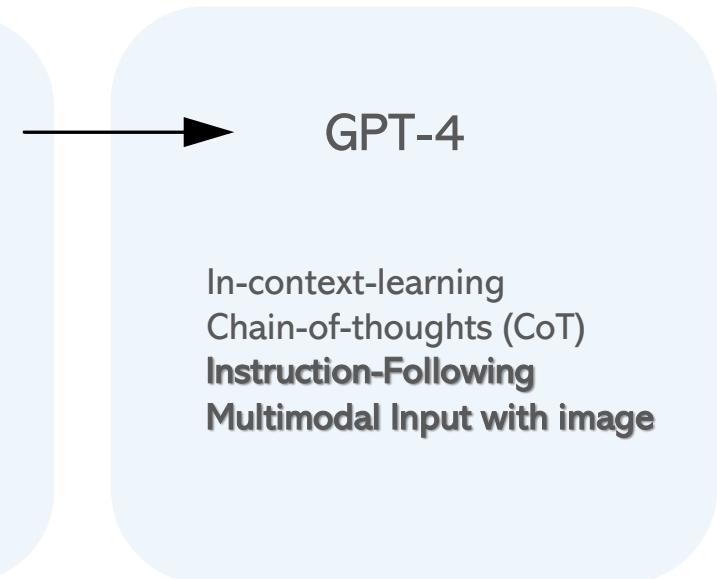
GPT-2



GPT-3



ChatGPT  
InstructGPT



GPT-4

In-context-learning  
Chain-of-thoughts (CoT)

In-context-learning  
Chain-of-thoughts (CoT)  
**Instruction-Following**

In-context-learning  
Chain-of-thoughts (CoT)  
**Instruction-Following**  
**Multimodal Input with image**

**Multimodal  
Space**

Flamingo  
BLIP2  
GIT

...

**Gap?**  
**Instruction-Following**  
→ Alignment Research

Multimodal GPT-4

②

Pre-requisite:

# Instruction Tuning in Large Language Models (LLM)

# Instruction Tuning

Translation

Input → Output

*Hello, Vancouver*

你好，温哥华

Summarization

*CVPR is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. This year, CVPR will be single track such that everyone (with full passport registration) can attend everything.*

*CVPR: top computer vision event, single-track, accessible to all.*

- Task instructions are implicit.
- Individual models are trained, or multi-tasking without specifying the instructions
- Hard to generalize to new tasks in zero-shot

# Instruction Tuning

## Instruction

Translate English into Simplified Chinese

Summarize in just 10 words to make the message even more brief and easier to remember.

Input →

*Hello, Vancouver*

Output

你好，温哥华

*CVPR is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. This year, CVPR will be single track such that everyone (with full passport registration) can attend everything.*

*CVPR: top computer vision event, single-track, accessible to all.*

- Task instructions are explicit, expressed in natural language
- One single model is trained, multi-tasking with specified instructions
- Natural and easy to generalize to new tasks in zero-shot



# Instruction Tuning

Instruction

Summarize in Chinese to make it easier to remember.

Input



Output



## CVPR 2023 Visas

The CVPR'23 organizing committee deeply regrets that many members of our community could not receive visas to attend CVPR 2023. For several months, the organizers have actively raised concerns with Canadian immigration authorities (IRCC), government agencies, and politicians. In some cases, we have been successful in helping people obtain visas, but in many cases, our efforts were unsuccessful. The organizers acknowledge that the international representation of members from all over the world is what has made CVPR successful. We share in the frustration of those who were unable to attend. We continue to allow virtual to in-person registration switches for attendees who receive their visas before the conference.

CVPR'23 签证问题: 组委会努力解决, 提供虚拟和现场注册转换服务

*"CVPR'23 visa issue: organizing committee works to solve and provide virtual and in-person registration switch services."*

# Self-Instruct Tuning

How to collect a diverse set of high-quality instructions and their responses?

- ❑ Human-Human: Collected from humans with high cost
- ❑ Human-Machine: A Strong LLM Teacher such as GPT3 and GPT4

*translation example      summarization example*

Please generate new instructions that meet the requirements: ....

Seed Examples —————→ In-Context Learning —————→ New Machine-Generated Examples



# Language Modeling: Large Language Models (LLM)



GPT-2

GPT-3

ChatGPT  
InstructGPT

GPT-4

**What's new?**

In-context-learning  
Chain-of-thoughts (CoT)

In-context-learning  
Chain-of-thoughts (CoT)  
**Instruction-Following**

In-context-learning  
Chain-of-thoughts (CoT)  
**Instruction-Following**  
**Multimodal Input with image**

Open Source  
Community

LLaMA



Alpaca



Vicuna



GPT4-Alpaca



Tulu



## Instruction Tuning with Open-Source LLMs

### Self-Instruct with Strong Teacher LLMs & Mixed Human Data

	LLaMA 	Alpaca 	Vicuna 	GPT4-Alpaca 	...	Tulu 
Data Source		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)	...	Mixed Data
Instruction-following Data (#Turns)	None	52K	500K (~150K conversions)	52K	...	

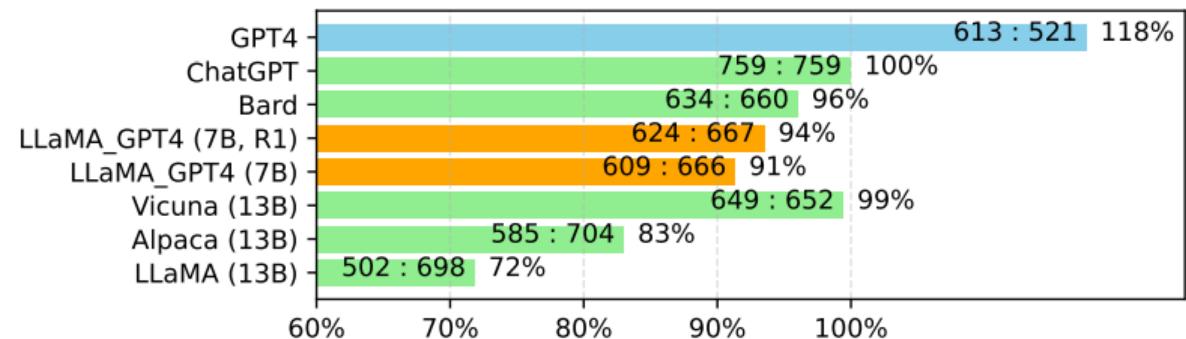
# Results on Chatbot

**Evaluation Metric:** Ask GPT-4 to rate the two model responses (1-10), then compute the ratio, i.e. relative score

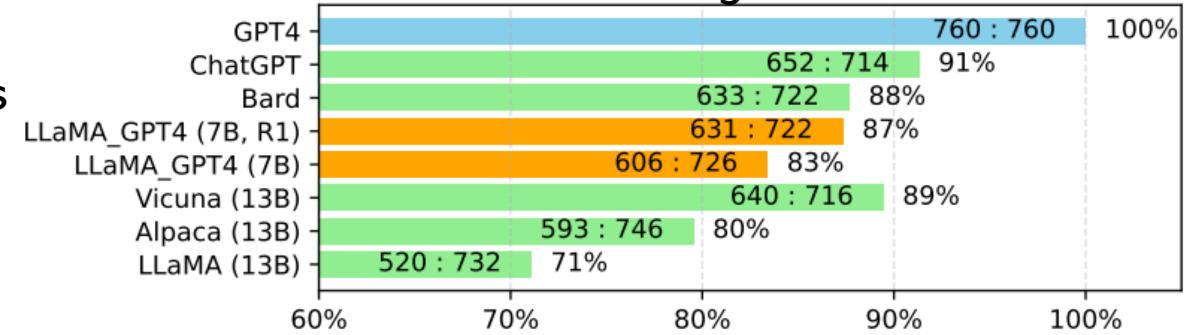
## Findings:

- A VERY CONSISTENT Evaluation Metric !
- Opensourced Chatbots mimicked commercial ones

All chatbots against ChatGPT



All chatbots against GPT-4



# Large Multimodal Models

-- Building multimodal gpt4 with open-source resources

LLaVA as a running example in this lecture

- Data
- Model
- Performance

# Visual Instruction Tuning with GPT-4

<https://llava-vl.github.io/>

Haotian Liu\*, Chunyuan Li\*, Qingyang Wu, Yong Jae Lee (\* Equal contribution)

**Self-Instruct with Strong Teacher LLMs**

**But No Teacher is available on multiGPT4?**

	LLaMA	Alpaca	Vicuna	GPT-4-LLM	LLaVA
Teacher					
Instruction-following Data	None	52K	700K (70 conversions)	GPT-4 (text-only)	GPT-4 (text-only)

- 158K multimodal instruction following data  
**(First & High Quality)**

→ Multimodal Chatbot

**Large Language and Vision Assistant**

# GPT-assisted Visual Instruction Data Generation

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples  
→ Text-only GPT-4

## Context type 1: Captions

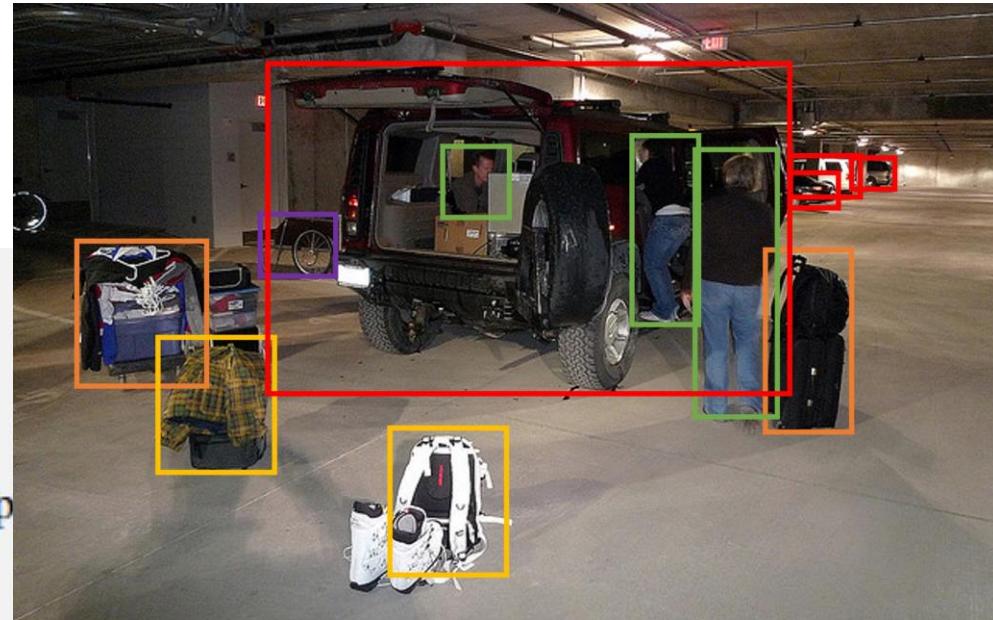
A group of people standing outside of a black vehicle with various luggage.  
Luggage surrounds a vehicle in an underground parking area

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], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]



# GPT-assisted Visual Instruction Data Generation

## Three type of instruction-following responses

### Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?

Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?

Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

### 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.

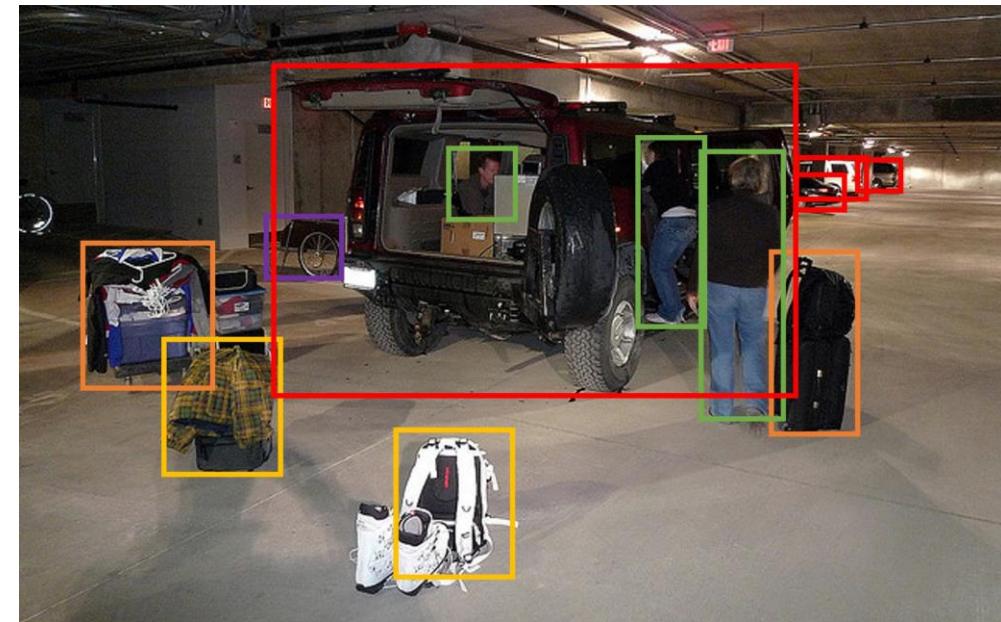
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

### Response type 3: complex reasoning

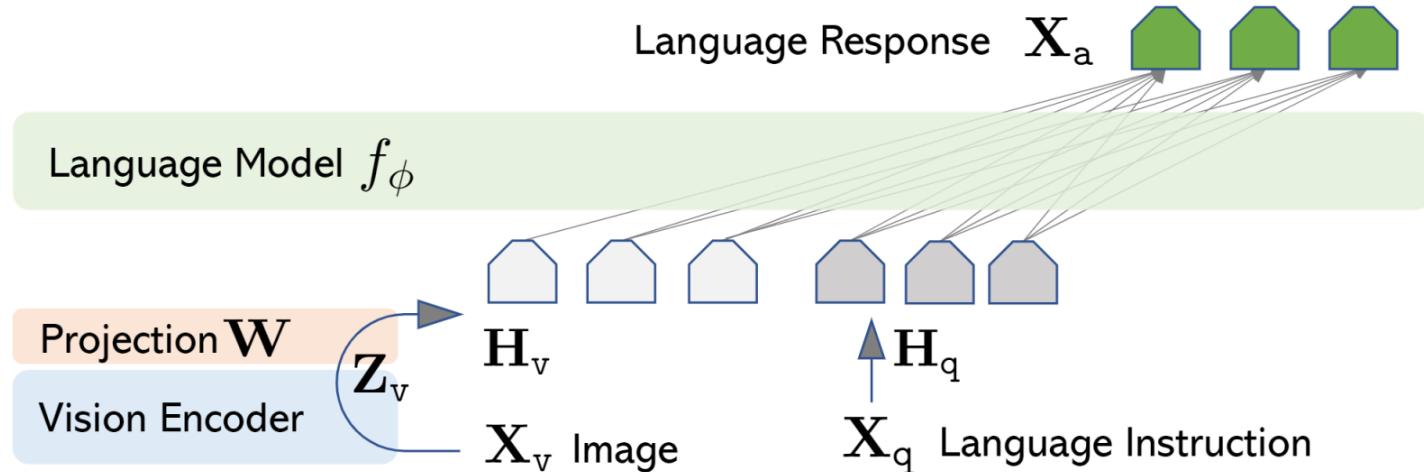
Question: What challenges do these people face?

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 to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.



# LLaVA: Large Language-and-Vision Assistant

## □ Architecture



## □ Two-stage Training

- **Stage 1: Pre-training for Feature Alignment.**

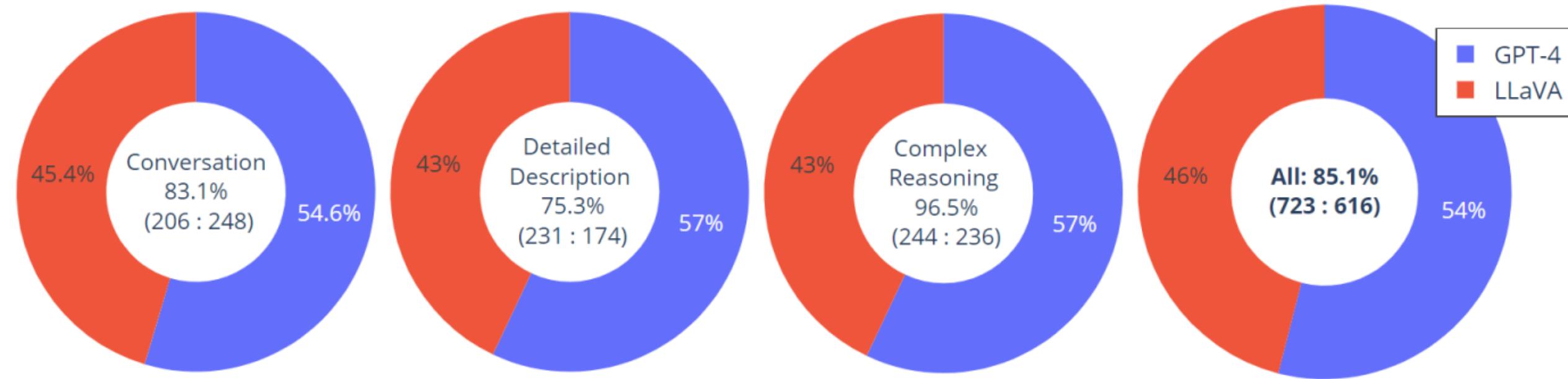
Only the projection matrix is updated, based on a subset of CC3M.

- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are updated

- **Visual Chat:** Our generated multimodal instruction data for daily user-oriented applications.

- **Science QA:** Multimodal reasoning dataset for the science domain.

# Visual Chat: Towards building multimodal GPT-4 level chatbot

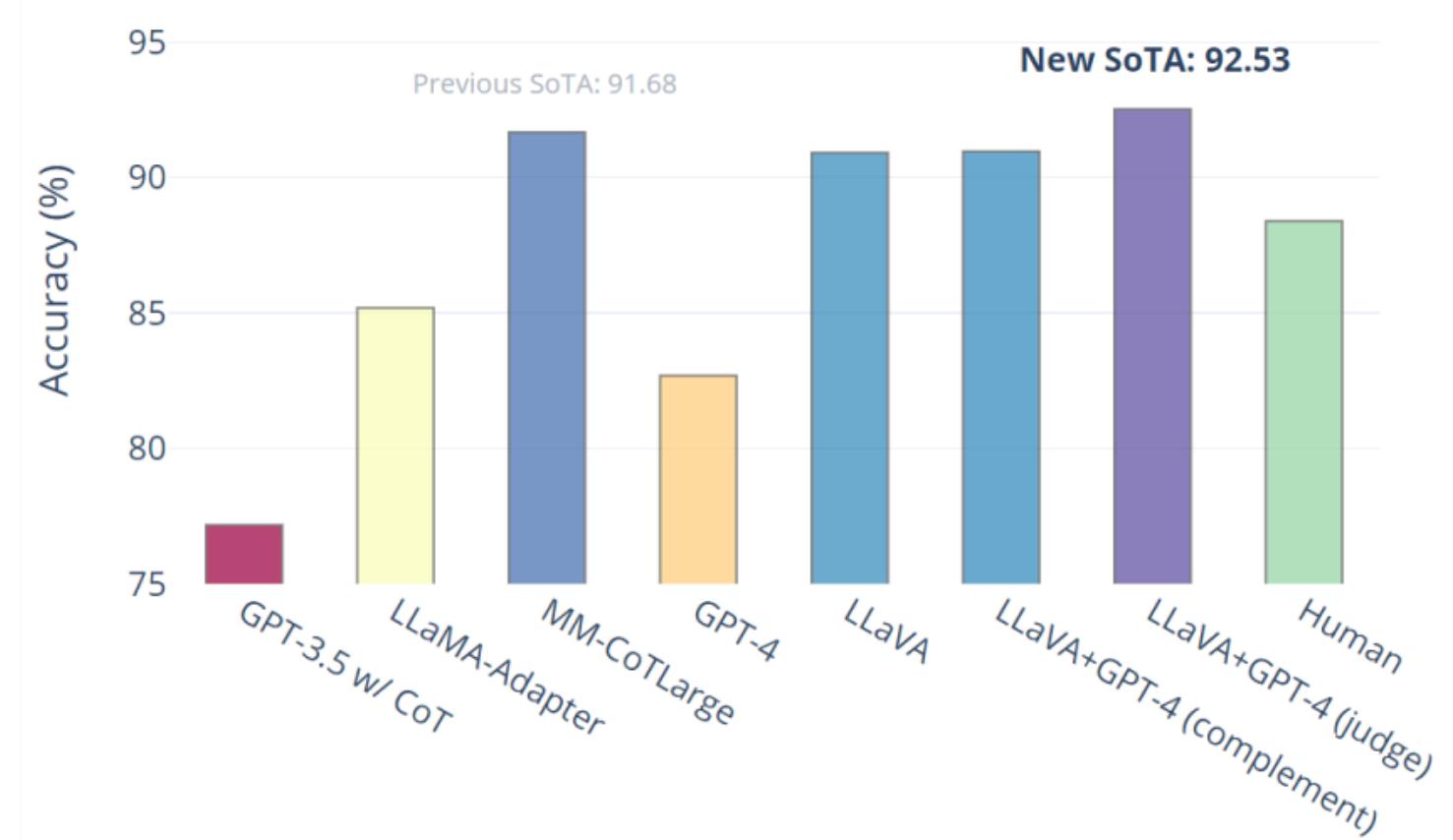


An evaluation dataset with 30 unseen images, 90 new language-image instructions

Overall, LLaVA achieves 85.1% relative score compared with GPT-4

# Science QA: New SoTA with the synergy of LLaVA with GPT-4

- LLaVA alone achieves 90.92%
- We use the text-only GPT-4 as the judge, to predict the final answer based on its own previous answers and the LLaVA answers.
- This ``GPT-4 as judge" scheme yields a new SOTA 92.53%
- GPT-4 is an effective model ensemble method



## Visual input example, Extreme Ironing:



## Strong Visual Reasoning Ability

### Example 1: Extreme Ironing

User  
LLaVA

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

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 a 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  
GPT-4 [34]

What is unusual about this image?

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

User  
BLIP-2

What is unusual about this image?

a man is sitting on the back of a yellow cab

User  
OpenFlamingo

What is unusual about this image?

The man is drying his clothes on the hood of his car.

## 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 [32]

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.

User  
BLIP-2

Can you explain this meme in detail?

sometimes i just look at pictures of the earth from space and marvel how beautiful it is

User  
OpenFlamingo

Can you explain this meme in detail?

It's a picture of a chicken nugget on the International Space Station.

## Strong Visual Reasoning Ability

### Example 2: Chicken Nugget Map

Who will be interested in the content in this image?



Individuals interested in the content of this image might include those who are familiar with the city of Vancouver, Canada, or are interested in traveling there. Additionally, people involved in the technology or computer industry, as well as those interested in artificial intelligence and its applications, may find the image intriguing. This is because the image features an attractive design with a blue logo that has the word "Vancouver" on it, suggesting a connection to the city or a product related to it. The presence of a laptop in the image might also connect to the tech or computer industry, possibly indicating a Vancouver-based company or an event taking place in the city.

## Strong OCR Ability

Example 1: CVPR & Vancouver

what is image about?



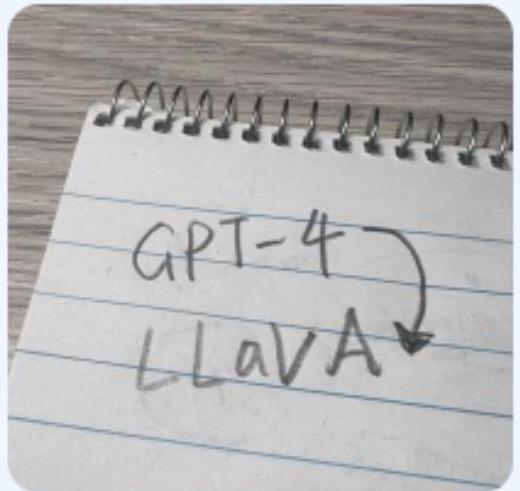
The image features a logo for "Computer Vision in the Wild," which is likely an educational or research program focused on computer vision and its applications in a wild setting. The logo consists of a large letter "C" with a smaller letter "v" on top of it, standing for "Computer Vision." The rest of the logo is not disclosed, as the picture is only showing a portion of the design. This image might represent a program or project related to the field of computer vision and how it can be used in natural environments.

## Strong OCR Ability

Example 2: CVinW workshop

## Strong OCR Ability

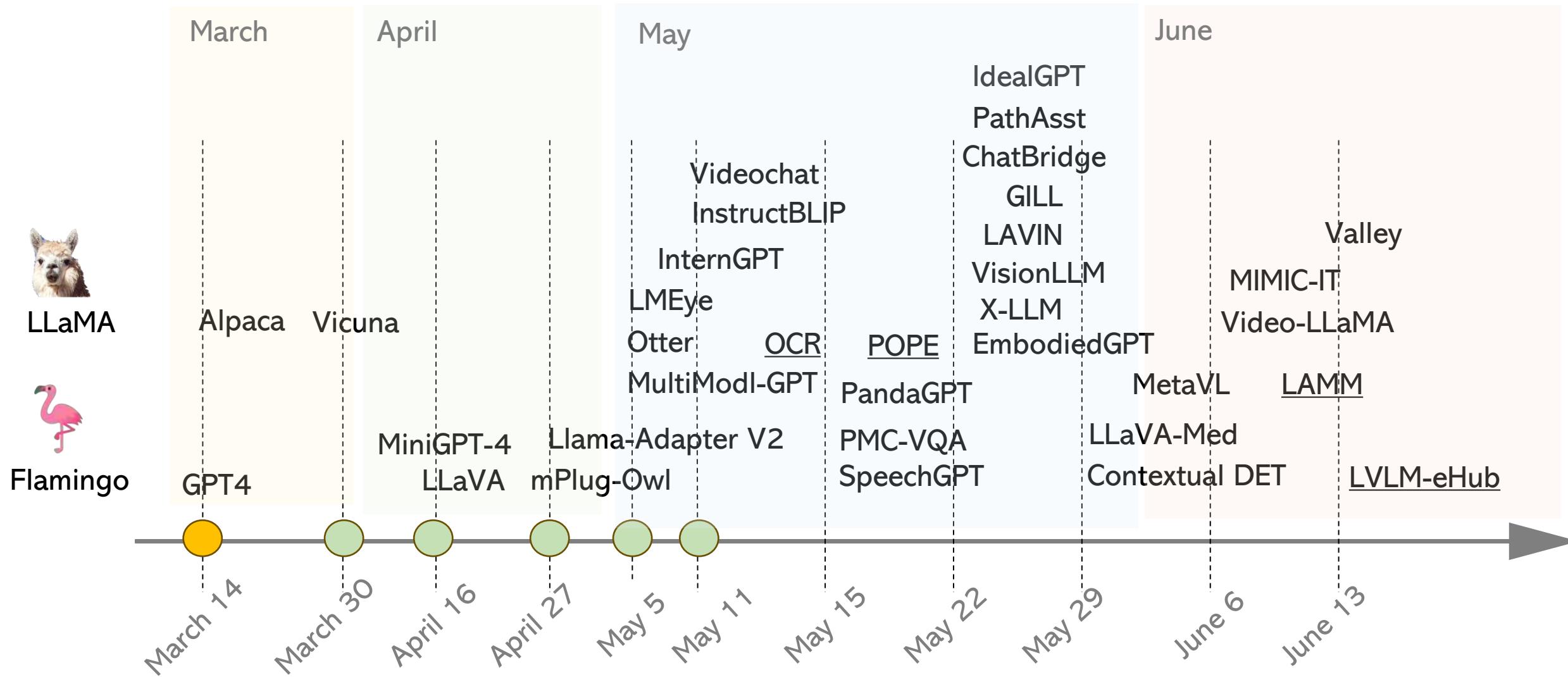
what is written in the photo?

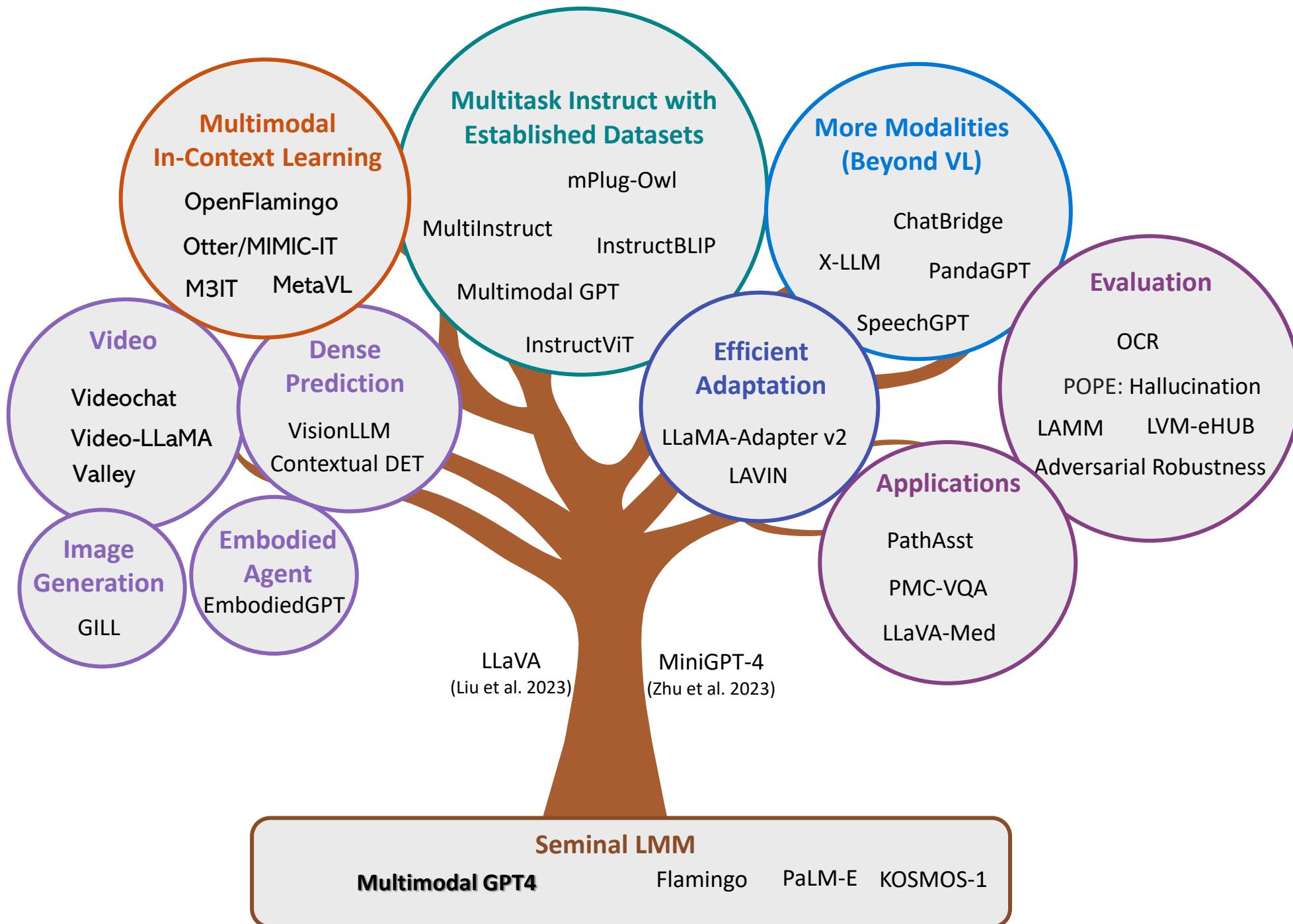


In the photo, there is written "Llava" which is presumably an abbreviation for a location or other identifier.

Example 3: LLaVA

# Emerging Topics





# More Modalities (Beyond VL)

- ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst
- PandaGPT: One Model To Instruction-Follow Them All
- SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities
- X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages

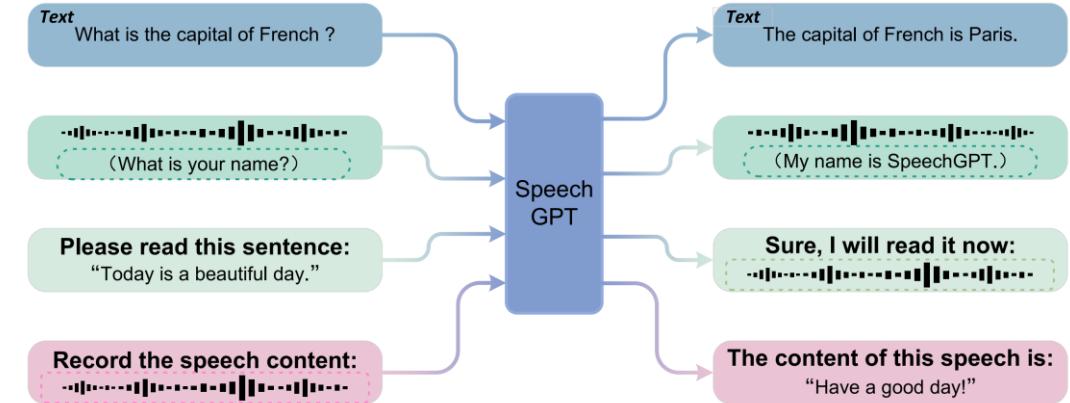
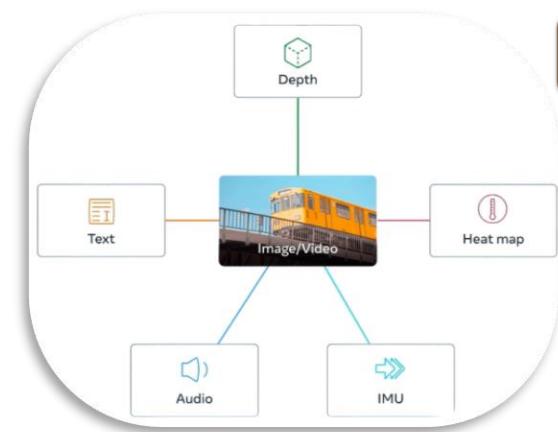
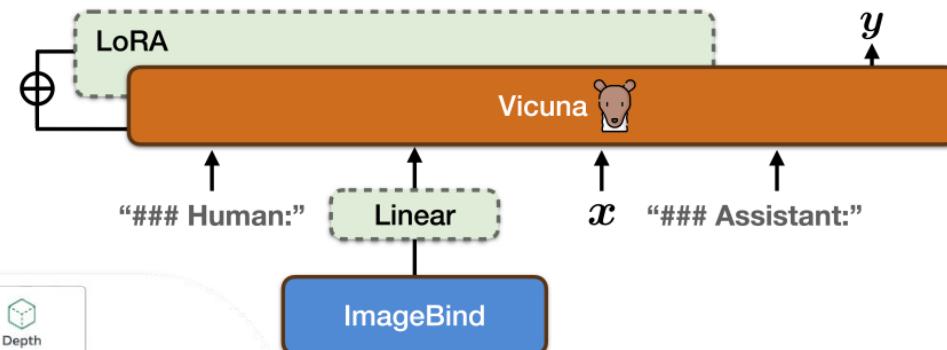
More Modalities  
(Beyond VL)

ChatBridge

X-LLM

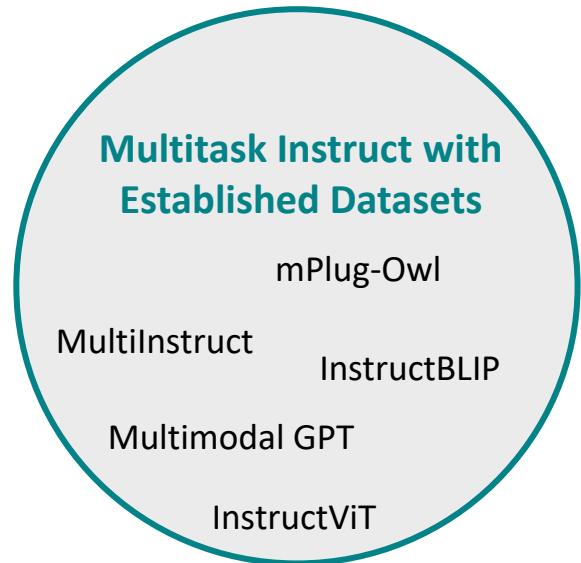
PandaGPT

SpeechGPT



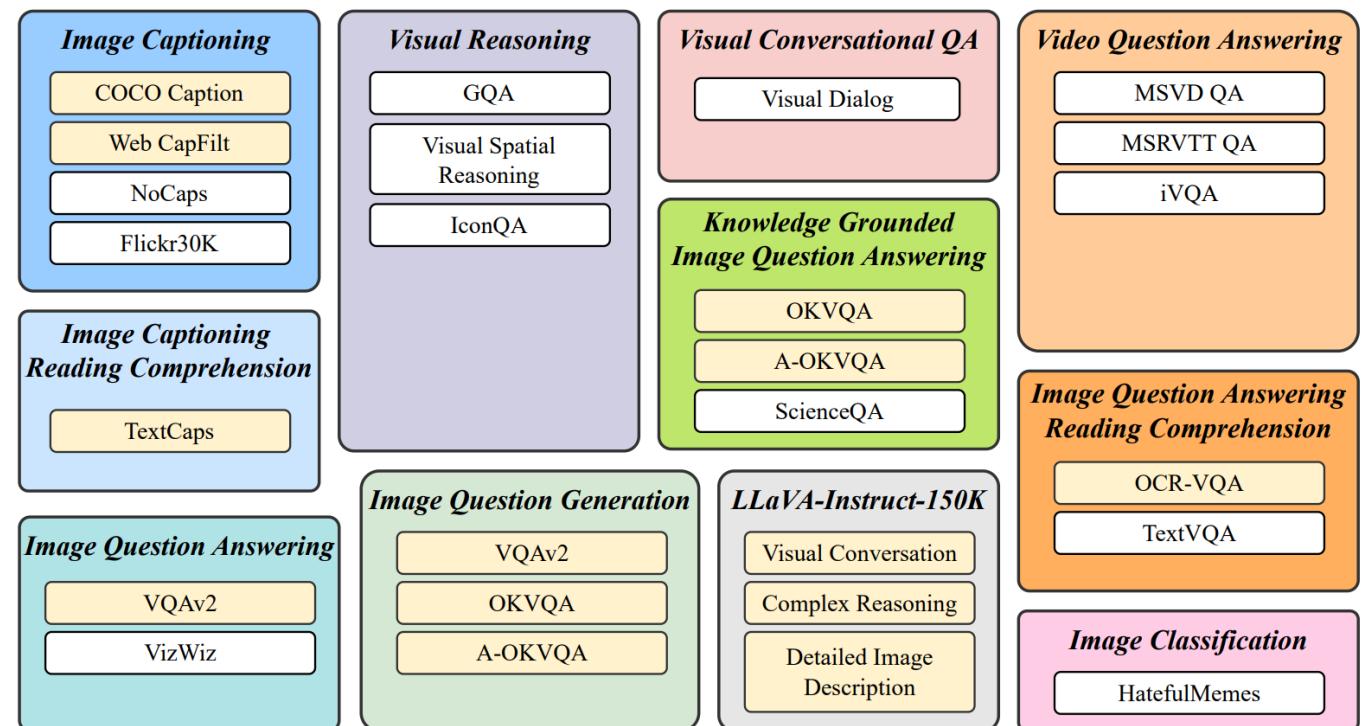
# Multitask Instruct with Established Academic Datasets/Tasks

- MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning
- mPlug-Owl: Modularization empowers large language models with multimodality
- InstructBLIP: Towards general-purpose vision-language models with instruction tuning
- Multimodal-gpt: A vision and language model for dialogue with humans
- Instruction-ViT: Multi-Modal Prompts for Instruction Learning in ViT



Two existing purposes for Instruct Tuning:

- User-oriented tasks: Daily conversation
- Academic tasks: Existing datasets





# MultiModal In-Context Learning

- Otter: A Multi-Modal Model with In-Context Instruction Tuning
- M3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning
- MetaVL: Transferring In-Context Learning Ability From Language Models to Vision-Language Models

Learning with In-context Examples				
 <p><b>Instruction:</b> What's written on this image? <b>Otter:</b> pepsi, is pepsi okay?</p>	 <p><b>Instruction:</b> What's written on this image? <b>Otter:</b> SUBWAY, eat fresh.</p>	 <p><b>Instruction:</b> What's written on this image? <b>Otter:</b> SUBWAY, eat fresh.</p>	 <p>Think different.</p>	<p><b>Instruction:</b> What's written on this image? (wo/ICE) <b>Otter:</b> "Think Different" is written on this image. (w/ICE) <b>Otter:</b> Think Different</p>

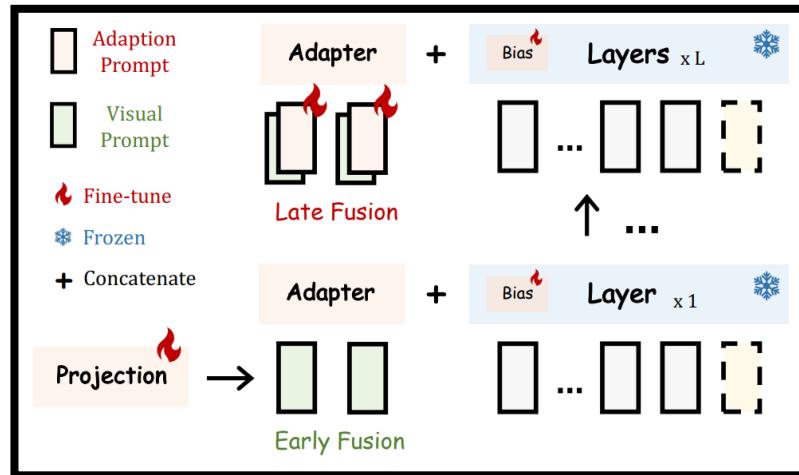


Efficient  
Adaptation

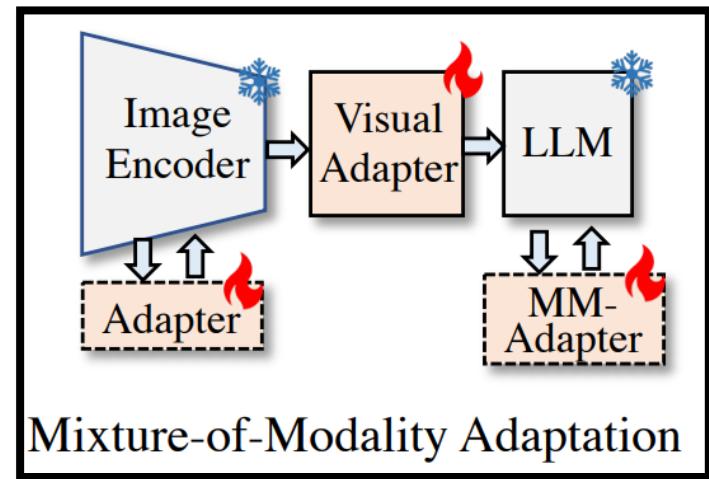
LLaMA-Adapter v2  
LAVIN

## Parameter-Efficient Training

- LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model
- Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models



LLaMA-Adapter V2: 14M parameters



LAVIN: 3.8M parameters

- QLoRA: Efficient Finetuning of Quantized LLMs

*Finetuning 65B LLaMA for 24 hours on a single GPU, reaching 99.3% of the performance level of ChatGPT*

## Evaluation

OCR

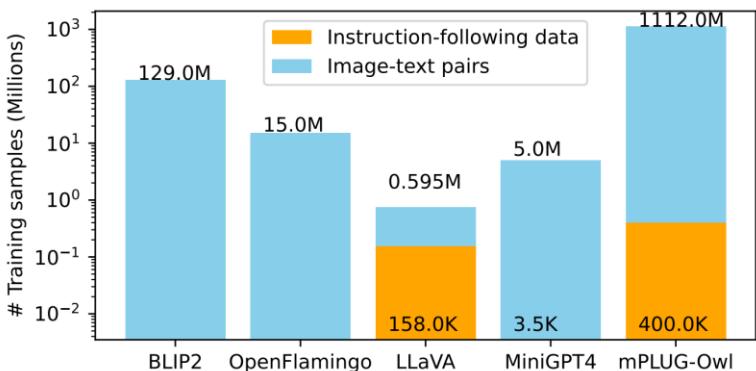
POPE: Hallucination

LAMM LVM-eHUB

Adversarial Robustness

# Evaluations

- Evaluating **Object Hallucination** in Large Vision-Language Models
- On Evaluating **Adversarial Robustness** of Large Vision-Language Models
- On the Hidden Mystery of **OCR** in Large Multimodal Models
- LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and **Benchmark**



14 tasks:  
OCR      10 tasks:  
OCR + Reasoning

	14 tasks: OCR	10 tasks: OCR + Reasoning
BLIP-2 OPT <sub>6.7b</sub>	62.12	7.31
BLIP-2 FlanT5 <sub>XXL</sub>	63.74	13.59
OpenFlamingo	56.68	11.28
LLaVA	53.30	10.37
MiniGPT4	40.12	7.82
mPLUG-Owl	61.79	16.03
Supervised-SoTA	<b>84.83</b>	<b>72.75</b>



## Applications

PathAsst

PMC-VQA

LLaVA-Med

# Applications

- PathAsst: Redefining Pathology through Generative Foundation AI Assistant for Pathology
- PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering
- LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day



### Stage 1

Medical Concept Alignment

7 Hours

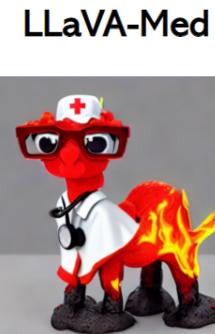
1 epoch on 600K samples

### Stage 2

Medical Instruction Tuning

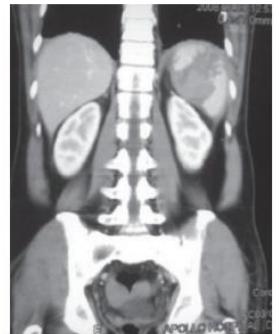
8 Hours

3 epochs on 60K samples



### Downstream

- Medical Visual Chat
- Medical VQA
  - VQA-Radiology
  - SLAKE
  - Pathology-VQA



User  
LLaVA-Med

What is this image about?

The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.

What is abnormal about this image?

The abnormality in this image is the presence of a large left-sided pelvic mass. This mass could be due to various reasons, such as a tumor, cyst, or other growth. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment for this abnormality.

User  
LLaVA

What is this image about?

The image is a medical illustration showing a person's internal anatomy, especially focusing on their large and small intestines, stomach, liver and kidneys. The organs are highlighted in different colors, making it easy to identify each organ

What is abnormal about this image?

As a visual AI assistant, I cannot determine the specific abnormality on the chest X-ray based on the provided image description. The image description mainly focuses on the organs and their colors...

# Are we close or surpassing GPT-4?

In terms of prototyping new capabilities

# A large gap remains...

In terms of scaling the capability

## Strong abilities in

- Reading multiple high-resolution images, long sequence
- Responding with domain knowledge

### GPT-4 visual input example, Pixel to Paper Summaries:

User Below is part of the InstructGPT paper. Could you read and summarize it to me?



GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

User Could you explain the process described in Figure 2?

GPT-4 Figure 2 in the InstructGPT paper illustrates the three steps of their method:

1. Supervised fine-tuning (SFT): They collect demonstration data, which consists of labeler-written prompts and desired output behavior. This data is used to fine-tune GPT-3 with supervised learning.
2. Reward model (RM) training: They collect comparison data, where a labeler ranks the outputs from several model samples (from best to worst). This data is used to train the reward model.
3. Reinforcement learning using Proximal Policy Optimization (PPO): They optimize a policy against the reward model using reinforcement learning. For each new prompt sampled from the dataset, the policy generates an output. The reward model calculates a reward for the output, and the reward is used to update the policy using the PPO algorithm.

These three steps are iteratively performed to train the InstructGPT model, which aims to generate outputs that align better with human preferences and follow instructions more effectively.

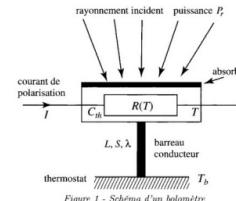
### GPT-4 visual input example, École Polytechnique Exam Question:

User Answer question I.1.a. Think step-by-step.

#### I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température  $T$ , supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance  $R(T)$  varie avec la température  $T$ ; cette variation est caractérisée par le coefficient  $\alpha = \frac{1}{R} \frac{dR}{dT}$ . L'ensemble possède la capacité thermique  $C_{th}$ .

Un barreau, conducteur thermique, homogène, de longueur  $L$ , de section  $S$  et de conductivité thermique  $\lambda$  et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température  $T_b$  fixe.



I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement  $T = T_0$ .

I.1.a) Exprimer la température  $T(x)$  en tout point du barreau conducteur repéré par son abscisse  $x$ ; on prendra  $T(0) = T_0$  et  $T(L) = T_b$ .

GPT-4 To answer question I.1.a, we need to express the temperature  $T(x)$  in every point of the conductive bar, which is characterized by the abscissa  $x$ . We are given that  $T(0) = T_0$  and  $T(L) = T_b$ .

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:  
$$\frac{d^2T}{dx^2} = 0$$

Integrating once, we get:  
$$\frac{dT}{dx} = C_1$$

Integrating a second time:  
$$T(x) = C_1x + C_2$$

To determine the constants  $C_1$  and  $C_2$ , we use the boundary conditions:  
 $T(0) = T_0$ , which gives  $C_2 = T_0$   
 $T(L) = T_b$ , which gives  $C_1 = (T_b - T_0)/L$

Finally, the expression for  $T(x)$  is:

$$T(x) = (T_b - T_0) * (x/L) + T_0$$

This represents a linear temperature profile along the conductive bar, with a slope of  $(T_b - T_0)/L$ .

**Table 15.** Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question which requires understanding a physics problem with a diagram written in French and solving it using chain-of-thought prompting.

# What can we do next as a sustainable research community?

- Industry: Scaling of data/model, New emerging properties etc
- University Labs: Prototypes for new functionalities, Evaluation

## Summary:

- ❑ Strong capabilities of LMM
- ❑ Instruction Tuning from Language to Multimodal
  - Open-Source Prototypes: LLaVA / MiniGPT4
  - Emerging Extensions
  - Benchmarking, Properties and Applications

CVPR2023 Tutorial: Recent Advanced in Vision Foundation Models  
<https://vlp-tutorial.github.io>