



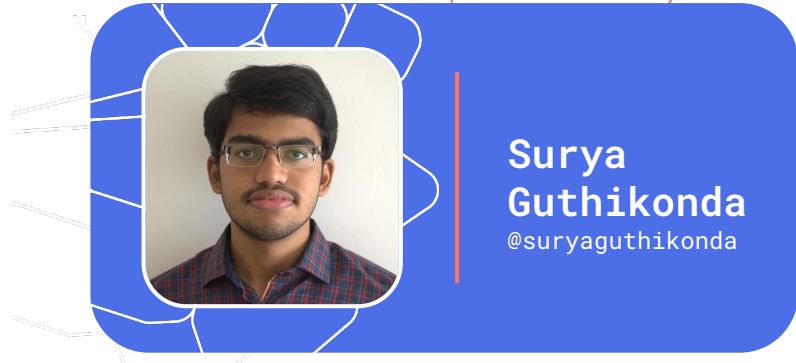
# MULTIMODAL AI SESSION #2

28th June 2024

## About Us



**Henry Vo**  
 @\_lowkeyboi



**Surya  
Guthikonda**  
 @suryaguthikonda

# What to expect?

#multimodal-ml - Channel  
@multimodal - Role



## NEWSLETTER

### Monthly Newsletter

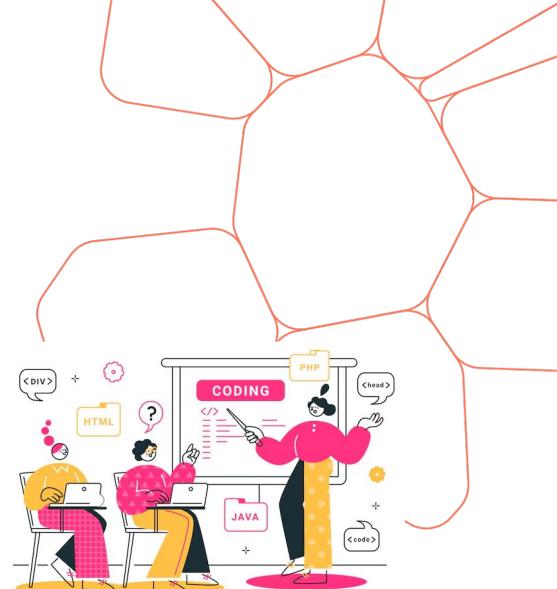
- Industry
- Research
- Learning Resources



## Paper Reading Session



## Guest Speaker Session



## Live Coding Session

First Monday

Second and Fourth Friday 10 AM ET

And More...

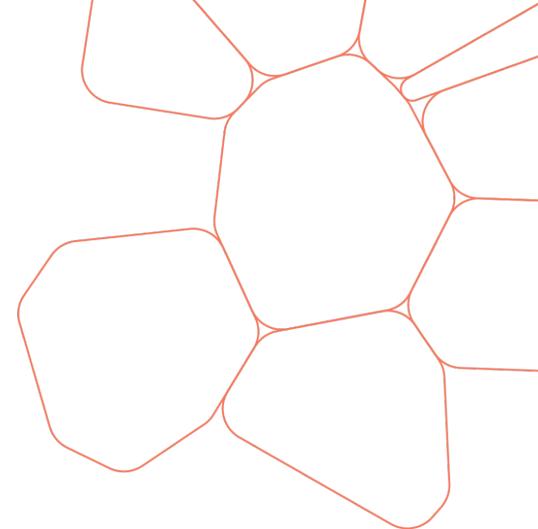


# Multimodal Foundation Models: From Specialists to General-Purpose Assistants

[Source:CVPR 2023](#)

# What will we cover today ?

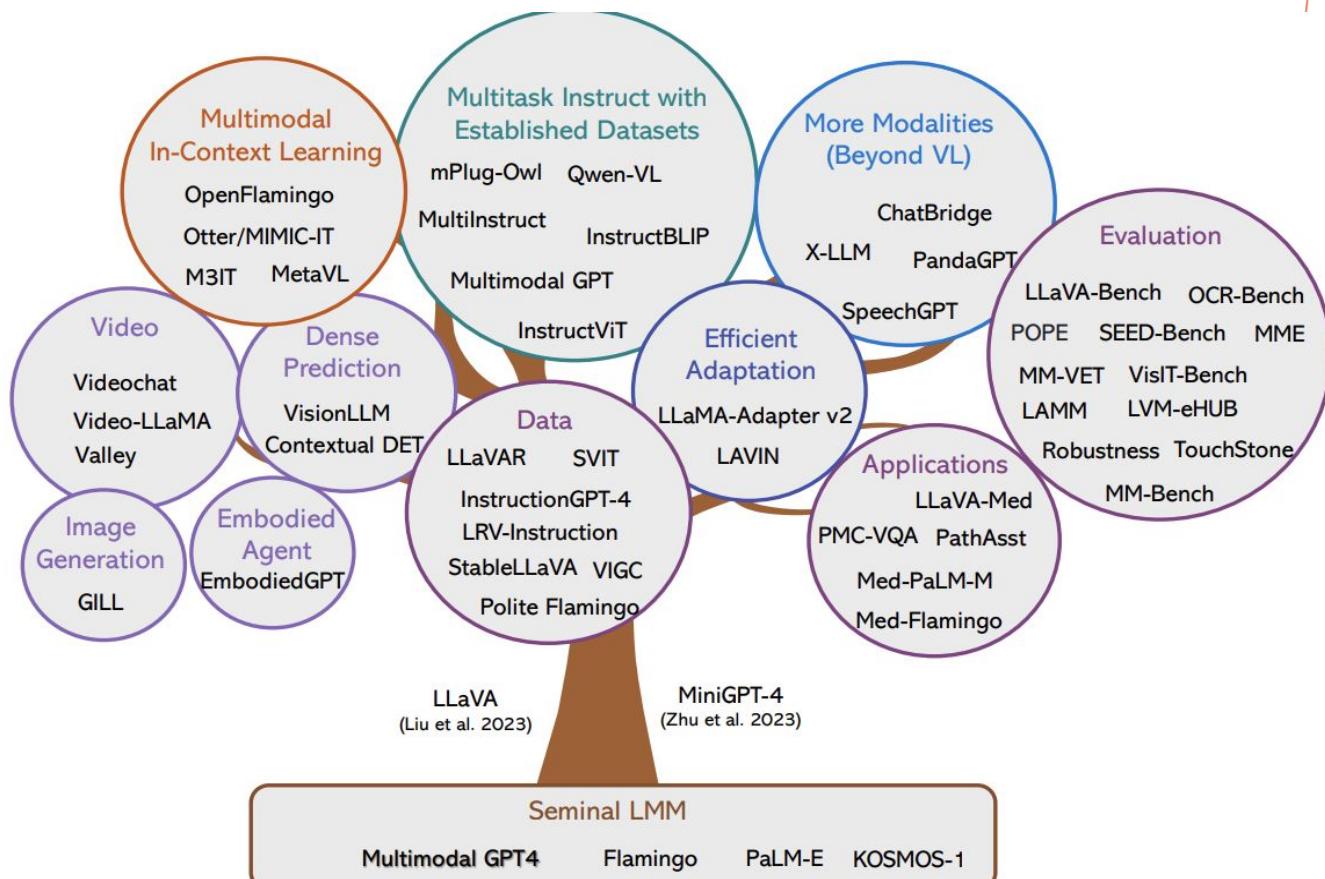
- Training with Multimodal LLM
- Multimodal Agentics
- Research Trends



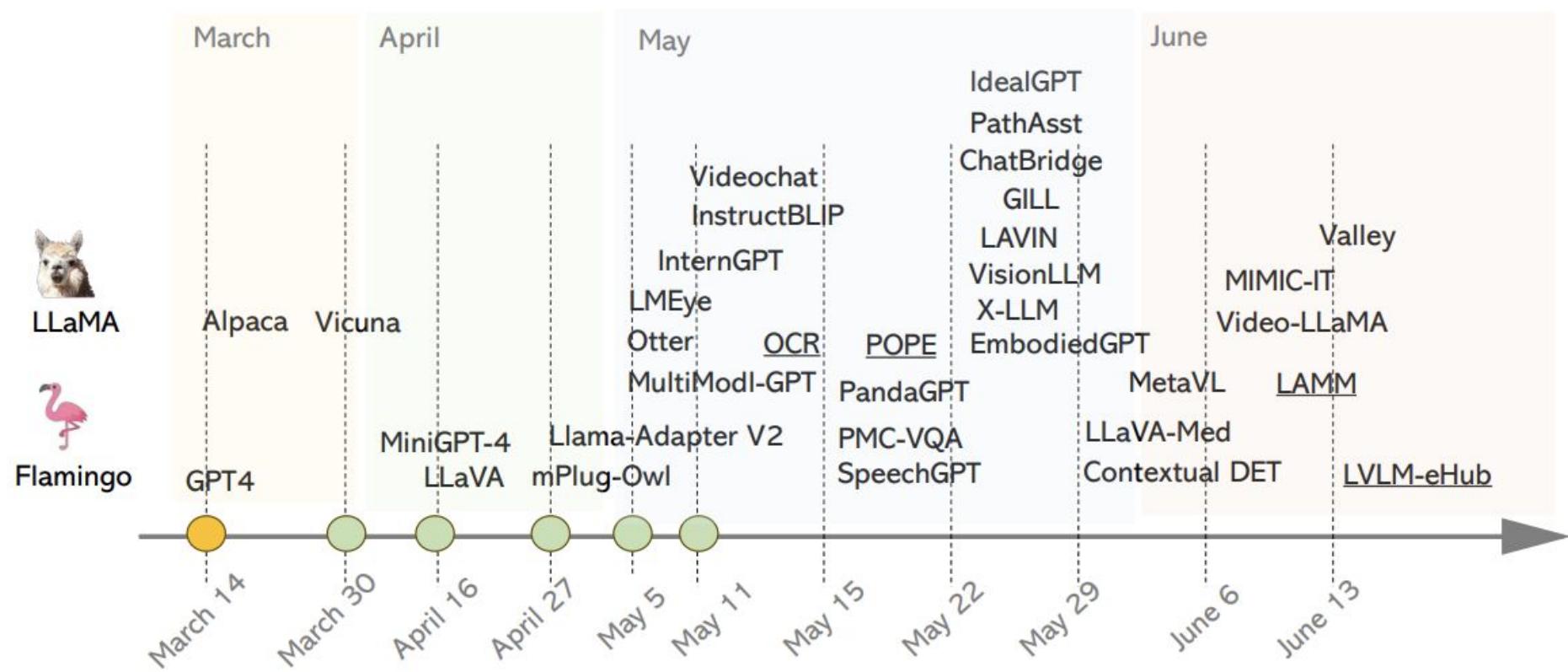
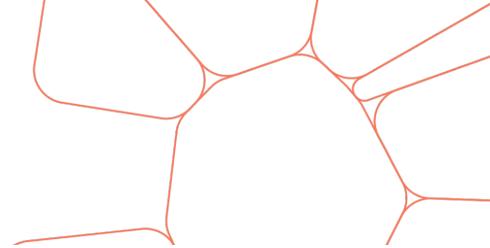
# Large Multimodal Models: Training with LLM



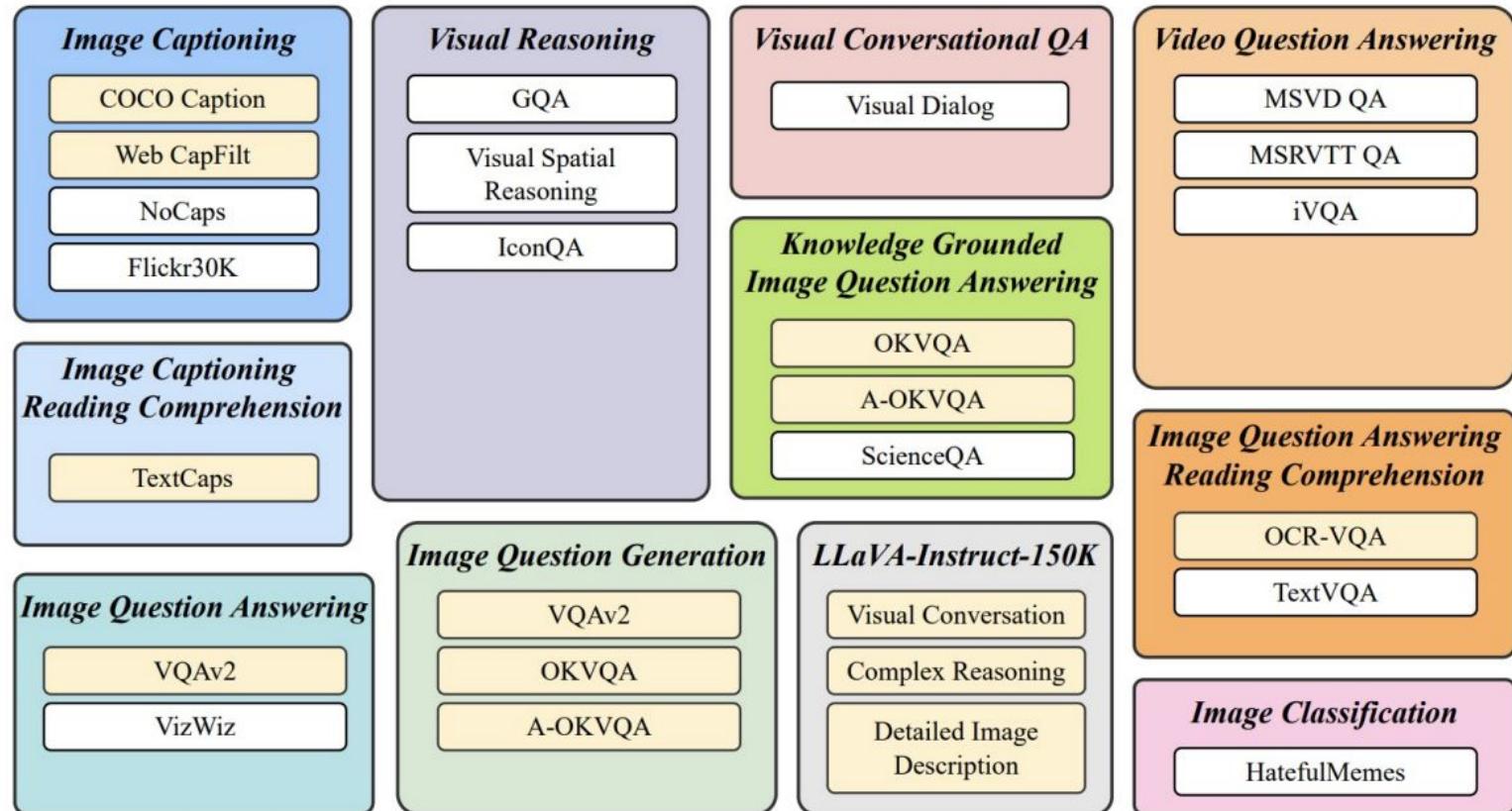
# Summarization of Multimodal LLM



# The Growth of Multimodal LLM



# Task Cover in Multimodal LLM



# Image-to-Text (Image Captioning) Deep Learning

Language

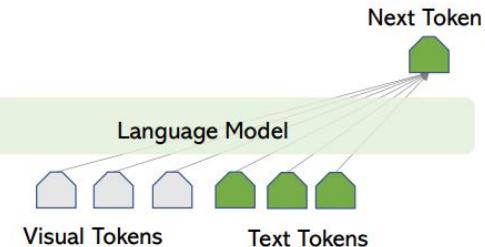
Language Model

Connection Module

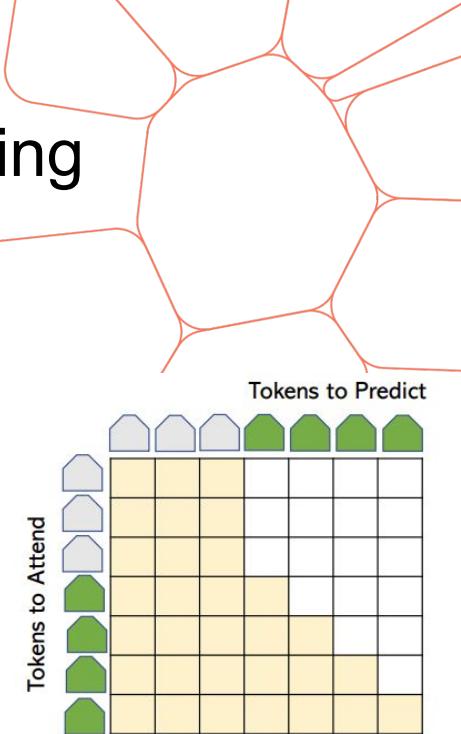
Image

Vision Encoder

Model Architecture of  
Image Captioning Task



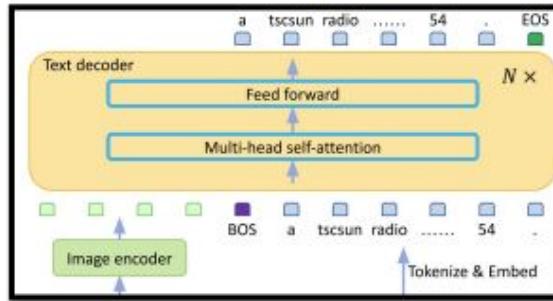
Training Objectives and  
Attention Mask



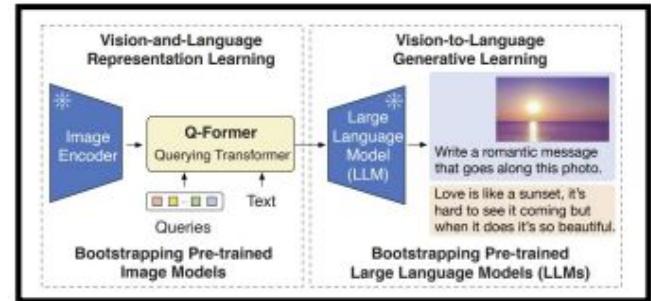
# Deep Learning with Multimodal Training



- GIT



- BLIP2



Language Model

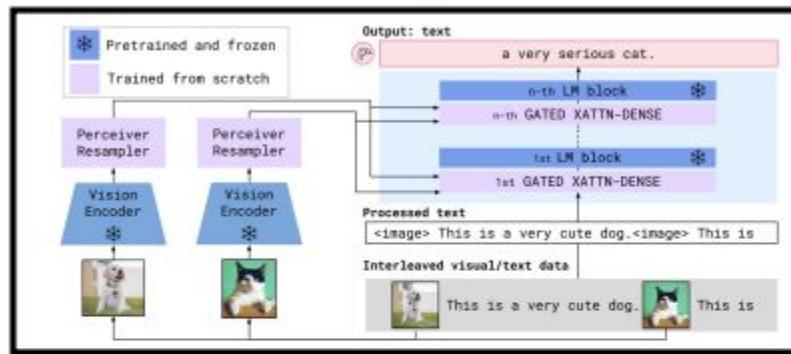
Connection Module

Vision Encoder

	From Scratch	Pre-trained: FLAN-T5/OPT
		Q-Former: Lightweight Querying Transformer
	Contrastive pre-trained: Florence/CLIP	Contrastive pre-trained: EVA/CLIP

# Large Language Model with Multimodal Training

- Flamingo



Language Model

Connection Module

Vision Encoder

Pre-trained: 70B Chinchilla

Perceiver Resampler  
Gated Cross-attention + Dense

Pre-trained: Nonrmalizer-Free ResNet (NFNet)

# Image-to-Text Generatives Models

## OpenAI MultiModal GPT-4

- *Model Details: Unknown*
- *Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild*

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

User     What is unusual about this image?



A man in a yellow shirt is ironing a blue shirt on an ironing board attached to the roof of a yellow taxi cab. The taxi is driving down a city street.

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.

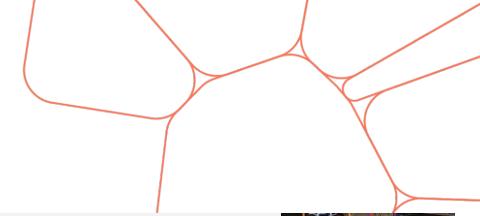


A photograph of several pieces of fried chicken nuggets arranged on a dark surface to resemble the shape of the world map.

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.

# Data Curation

- Multi-turn conversations so that users can chat with the model
- Detailed description so that long-form responses can be generated from the model
- Complex reasoning



## 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]



(a) The sequence representation of the image data.

## 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.

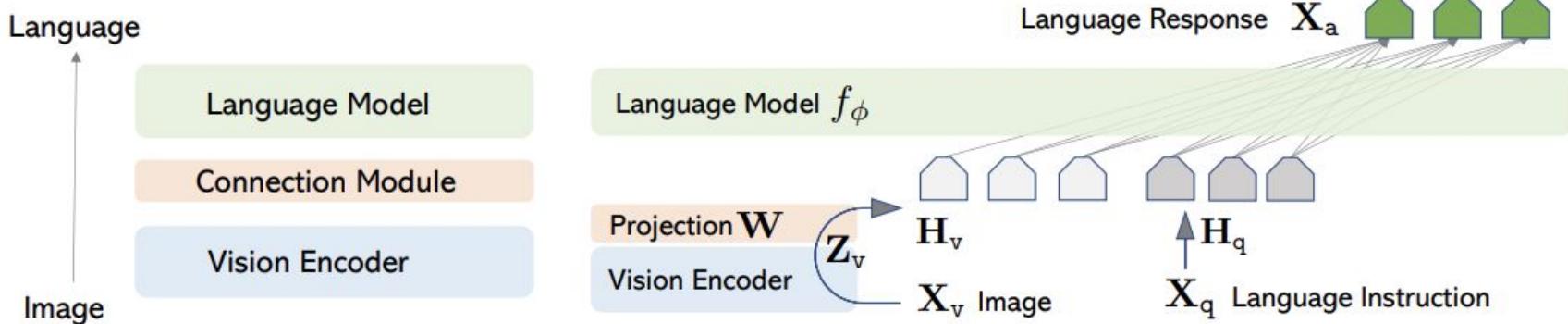
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

## Response type 3: complex reasoning

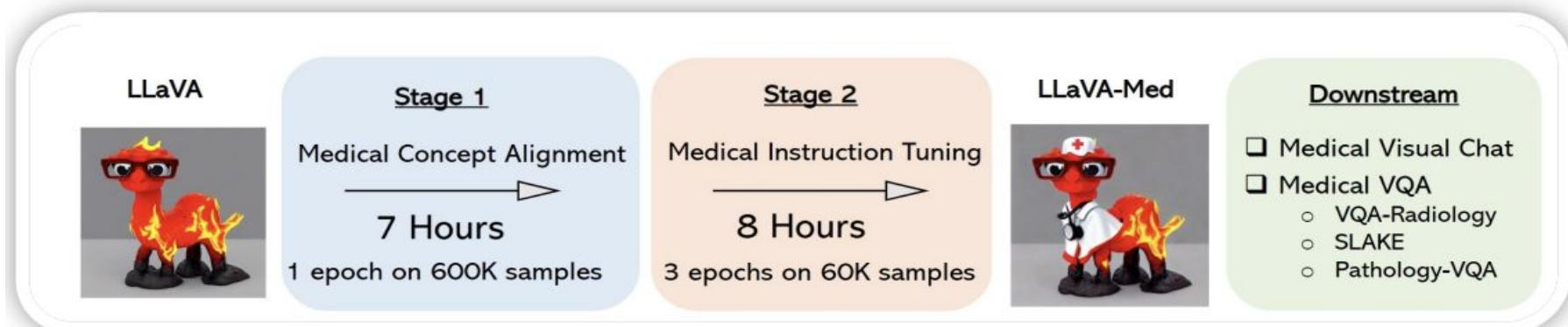
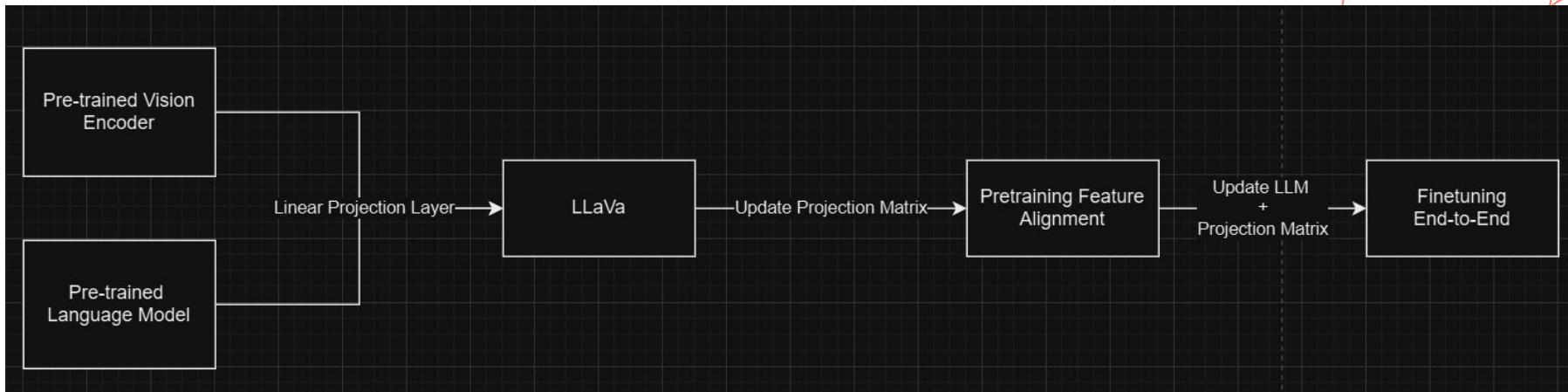
Question: What challenges do these people face?

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.

# Network Architecture

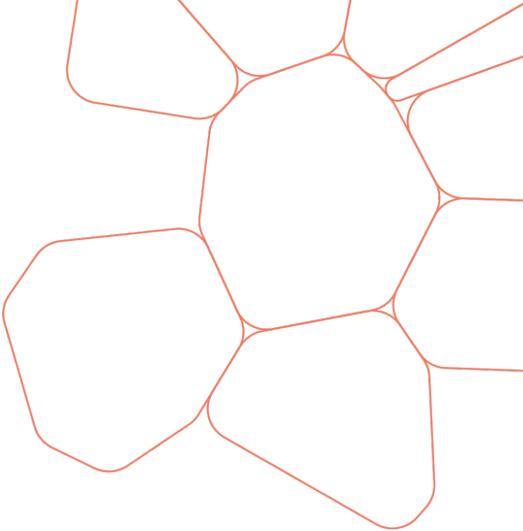


# Training Process



# How to evaluate ?

Benchmark	Capability to Evaluate	Statistics	Metric
LLaVA-Bench (Liu et al., 2023c)	Multi-turn QA, detailed description, reasoning	Two subsets: 90 samples on COCO and 60 samples on In-the-Wild	Relative score via GPT-4 evaluation
OCR-Bench (Liu et al., 2023k)	Zero-shot OCR	A suite of 23 OCR-related academic tasks	Accuracy
MMBench (Liu et al., 2023j)	Perception (coarse, fine-grained single-instance and cross-instance) and Reasoning (attribute, relation, logic)	2974 multiple-choice samples in 20 ability dimensions	Circular evaluation via ChatGPT answer extraction
M3Exam (Zhang et al., 2023i)	Multilingual, multimodal, and multi-level assessment	12,317 questions in 9 languages, with 2,816 questions involving images	Accuracy on multiple-choice questions
MME (Fu et al., 2023)	Perception and Cognition	14 tasks	Accuracy on "yes" or "no"
LAMM (Yin et al., 2023)	Various 2D/3D vision tasks	9 image task with 62K samples, and 3 point cloud tasks with 12K samples	Traditional CV task metrics
LVLM-eHub (Xu et al., 2023b)	six multimodal capabilities such as VQA and embodied AI	47 standard text-related visual benchmarks	CIDEr and accuracy; Arena with human judgment
SEED-Bench (Li et al., 2023b)	Comprehension of both the image and video modality	19K multiple choice questions in 12 dimensions	Accuracy on multiple-choice questions
VisIT-Bench (Bitton et al., 2023)	Real-life vision-language instructions	592 samples in 70 tasks	Elo, matches
MM-VET (Yu et al., 2023d)	Integrated capabilities in recognition, OCR, spatial, knowledge, math, language	200 samples	GPT-4 evaluation
TouchStone (Bai et al., 2023b)	Five abilities: basic description, visual recognition, visual comprehension, visual storytelling, and multi-image analysis	908 dialogues in 27 tasks	GPT-4 evaluation
SciGraphQA (Li and Tajbakhsh, 2023)	Scientific graph question-answering	3K test samples	CIDEr, BLEU-4, and ROUGE



## OpenVLM Leaderboard

Welcome to the OpenVLM Leaderboard! On this leaderboard we share the evaluation results of VLMs obtained by the OpenSource Framework:

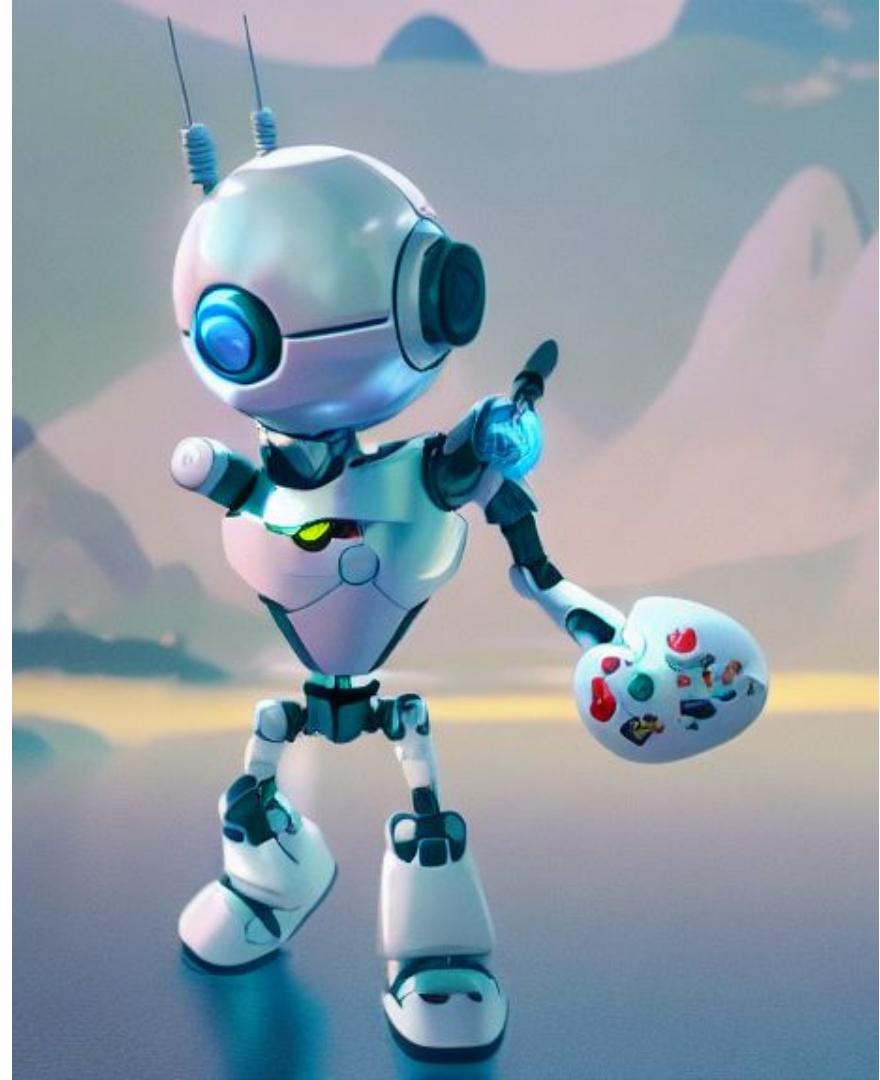
[VLMEvalKit](#): A Toolkit for Evaluating Large Vision-Language Models 🎉

Currently, OpenVLM Leaderboard covers 76 different VLMs (including GPT-4v, Gemini, QwenVLP+, LLaVA, etc.) and 23 different multi-modal benchmarks.

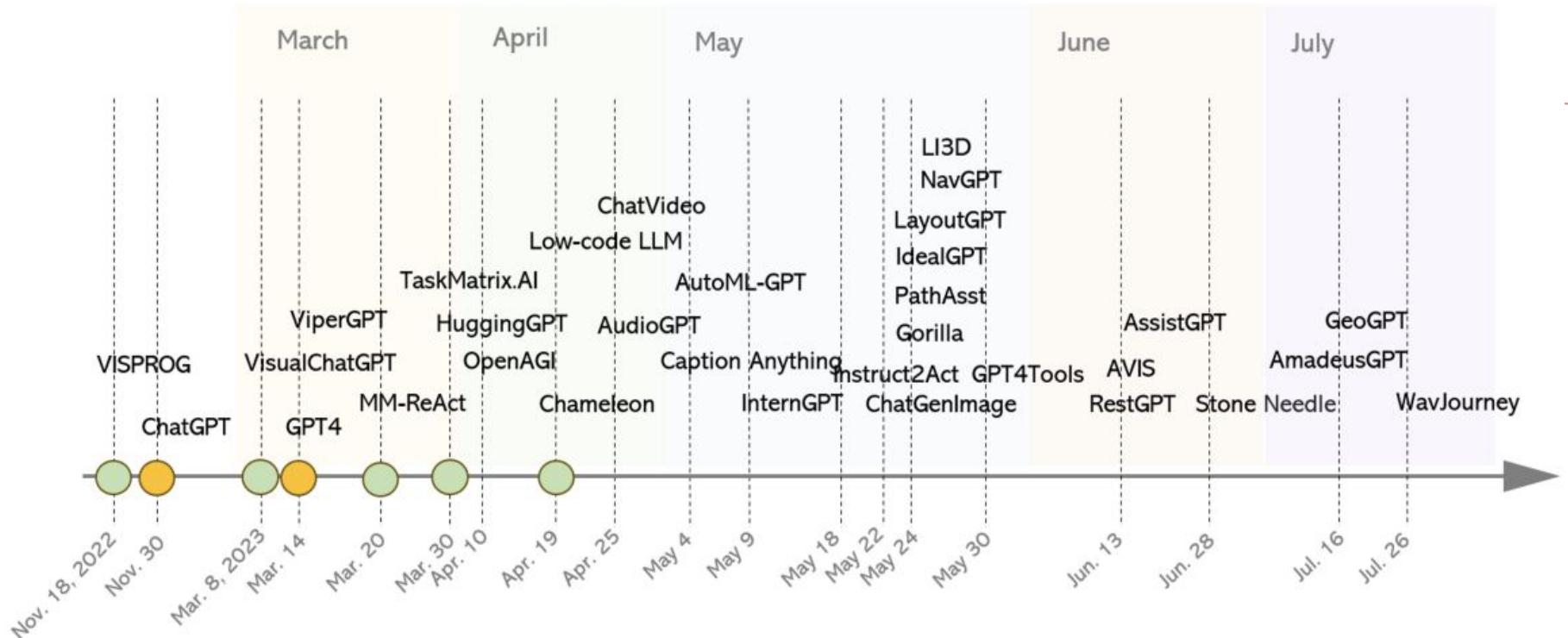
This leaderboard was last updated: 24.06.26 11:23:44.

OpenVLM Leaderboard only includes open-source VLMs or API models that are publicly available. To add your own model to the leaderboard, please create a PR in [VLMEvalKit](#) to support your VLM and then we will help with the evaluation and updating the leaderboard. For any questions or concerns, please feel free to contact us at [opencompass.duanhaodong@pjlab.org.cn].

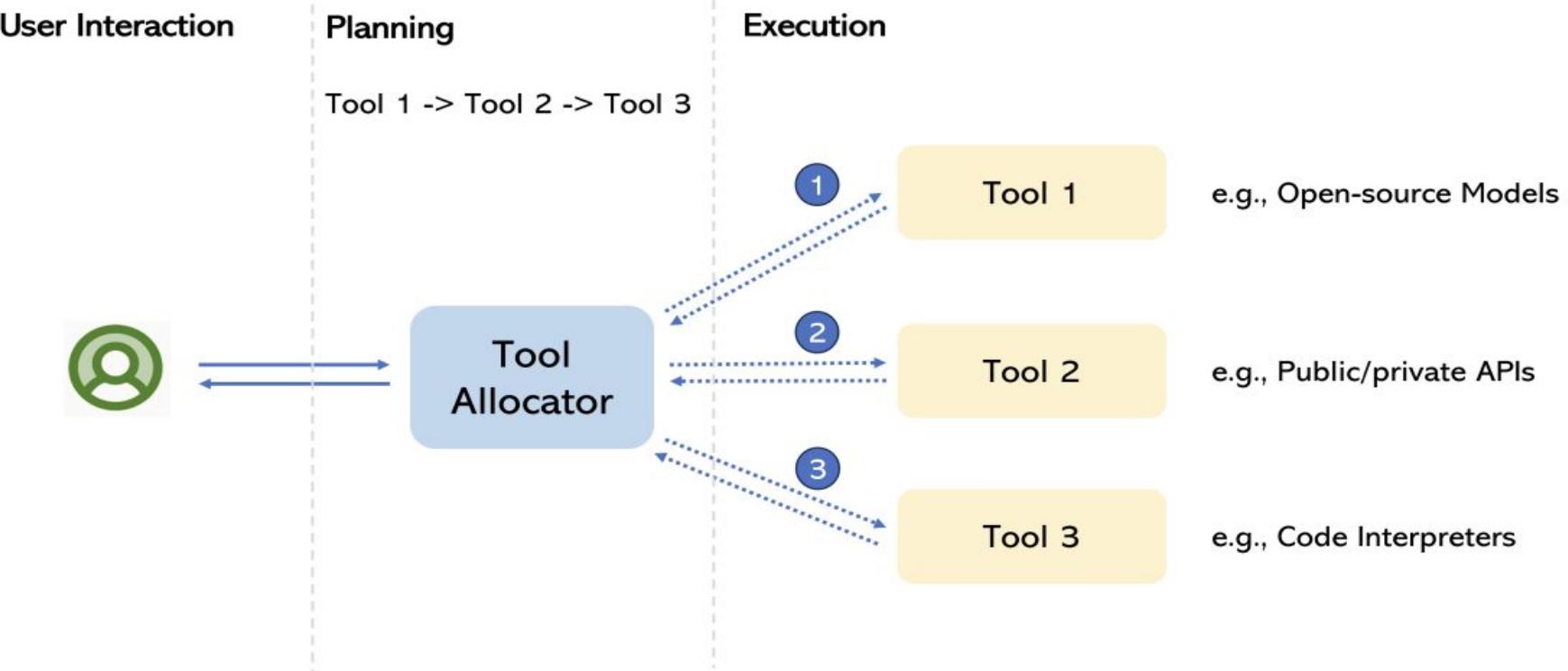
# Multimodal Agents: Chaining Tools with LLM



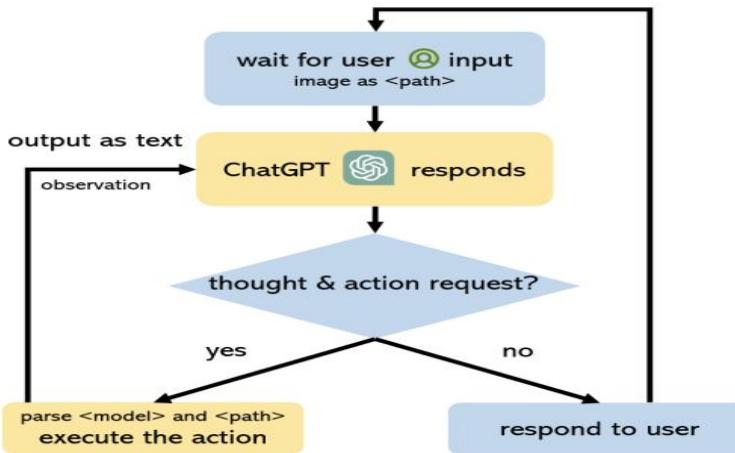
# The Growth of Multimodal Agent



# Generic Framework



# Case Study: MM-ReAct



This image contains two celebrities, Kobe Bryant and Paul Pierce. They are both basketball players.



Thought 1 This is an image.

Action request 1 Assistant, what objects do you see in this image?  
<ImagePath>

Action execution 1 Invoke vision experts over <ImagePath>

Observation 1  
Image Captioning:  
This is an image (694 x 720) with description a basketball player in a yellow jersey with a ball in his hand.

Dense Captioning:  
List of object descriptions, and their locations in this image:  
a close up of a basketball player (x1, y1, x2, y2) = (63, 535, 190, 663)  
...

Object Tagger:  
List of object tags seen in this image:  
sport, athletic game, basketball, person, sports uniform, ball, ...

Face Detection:  
Detected faces, and their location in this image:  
man (x1, y1, x2, y2) = (365, 278, 439, 352)  
man (x1, y1, x2, y2) = (548, 308, 620, 380)

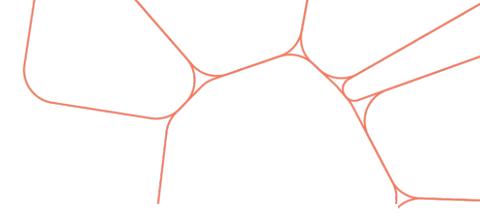
Thought 2 There are two faces detected in this image.

Action request 2 Assistant, are there any celebrities in this image?  
<ImagePath>

Action execution 2 Invoke vision experts over <ImagePath>

Observation 2  
Celebrity Recognition:  
This image contains celebrities  
List of celebrities, and their location in this image:  
Kobe Bryant (x1,y1,x2,y2) = (365, 278, 439, 352)  
Paul Pierce (x1,y1,x2,y2) = (548, 308, 620, 380)

# Case Study: MM-ReAct



## Visual Math and Text Reasoning

What is "a" in the first equation?  
What is y in the last equation?

$5 + a = 7$	$40 + y = 45$
$10 + y = 15$	$39 + a = 41$
$20 + a = 22$	$24 + y = 29$
$48 + a = 50$	$15 + a = 17$
$10 + a = 12$	$34 + y = 39$

The value of "a" in the first equation " $5 + a = 7$ " is 2.

The value of "y" in the last equation " $34 + y = 39$ " is 5.

## Visual-Conditioned Joke/Meme

What is funny about it?



The humor in the comic may come from the situation, the characters, or the dialogue. The person is feeling unwell and is asked if they are pregnant or have the coronavirus, which is a current and relevant topic. The humor may come from the absurdity of the situation or the unexpectedness of the question.

## Visual Planning and Prediction

What should I do after kneading the dough, before dividing it into pieces? For how long?



After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image.

You should put the dough in a warm place for 1 hour.

## Multi-Image Reasoning

How much in total did I spend on travel?  
How much in total did I pay for taxes?



You spent \$1299.90 on airfare and \$43.83 on an Uber ride for a total of \$1343.73 on travel.

You paid \$323.23 in total for taxes. This includes \$321.90 for taxes and fees for the flight purchase and \$1.33 for sales tax for the Windmill Family Restaurant.

## Open-World Concept Understanding

What is the brand of these drink?  
Which one is unsweetened?



The brand of these drinks is "BETTER THAN MILK". The unsweetened one is "BETTER THAN MILK organic OAT DRINK".

Who are the celebrities in this image?



Keira Knightley and Natalie Portman are the celebrities in this image.

Is it safe to eat this mushroom?



Morel mushrooms are generally considered safe to eat when properly cooked.

## Spatial/Coordinate Understanding

Where is the frisbee?  
What is the relative position of the frisbee and the man?

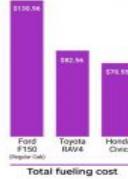


The frisbee is located at coordinates (63,119,133,177).

The frisbee is located above the man in the image. The man is falling on the ground while the frisbee is in the air. The frisbee is also closer to the top of the image than the man.

## Multi-Hop Document Understanding

Excluding the regular cab, what is the average total fueling cost?



The average total fueling cost excluding the Ford F150 (Regular Cab) is \$76.55. This is calculated by taking the sum of the total fueling cost of Toyota RAV4 and Honda Civic and dividing it by 2. ( $\$82.56 + \$70.55$ ) / 2 = \$76.55.

## Video Summarization/Event Localization

Can you breakdown this video tutorial to step-by-step instructions, with start and end time?

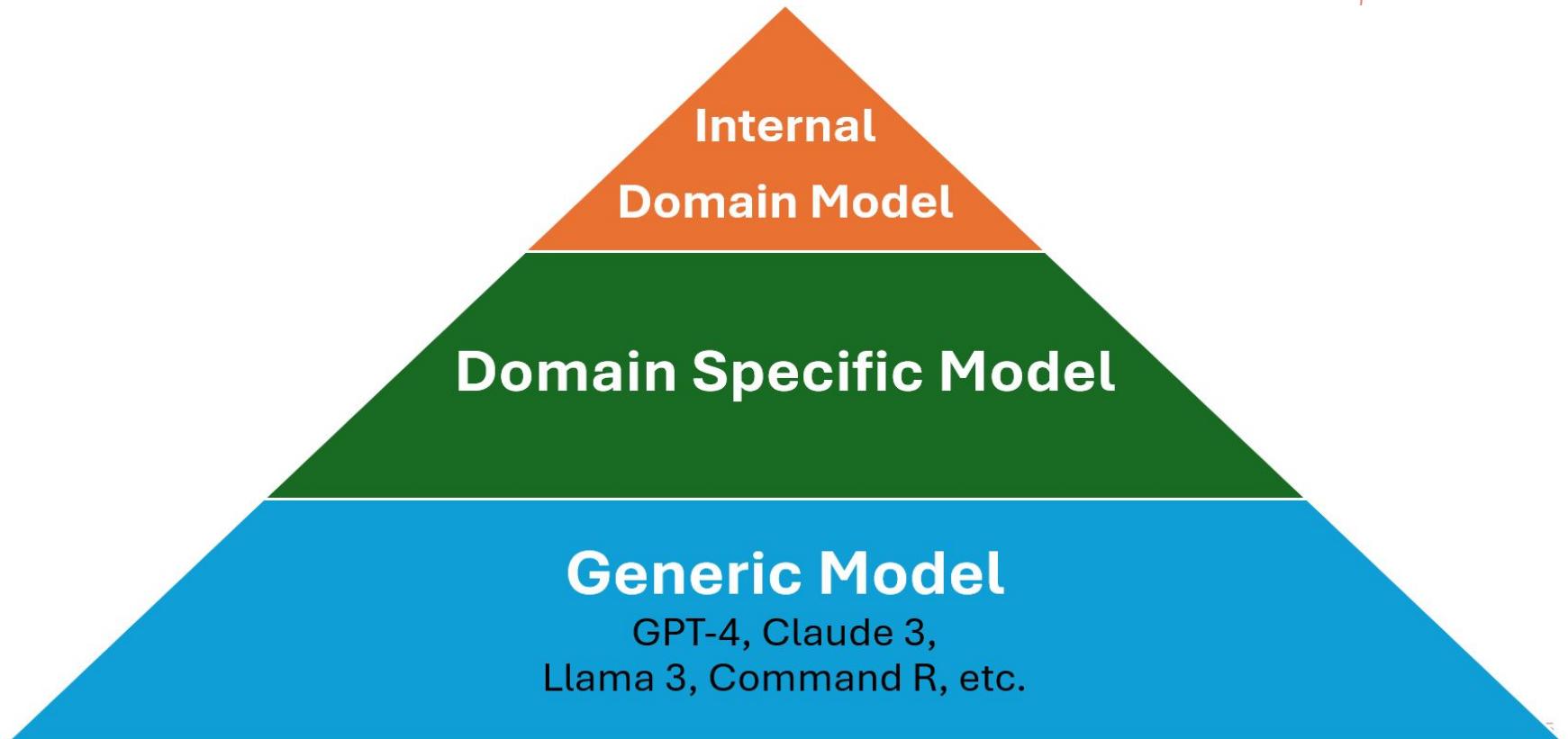


Here's a step-by-step breakdown of the tutorial with start and end time.



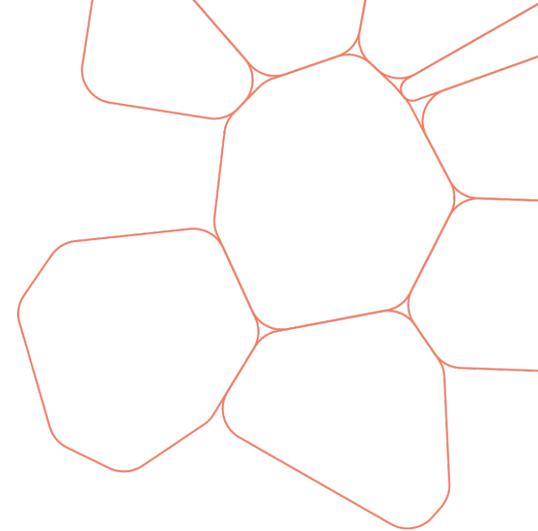
\*\*\*Add Lettuce Add Tomato Add Bacon \*\*\*

# How to improve ?



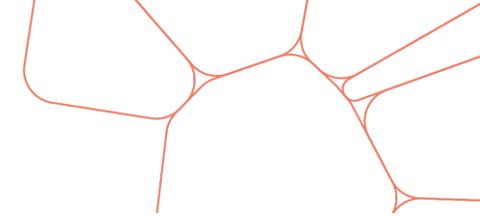
# How to evaluate ?

- Tool Choosing / Using / Creating
- Emergent capabilities benchmarking dataset
- Retrieval Augmented Generation Multimodal Agents



# Research Trends





# Specialise vs Generalise Multimodal

Models	Advantages	Disadvantages
Specific-Purpose Models	Visual Understanding	Well studied and scalable solutions on image-level understanding; Emerging interests and success on region-level and pixel-level visual understanding
	Visual Generation	Well studied and scalable solutions on image-level generation; Emerging interests and success in controllable/customized image generation
	Unified Vision Models	Promises to unlock new emerging capabilities and scenarios
	Training with LLM	Some new emerging capabilities and scenarios are enabled with light model training
	Chaining with LLM	Fast system development cycles with low cost as no training is involved
General-Purpose Assistants		High training cost; No successful scalable recipe beyond the billion-image level
		High training and inference cost; Debate between diffusion and auto-regressive solutions for the best recipe; More studies are needed for video generation
		High risks in modeling and high training cost
		The performance is bounded by LLM
		Low flexibility in improving system performance; No new emerging capabilities

# General-Purpose AI Agents

- Agents with Multi-modality
- Align human's intention
- Planning, memory and tool usage

