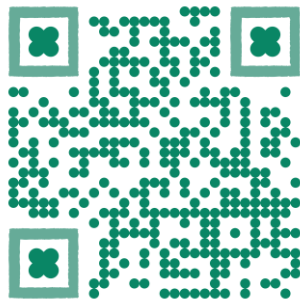


# From Multimodal LLM to Human-level AI

*Architecture, Modality, Function, Instruction, Hallucination, Evaluation, Reasoning and Beyond*

<https://mllm2024.github.io/ACM-MM2024/>



ACM Multimedia 2024

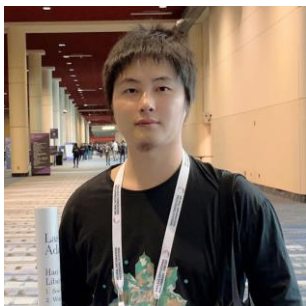


Melbourne, Australia



CC BY 4.0 DEED

Attribution 4.0 International



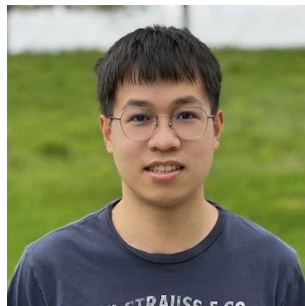
**Hao Fei**

*National University of Singapore*



**Xiangtai Li**

*ByteDance/Tiktok*



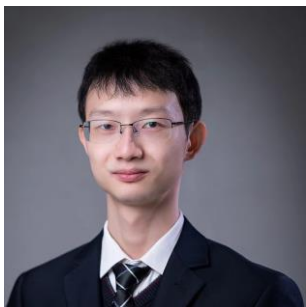
**Haotian Liu**

*xAI*



**Fuxiao Liu**

*University of Maryland, College Park*



**Zhuosheng Zhang**

*Shanghai Jiao Tong University*



**Hanwang Zhang**

*Nanyang Technological University*



**Kaipeng Zhang**

*Shanghai AI Lab*



**Shuicheng Yan**

*Kunlun 2050 Research, Skywork AI*

## \* Part-III

# MLLM Functionality & Advance

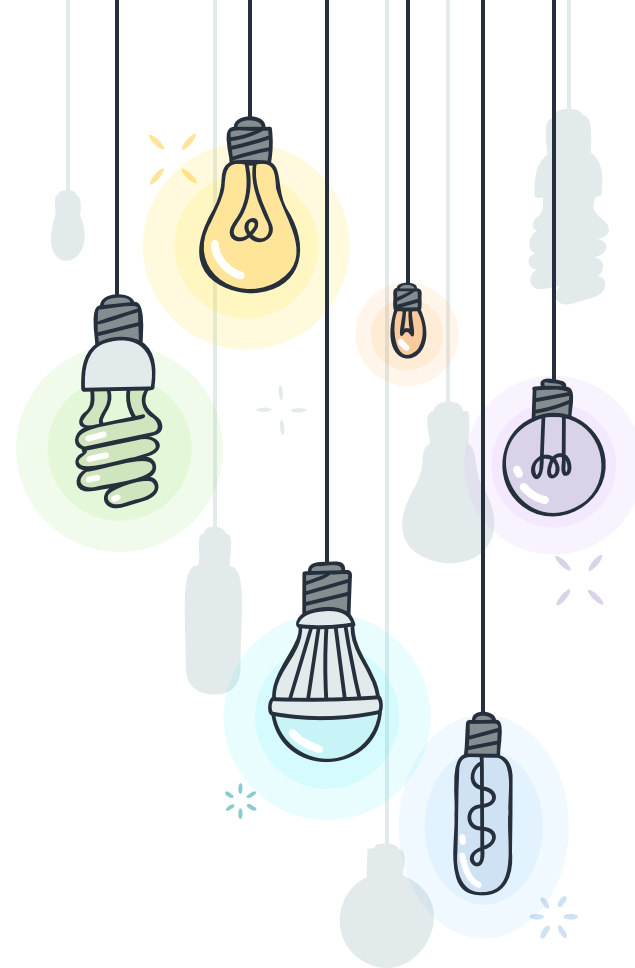


**Xiangtai Li**

**Research Scientist**

*ByteDance/Tiktok, Singapore*

<https://lxtgh.github.io/>



# \* MLLM Functionality & Advance

---

## + Fine-Grained MLLM Design

- × Overview
- × With Visual Grounding.
- × With Visual Segmentation.
- × Video and 3D Fine-Grained MLLM.

## + Advanced MLLM Design

- × Overview
- × Unified Architecture Designs.
- × MLLM For Long Video Analysis.
- × MLLM With MOE Design.

# \* MLLM Functionality & Advance

---

## + Fine-Grained MLLM Design

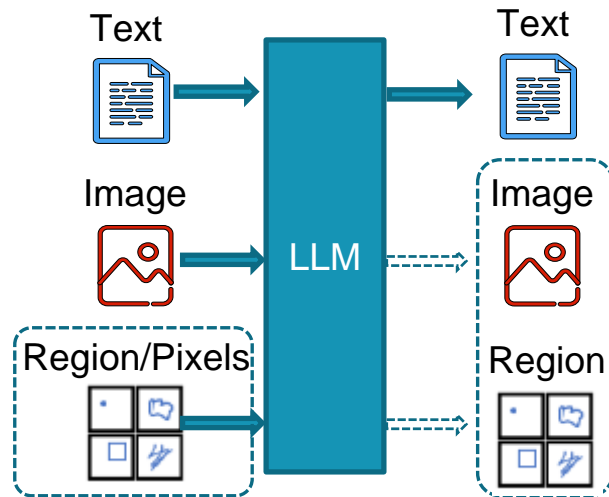
- × **Overview**
- × With Visual Grounding.
- × With Visual Segmentation.
- × Video and 3D Fine-Grained MLLM.

# \* Fine-grained Capability of MLLM

## Overview and Concepts

### Fine-Grained MLLM:

- 1, Region-level or Pixel-level visual prompts as inputs and outputs.
- 2, Aims at understanding multi-granularity concepts in image/video/3D.
- 3, Enhance the interactive features in MLLM. This is important in the real product.



# \* Fine-grained Capability of MLLM

---

## Motivation

### Why We need Fine-grained MLLM?

#### New Features:

- refer to specific regions/objects/masks and perform chat.
- understanding and reasoning region and pixel.

#### New Applications:

- VR / AR application.
- Medical image analysis.

#### New Model Designs:

- How to avoid hallucination.
- How to balance chat and localization ability.

# \* Fine-grained Capability of MLLM

---

## Overview of Fine-grained Models Before LLM

Visual Grounding

- Various tasks that driven by **language**.

Referring Segmentation

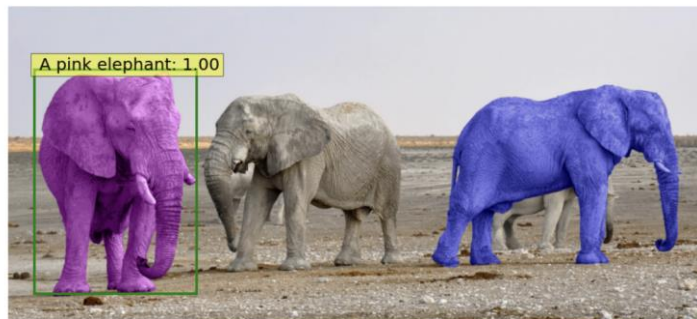
- Most works come from **vision** community.

Visual Segmentation

- **Dual** branches designs by connecting language model and vision model (detector or segmenter)

Video/3D Referring Segmentation

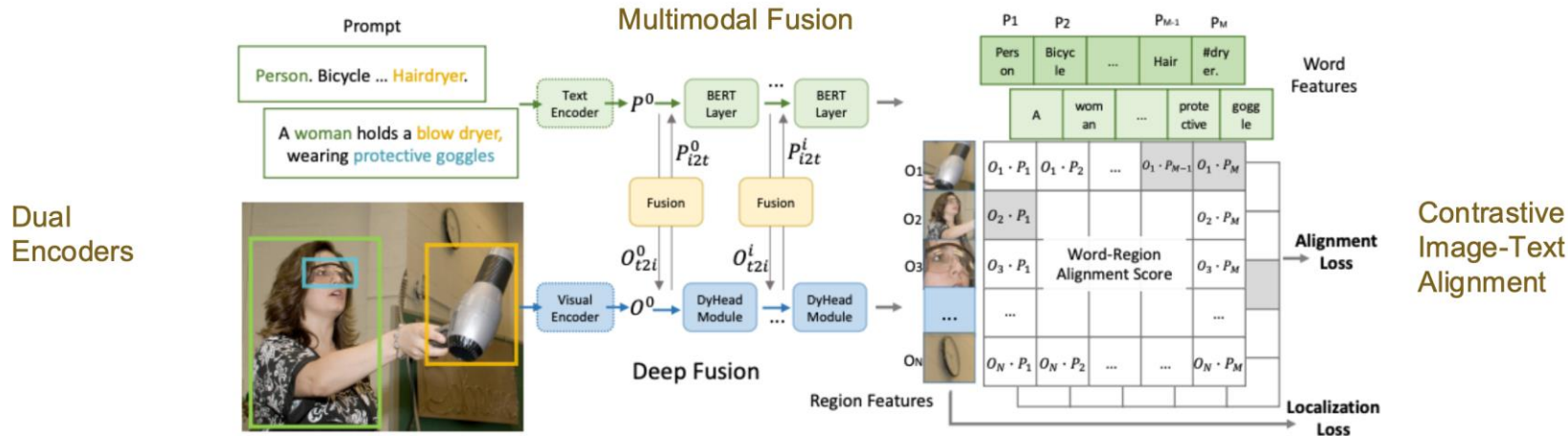
Visual Prompting.





# \* Fine-grained Capability of MLLM

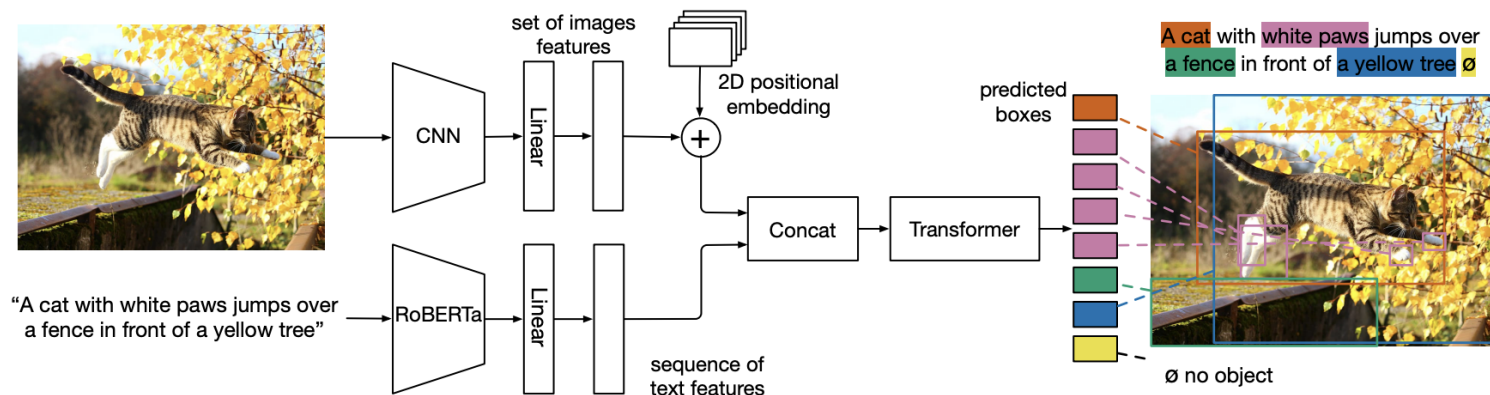
## Overview of Fine-grained Models Before LLM



GLIP: Grounded Language-Image Pre-training. 2022.

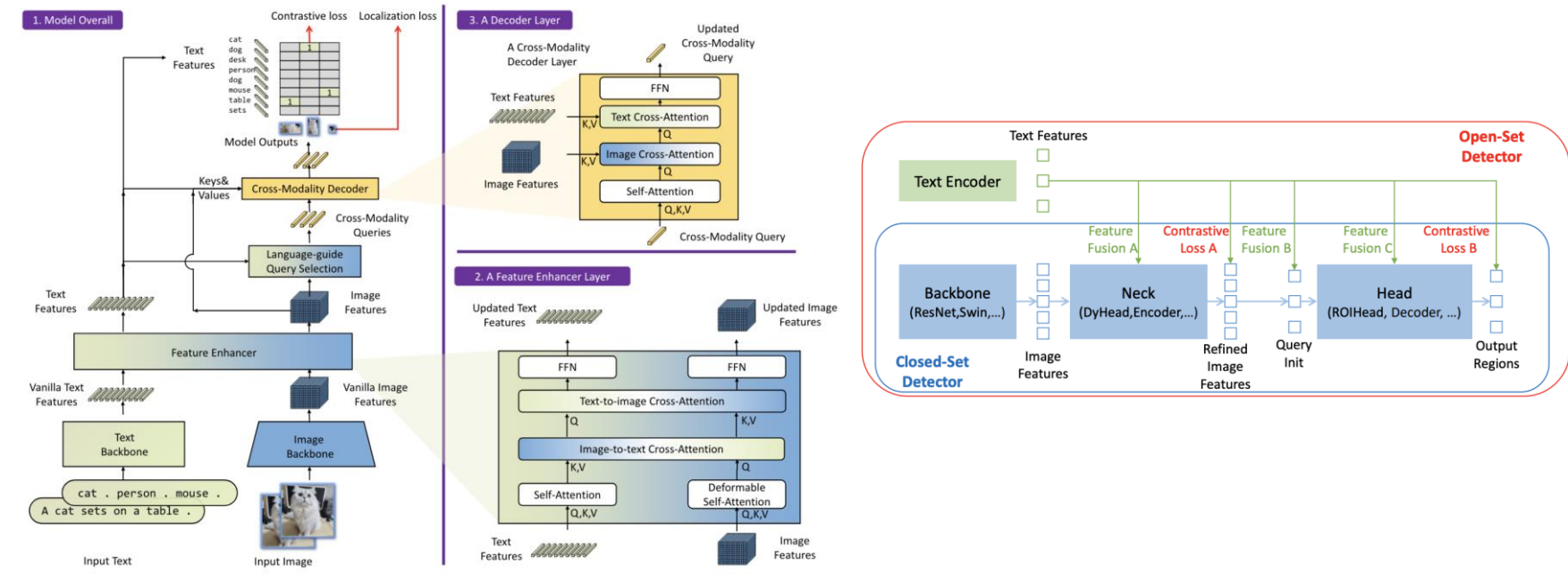
# \* Fine-grained Capability of MLLM

## Overview of Fine-grained Models Before LLM



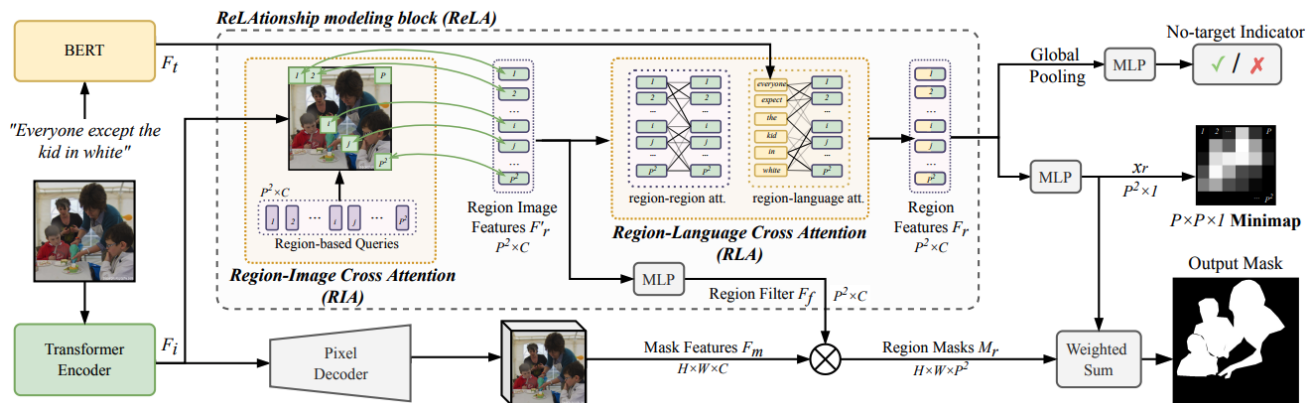
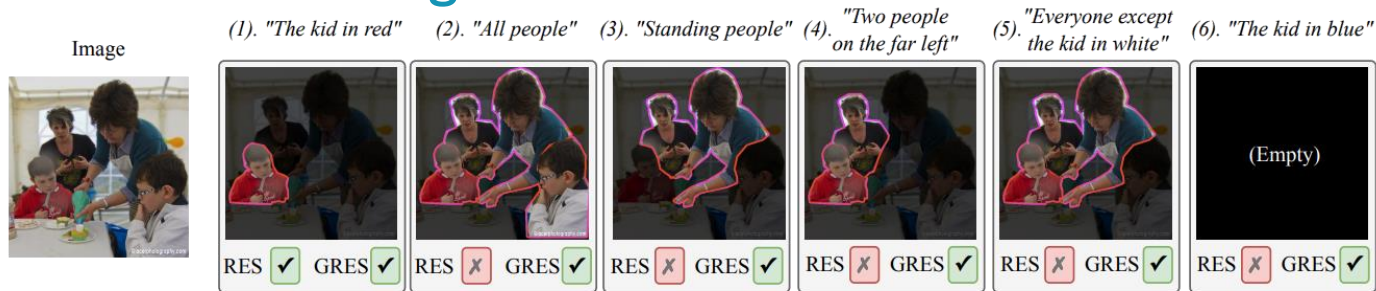
# \* Fine-grained Capability of MLLM

## Overview of Fine-grained Models Before LLM



# \* Fine-grained Capability of MLLM

## Overview of Fine-grained Models Before LLM



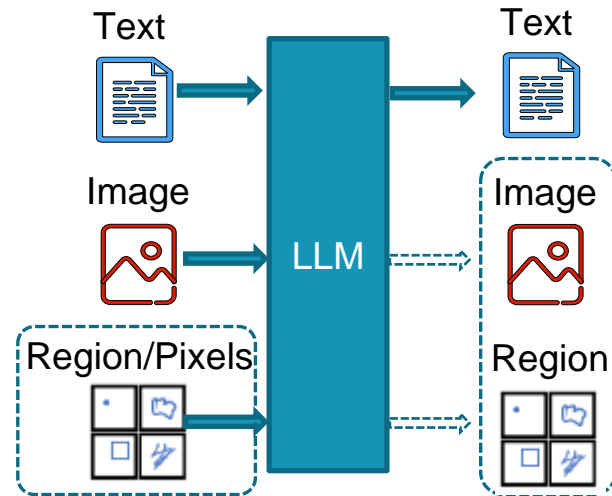
# \* Fine-grained Capability of MLLM

## Overview

- + GPT4RoI
- + NExT-Chat
- + MiniGPT-v2
- + Shikra
- + Kosmos-2
- + GLaMM
- + LISA
- + DetGPT
- + Osprey
- + PixelLM
- + OMG-LLaVA
- + VITRON
- + ...



*Users input an image (potentially specifying a region), and the LLM outputs content based on its understanding, grounding the visual content to specific pixel-level regions of the image.*



- [1] GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. 2023
- [2] NExT-Chat: An LMM for Chat, Detection and Segmentation. 2023
- [3] MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. 2023
- [4] Osprey: Pixel Understanding with Visual Instruction Tuning. 2023
- [5] GLaMM: Pixel Grounding Large Multimodal Model. 2023
- [6] Kosmos-2: Grounding Multimodal Large Language Models to the World. 2023
- [7] DetGPT: Detect What You Need via Reasoning. 2023
- [8] PixelLM: Pixel Reasoning with Large Multimodal Model. 2023
- [9] Lisa: Reasoning segmentation via large language model. 2023
- [10] Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. 2023

...

# \* Fine-grained Capability of MLLM

---

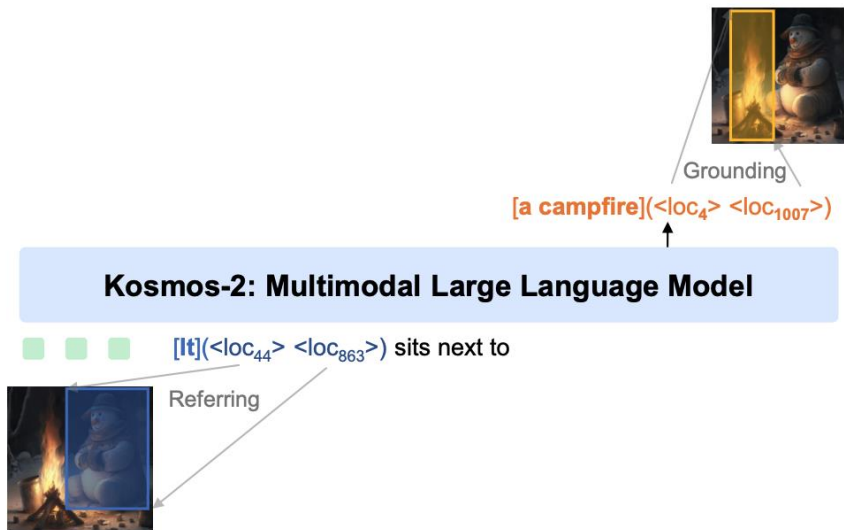
## + Fine-Grained MLLM Design

- × Overview
- × **With Visual Grounding.**
- × With Visual Segmentation.
- × Video and 3D Fine-Grained MLLM.

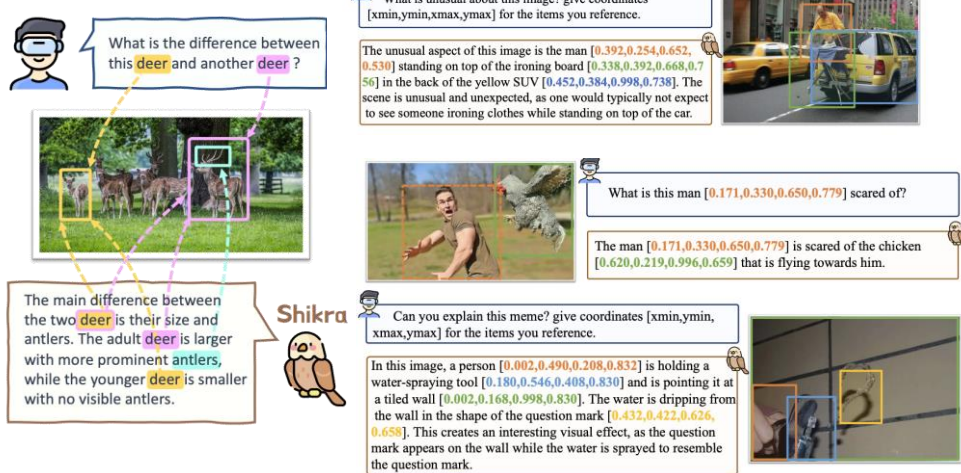
# \* Fine-grained Capability of MLLM

## With Visual Grounding.

Kosmos-2:



Shikra:

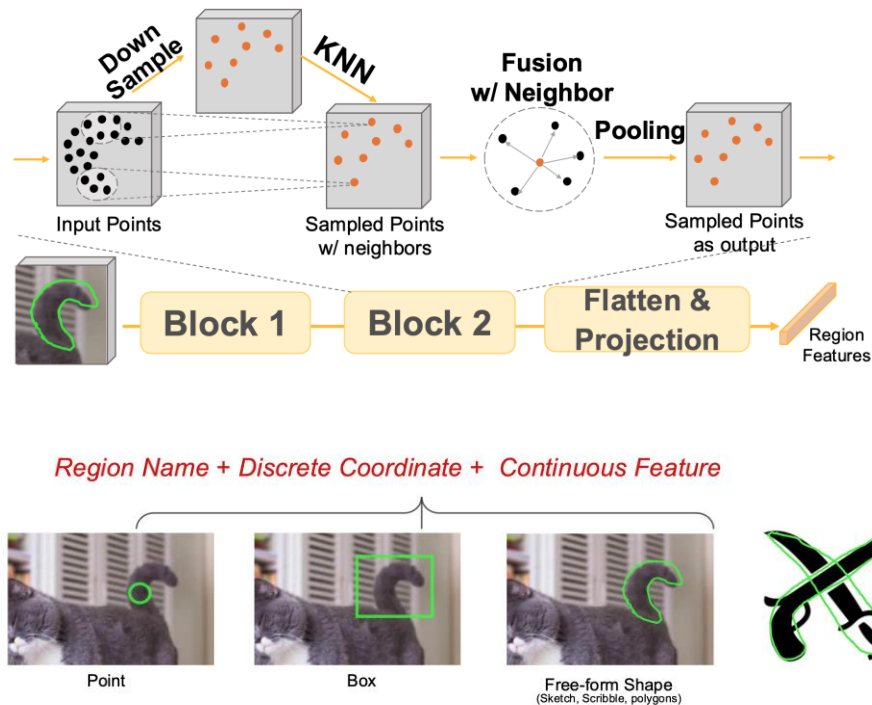
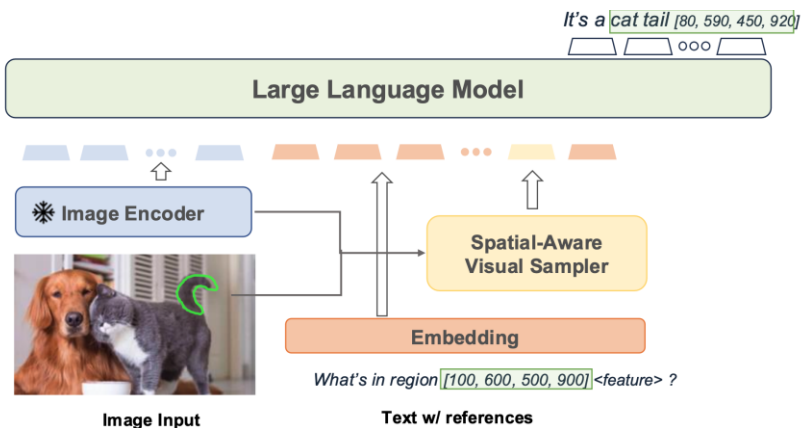


Kosmos-2: Grounding Multimodal Large Language Models to the World  
 Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic

# \* Fine-grained Capability of MLLM

With Visual Grounding.

Ferret



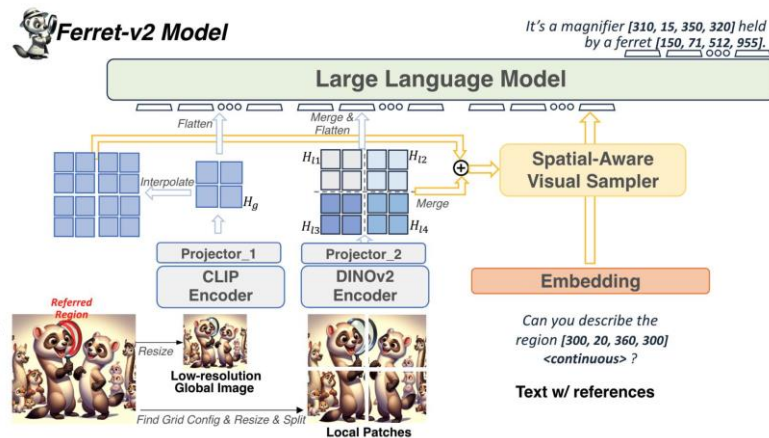
Ferret: Refer and Ground Anything Anywhere at Any Granularity, arxiv-2023.



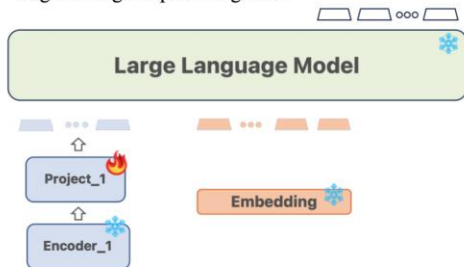
# \* Fine-grained Capability of MLLM

With Visual Grounding.

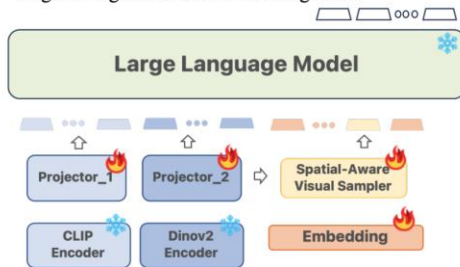
Ferret-v2



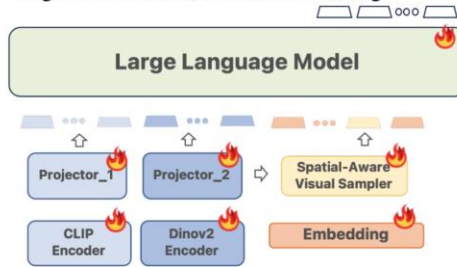
Stage I: Image-Caption Alignment



Stage II: High-resolution Dense Alignment



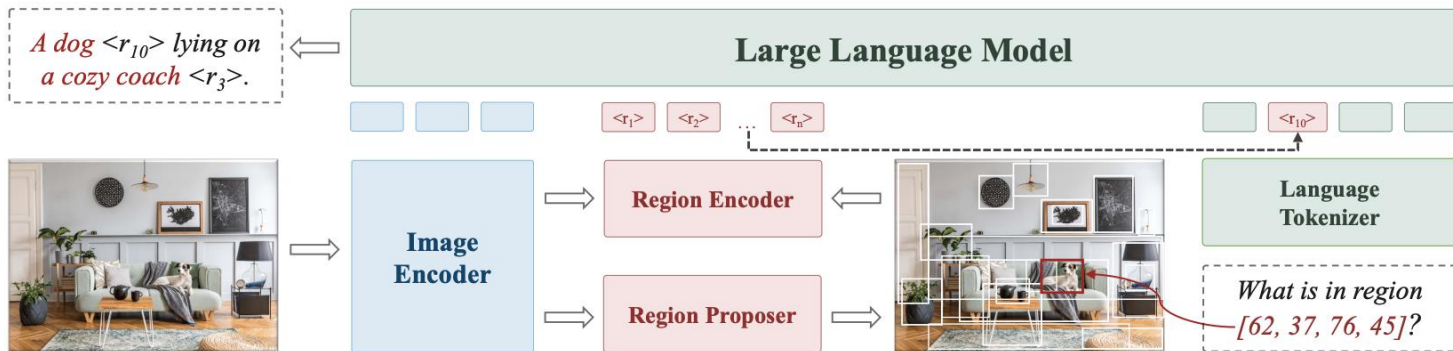
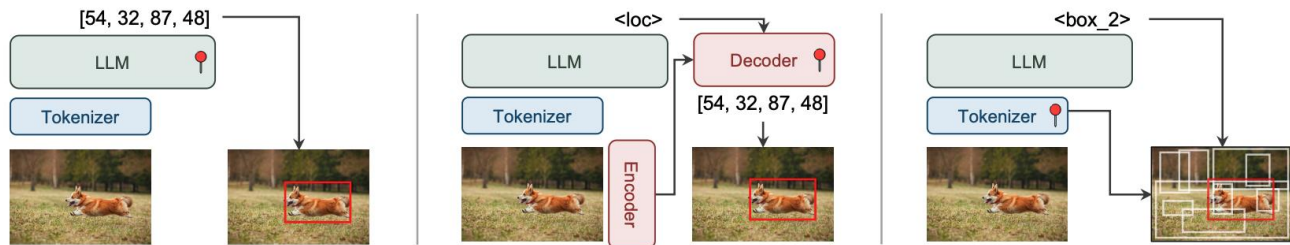
Stage III: Intent-Enhanced Instruction Tuning



# \* Fine-grained Capability of MLLM

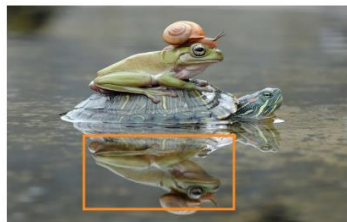
## With Visual Grounding.

Groma:



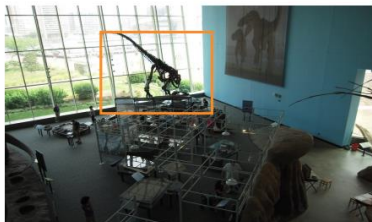
# \* Fine-grained Capability of MLLM

## With Visual Grounding.



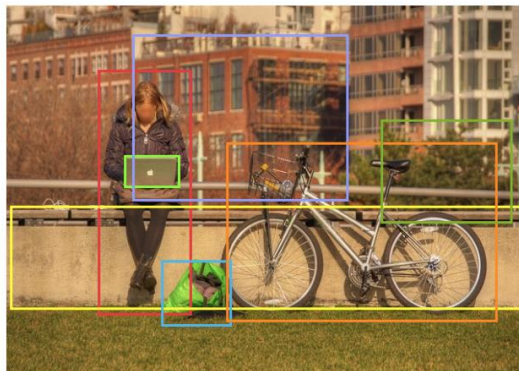
👤 Locate the the reflection of the frog.

🤖 It's located at <region>.

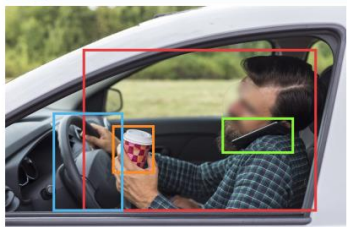


👤 Please give a brief description of <region>.

🤖 A large dinosaur skeleton.



👤 [grounding] Is it safe to drive like this?



🤖 It's not advisable. **The man** is holding **a cup** in one hand and talking on **the phone**, which means he's not using both hands on **the steering wheel**. This could be a distraction and increase the risk of an accident.

👤 [grounding] Can you describe this image in details?

🤖 In this image, we see **a woman** sitting on **a concrete bench** working on her **laptop**. She's surrounded by **a green bag** and **a white bicycle** that's leaning against the bench. Behind her, there's **a large brick building** and **a tree**.

# \* Fine-grained Capability of MLLM

---

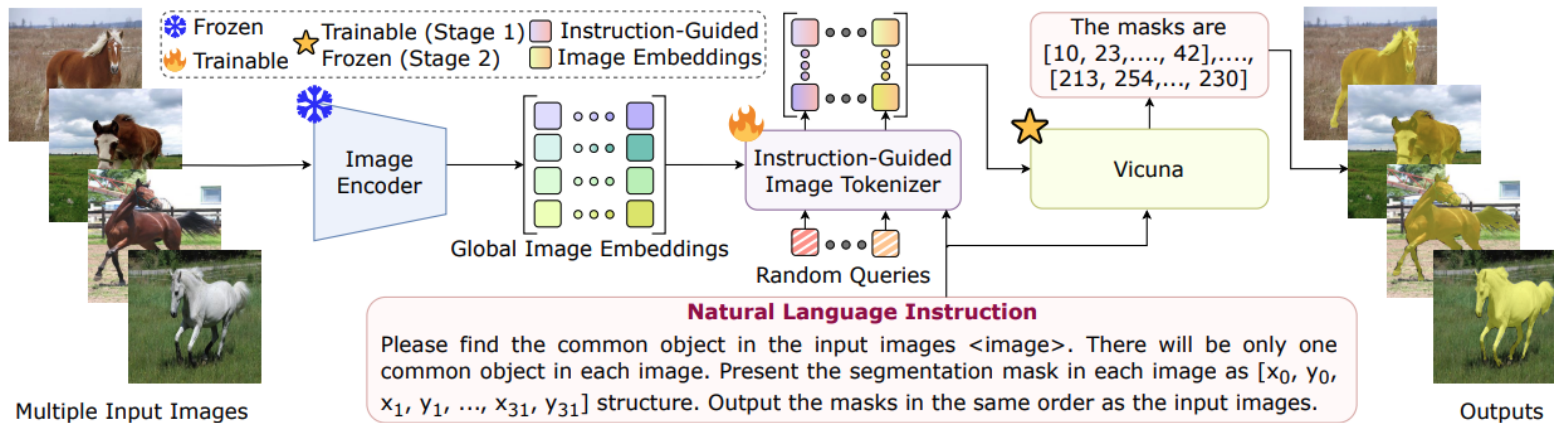
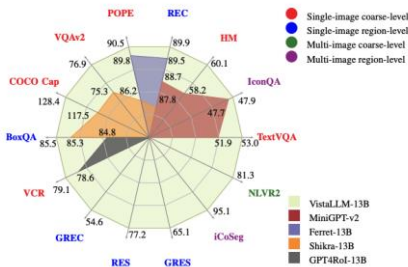
## + Fine-Grained MLLM Design

- × Overview
- × With Visual Grounding.
- × **With Visual Segmentation.**
- × Video and 3D Fine-Grained MLLM.

# \* Fine-grained Capability of MLLM

With Visual Segmentation.

VistaLLM







Jack of All Tasks, Master of Many: Designing General-purpose Coarse-to-Fine Vision-Language Model, arxiv-2023.

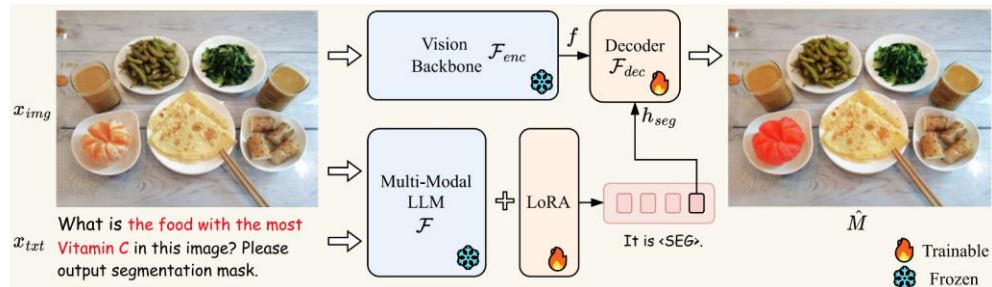
# \* Fine-grained Capability of MLLM

## With Visual Segmentation.

LISA: Large Language Instructed Segmentation Assistant

[Paper](#) | [Models](#) | [Training](#) | [Inference](#) | [Local Deployment](#) | [Dataset](#) | [Online Demo](#)


Input	Output	Input	Output
			
"Who was <u>the president of the US</u> in this image? Please output segmentation mask and explain the reason."	"Sure, the segmentation result is [SEG]. The President of the United States in the image is <u>President Obama</u> ."	"Who was <u>the president of the US</u> in this image? Please output segmentation mask and explain why."	"Sure, [SEG]. In the image, the President of the United States is <u>President Trump</u> ."



- **USER:** <IMAGE> Can you segment the tyre that does not touch the ground in this image?
- **ASSISTANT:** Sure, it is <SEG>.



- **USER:** <IMAGE> There are two washing machines as shown in the picture. If I need to do laundry, where in the picture would I put the clothes? Please output segmentation mask.
- **ASSISTANT:** <SEG>.



- **USER:** <IMAGE> Can you segment the food that tastes not spicy in this image?
- **ASSISTANT:** <SEG>.



- **USER:** <IMAGE> What is the place where the driver can observe the speed in this image? Please output segmentation mask.
- **ASSISTANT:** <SEG>.



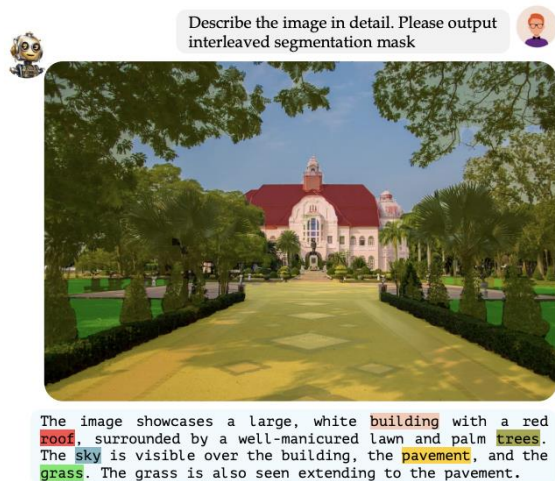
# \* Fine-grained Capability of MLLM

With Visual Segmentation.

 <p>the most likely object that someone else has left behind</p>							
 <p>the object used for stirring milk or coffee</p>							
 <p>In cold days, dogs may need extra protection to keep them warm. What object in the picture can a dog wear to provide warmth during snowy walks?</p>							
 <p>When a plane is ready to land on the airport runway, what area in the picture will it eventually land on?</p>							
Query	Image	OVSeg	GRES	X-Decoder	SEEM	Ours	GT

# \* Fine-grained Capability of MLLM

## With Visual Segmentation.



Method	Image	Input / Output		Region Enc. / Dec.	Pixel-Wise Grounding	Multi-turn Conversation	End-End Model
		Region	Multi-Region				
MM-REACT (arXiv-23) [51]	✓	✗/✗	✗/✗	✗/✗	✗	✓	✗
LLaVA (NeurIPS-23) [29]	✓	✗/✗	✗/✗	✗/✗	✗	✓	✓
miniGPT4 (arXiv-23) [61]	✓	✗/✗	✗/✗	✗/✗	✗	✓	✓
mPLUG-OWL (arXiv-23) [52]	✓	✗/✗	✗/✗	✗/✗	✗	✓	✓
LLaMA-Adapter v2 (arXiv-23) [8]	✓	✗/✗	✗/✗	✗/✗	✗	✓	✓
Otter (arXiv-23) [22]	✓	✗/✗	✗/✗	✗/✗	✗	✗	✓
Instruct-BLIP (arXiv-23) [6]	✓	✗/✗	✗/✗	✗/✗	✗	✓	✓
InternGPT (arXiv-23) [31]	✓	✓/✗	✗/✗	✗/✗	✗	✓	✗
Bubo-GPT (arXiv-23) [59]	✓	✗/✓	✗/✓	✗/✗	✗	✓	✗
Vision-LLM (arXiv-23) [44]	✓	✗/✓	✗/✓	✗/✗	✗	✗	✓
Det-GPT (arXiv-23) [36]	✓	✓/✓	✓/✓	✗/✗	✗	✓	✓
Shikra (arXiv-23) [5]	✓	✓/✓	✗/✓	✗/✗	✗	✗	✓
Kosmos-2 (arXiv-23) [35]	✓	✓/✓	✓/✓	✗/✗	✗	✗	✓
GPT4RoI (arXiv-23) [57]	✓	✓/✗	✓/✗	✓/✗	✗	✓	✓
ASM (arXiv-23) [45]	✓	✓/✗	✗/✗	✓/✗	✗	✗	✓
LISA (arXiv-23) [21]	✓	✗/✓	✗/✗	✗/✓	✓	✗	✓
GLaMM (ours)	✓	✓/✓	✓/✓	✓/✓	✓	✓	✓



# \* Fine-grained Capability of MLLM

## With Visual Segmentation.

### Objects and Attributes

- 1 dog, pub dog, a brown and white dog
- 2 dog collar, black color, chain collar
- 3 bell, cowbell
- 4 steps, stairs, the steps of a building
- 5 sack, a large white bag with black writing

### Relationships and Landmarks

A dog sitting on the steps

A large brown dog wearing a chain collar

Cowbell attached to dog collar

Landmarks: Outdoor – Urban Landscape



### Scene Graph



### Dense Grounded Caption

A large brown dog is sitting on the steps of a building. It is wearing a black chain dog collar. The collar has a cowbell attached to it. There is a bag in the background with black writings on it.

### Extra Context

Dogs, especially pugs and bulldogs, have been a part of human families for thousands of years, serving as loyal companions. They have been bred for specific traits, making them popular pets. Dogs have been trained for various tasks, including assisting people with disabilities and serving as search and rescue animals. Dog collars, often bearing identification tags, are essential for keeping pets safe and ensuring they can be returned home if lost. Cowbells, once used to signal the arrival of a cow, have been repurposed as dog collars, providing a distinct sound to help locate a dog if it wanders off. In outdoor urban landscape, dogs are often found sitting on steps, as they may choose to rest in spots that offer a good view of their surroundings.

### Level-1

Object localization and attributes

- Image Tagging and Object Detection
- Open Vocabulary Detection
- Region Attribute Detection

### Level-2

Relationships

- Short Captions and Phrase extraction
- Grounding expression
- Landmarks

### Level-3

Scene Graph & Dense Captioning

- Hierarchical Scene Graph
- In-contex Learning with LLM
- Verification Pipeline

### Level-4

Extra Contextual Insights

- Lanmark Details
- History and Background
- Precautionary Measures

Model	Validation Set					Test Set				
	M	C	AP50	mIoU	Recall	M	C	AP50	mIoU	Recall
BuboGPT [59]	17.2	3.6	19.1	54.0	29.4	17.1	3.5	17.3	54.1	27.0
Kosmos-2 [35]	16.1	27.6	17.1	55.6	28.3	15.8	27.2	17.2	56.8	29.0
LISA* [21]	13.0	33.9	25.2	62.0	36.3	12.9	32.2	24.8	61.7	35.5
GLaMM†	15.2	43.1	28.9	65.8	39.6	14.6	37.9	27.2	64.6	38.0
GLaMM	16.2	47.2	30.8	66.3	41.8	15.8	43.5	29.2	65.6	40.8

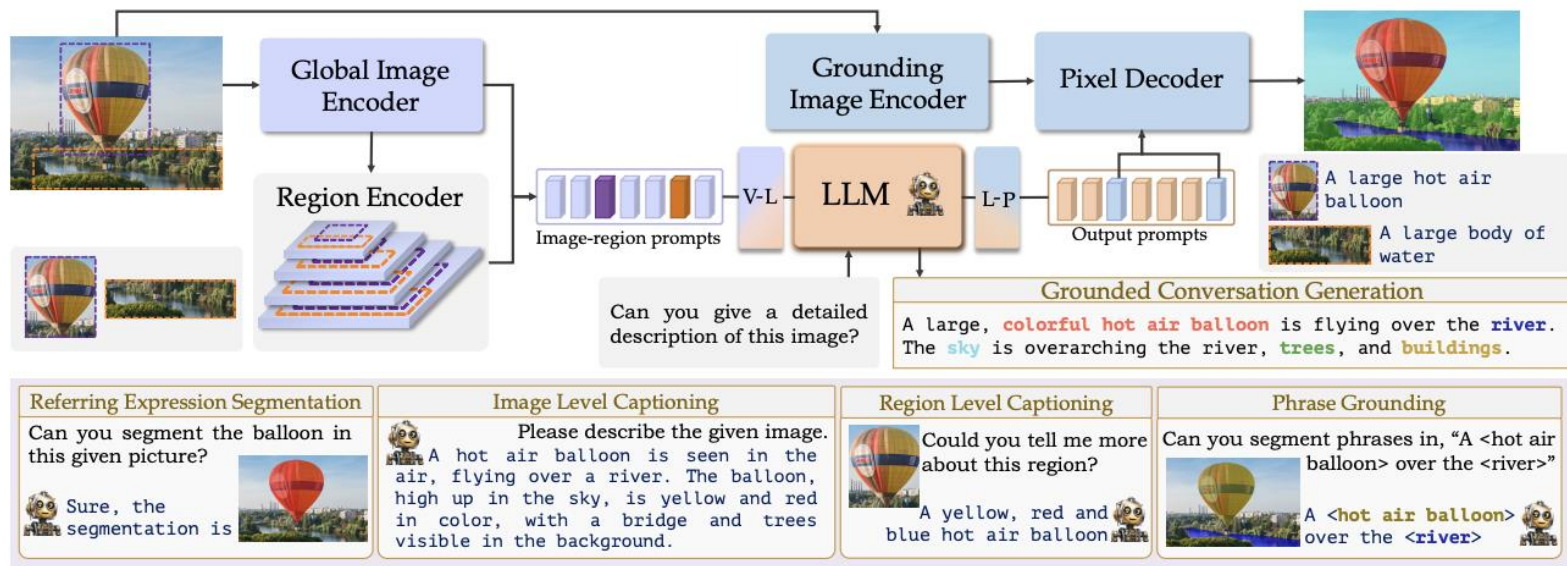
Table 3. Performance on GCG Task: Metrics include METEOR (M), CIDEr (C), AP50, mIoU, and Mask Recall. LISA\* denotes LISA adapted for GCG. GLaMM† denotes training excluding 1K human annotated images. GLaMM shows better performance.

Method	refCOCO			refCOCO+			refCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
CRIS [47]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT [50]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [26]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [63]	-	-	-	-	-	-	64.6	-
SEEM [64]	-	-	-	-	-	-	65.7	-
LISA-7B [21]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
GLaMM	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9

Table 4. Qualitative Assessment of GLaMM in Referring-Expression Segmentation: Performance across refCOCO, refCOCO+, and refCOCOg in generating accurate segmentation masks based on text-based referring expressions surpasses that of closely related work, including LISA which is specifically designed for this task.

# \* Fine-grained Capability of MLLM

## With Visual Segmentation.



# \* Fine-grained Capability of MLLM

---

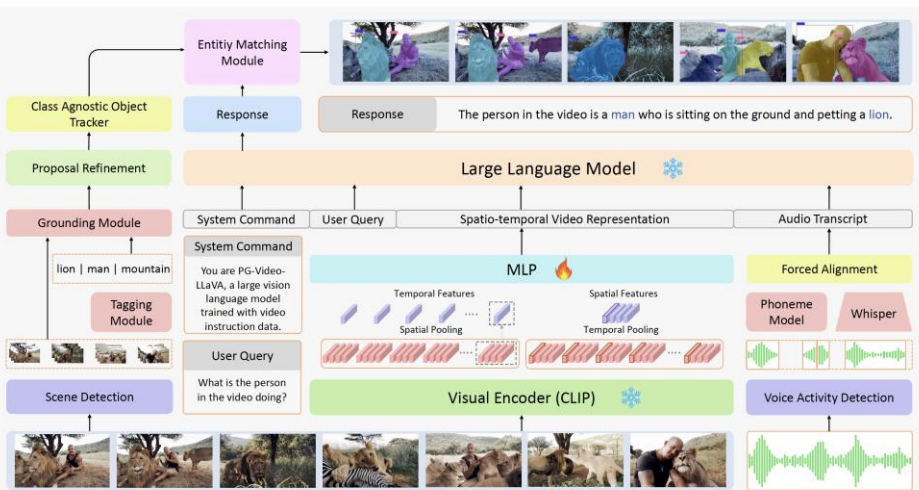
## + Fine-Grained MLLM Design

- × Overview
- × With Visual Grounding.
- × With Visual Segmentation.
- × **Video and 3D Fine-Grained MLLM.**

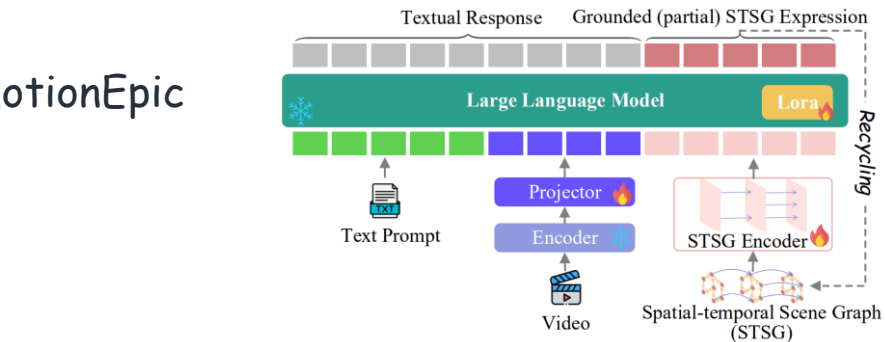
# \* Fine-grained Capability of MLLM

## Video and 3D Fine-Grained MLLM.

### PG-Video-LLaVA



### MotionEpic



**Question:** What is the least likely category for the animal in this video?  
A. Police Dog  Competitive Animal  Video-LLaVA  Circus Performer  Video-ChatGPT  D. Companion Pet  MotionEpic  E. Search and Rescue Dog

**Step-1:** The involved target is [dog].

**Step-2:** The partial STSG in tracking [dog] is:

**Step-3:** According to the video scene and STSG, the dog is **crossing multiple hurdles** with the dog being visible both before and after the hurdles. The accompanying man is observed providing instructions to guide the dog through the obstacles... Drawing on factual commonsense understanding, it might be inferred that the man is a trainer who is **impacting various commands and training the dog on a grassy field.**

**Step-4:**  
The video depicting professional training and complex actions suggests it might be a police dog performing daily training ... The rationality of the answer [A. Police Dog] is 2.  
The companion dog is to support companionship and emotional support to their owners rather than engaging in specialized tasks ... The answer [D. Companion Pet] has a coherence score of 8.

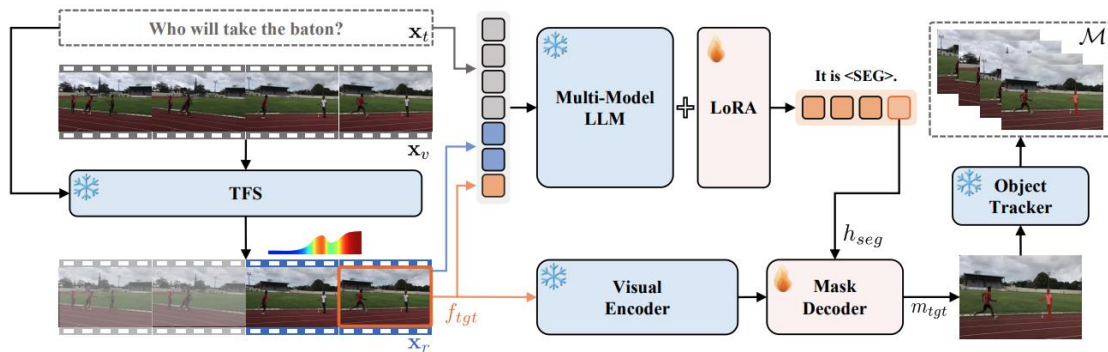
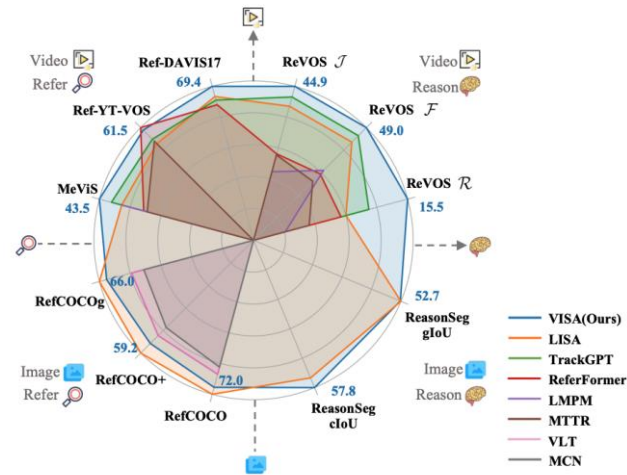
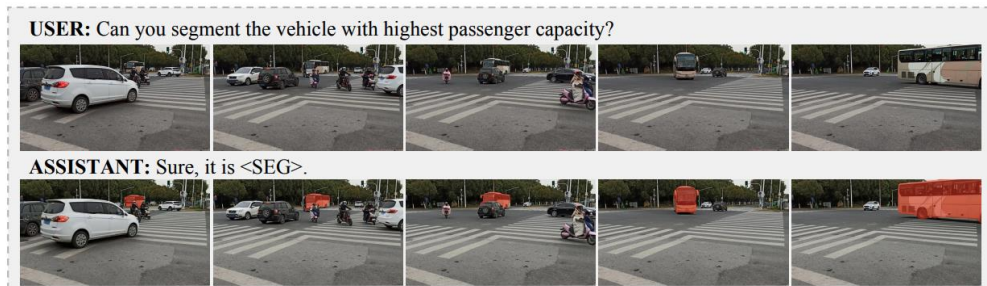
After ranking the rationale score, the final answer is [D. Companion Pet].

**Step-5:** Let's verify the [D. Companion Pet] based on visual perception ...  
1. Pixel Grounding Information Check: Based on the video scene, it depicts a training ground with a dog, so the answer is fitting.  
2. Commonsense Check: Observing the dog's energetic behavior during training aligns with the common understanding that companion pet are less likely to undergo such training, supporting the chose answer.  
Conclusion: The answer [D. Companion Pet] is supported both by ...

[1] PG-Video-LLaVA: Pixel Grounding in Large Multimodal Video Models. Arxiv-2023  
[2] Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. Arxiv-2024

# \* Fine-grained Capability of MLLM

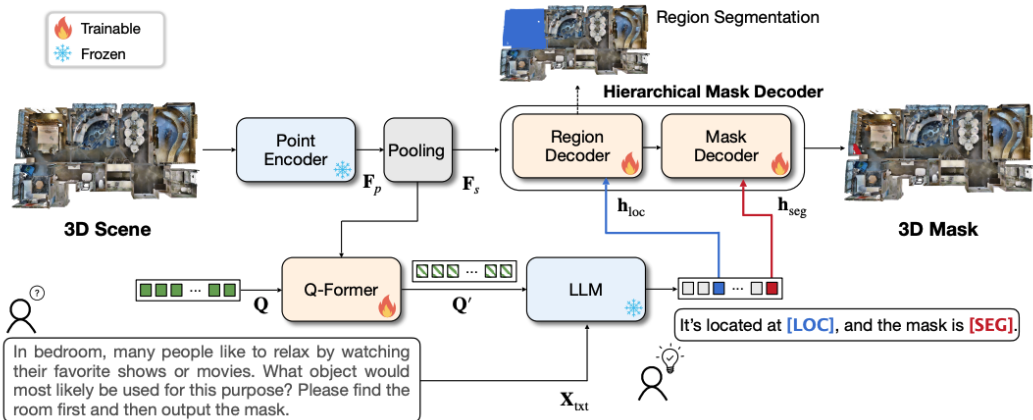
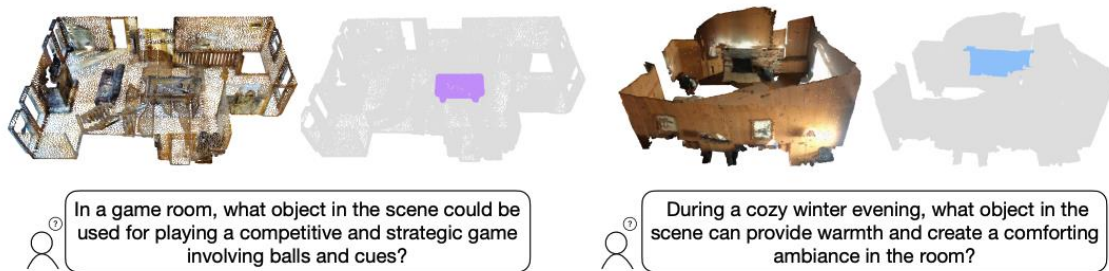
## Video and 3D Fine-Grained MLLM.



# \* Fine-grained Capability of MLLM


Video and 3D Fine-Grained MLLM.

Reason3D



# \* Fine-grained Capability of MLLM

## Video and 3D Fine-Grained MLLM.



**Grounded 3D-LLM**

**Language Grounding**

D<sub>o</sub> Your task involves locating something. Here's the detail: this is a reddish brown leather armchair; it is sitting in front of the door and just to the left of the closet.

A<sub>i</sub> I've pinpointed the `<p> object </p>` you're looking for.

**Dense Captioning**

D<sub>o</sub> Begin by outlining the visual characteristics of the `object <ref>`, followed by an account of its position relative to the objects around it.

A<sub>i</sub> this is a brown chair. it is against the brown table.

D<sub>o</sub> Can you outline the main features of the `object <ref>` succinctly?

A<sub>i</sub> A brown wooden desk with a flat top surface.

**Question Answering**

D<sub>o</sub> What is on a small table? Please answer with a single word or phrase.

A<sub>i</sub> lamp.

D<sub>o</sub> Where is the pillow on the bed? Please answer with a single word or phrase.

A<sub>i</sub> on top of bed.

**Embodied Dialog**

D<sub>o</sub> Can you please fetch the printer from the `nightstand <ref>` and place it on `<ref>`?

A<sub>i</sub> Certainly. I will move the `<p> the printer </p>` to `<p> the table </p>` right away.

D<sub>o</sub> Also, what is the design purpose of the cylindrical lamp with a white lampshade on `<ref>`? It looks intriguing.

A<sub>i</sub> `<p> The cylindrical lamp </p>` is designed to provide focused lighting on `<p> the desk </p>` for tasks like reading or working at the computer ...

**Embodied Planning**

D<sub>o</sub> I desire to enhance room lighting. Can you assist me with this?

A<sub>i</sub> Got it. Let's begin the planning process now.  
step 1. Turn off `<p> lamp </p>`.  
step 2. Dust `<p> lamp </p>` shade.  
step 3. Replace bulb in `<p> lamp </p>`.  
step 4. Turn on `<p> lamp </p>` to test light.  
Planning stage complete. What's our next objective?

**Object Detection**

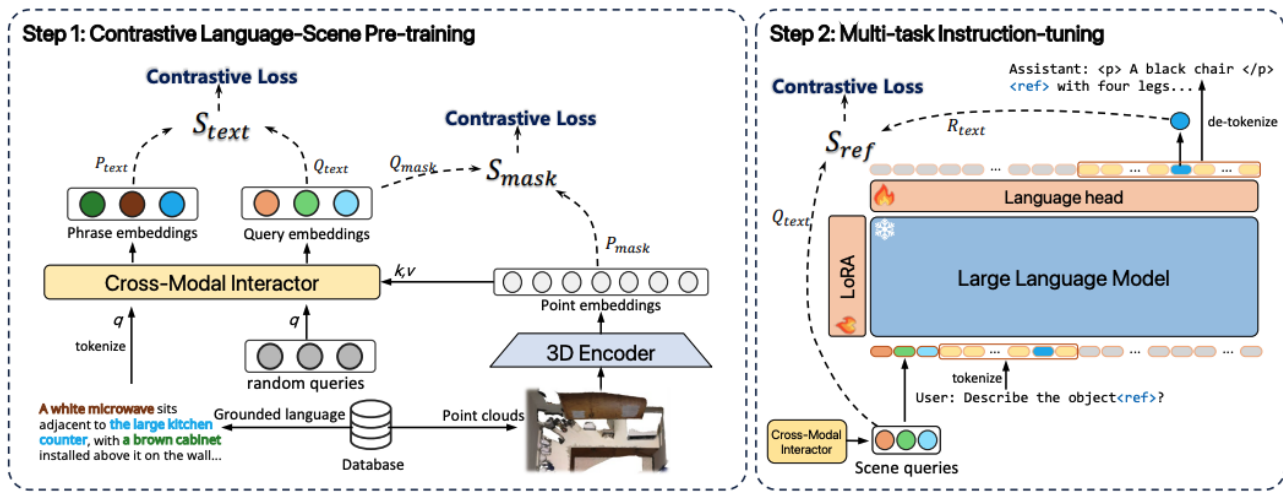
D<sub>o</sub> List all the chairs visible in this indoor area.

A<sub>i</sub> Yes, I was able to locate `<p> chairs </p>` in the space for you.

# \* Fine-grained Capability of MLLM

## Video and 3D Fine-Grained MLLM.

Method	LLM	Prompts		Tasks						
		Text	Vision	Inst.Seg.	Obj.Det.	Grd.	Point-Grd.	Multi-Obj Grd.	QA	Cap
PointGroup [37]	✗	-	-	✓	✓	✗	✗	✗	✗	✗
Mask3D [62]	✗	-	-	✗	✗	✓	✗	✓	✗	✗
Multi3DRef [83]	✗	-	-	✗	✗	✓	✗	✓	✗	✗
BUTD-DETR [36]	✗	-	-	✗	✓	✓	✗	✓	✗	✗
3D-VisTA [88]	✗	-	-	✗	✗	✓	✗	✗	✗	✓
Chat-3D [72]	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓
Chat-3D v2 [33]	✓	✓	✗	✗	✗	✓	✗	✓	✓	✓
3D-LLM [31]	✓	✓	✗	✗	✗	✓	✗	✓	✓	✓
LL3DA [11]	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓
Grounded 3D-LLM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓





# \* Recent Advanced MLLM Designs

---

## + Advanced MLLM Design

- × **Overview**
- × Unified Architecture Designs.
- × MLLM For Long Video Analysis.
- × MLLM With MOE Design.

# \* Recent Advanced MLLM Designs

---

## Overview of Recent Advanced MLLM Designs:

### 1, More Functionalities:

- One model For All Language Driven Vision Tasks.
- Mutual Cross-Task Benefits.

### 2, Long Video Analysis:

- Temporal Modeling For Extremely Long Video.
- Efficient Long Context Modeling.

### 3, Multi-Experts Models:

- Mixture of Experts (MoE) architecture.
- Better performance and enhanced capacity.

# \* Recent Advanced MLLM Designs

---

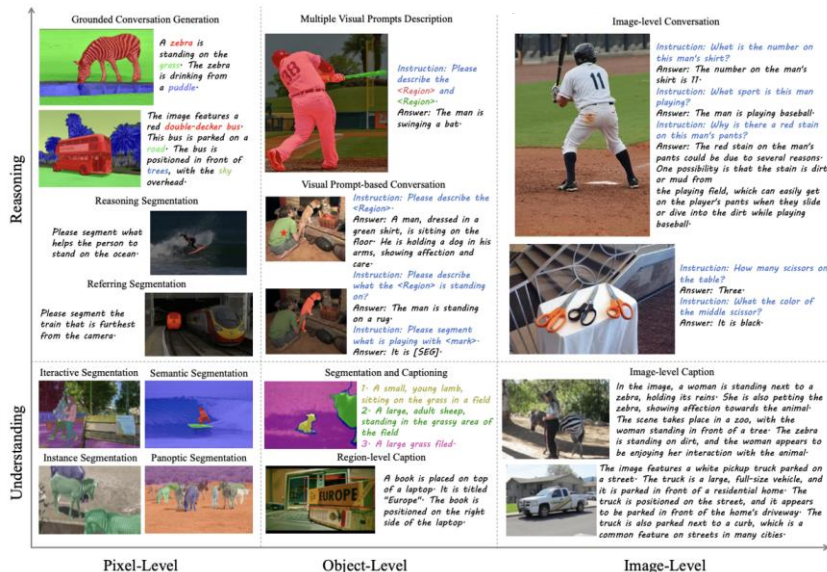
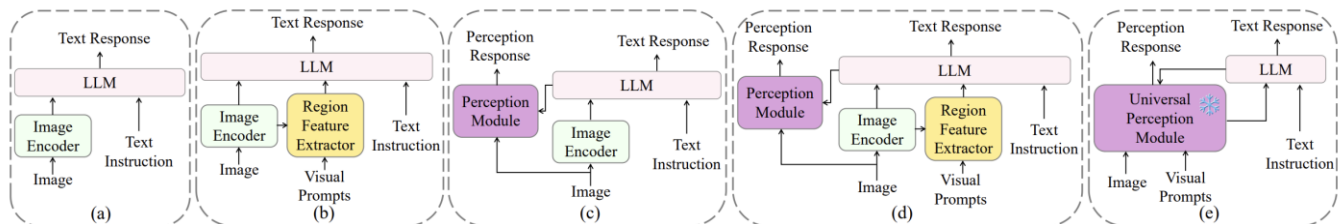
## + **Advanced MLLM Design**

- × Overview
- × Unified Architecture Designs.
- × MLLM For Long Video Analysis.
- × MLLM With MOE Design.

# \* Recent Advanced MLLM Designs

## Unified Architecture

### OMG-LLaVA



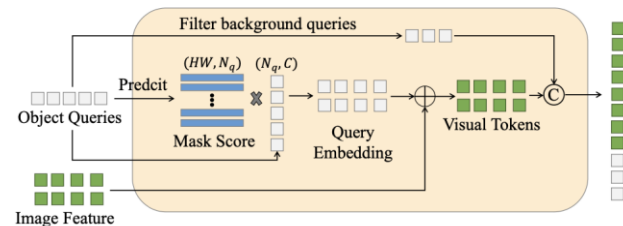
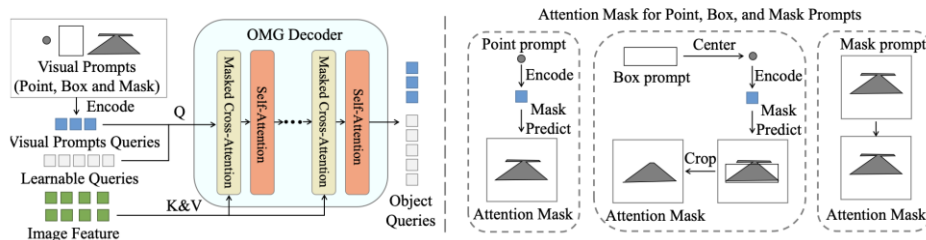
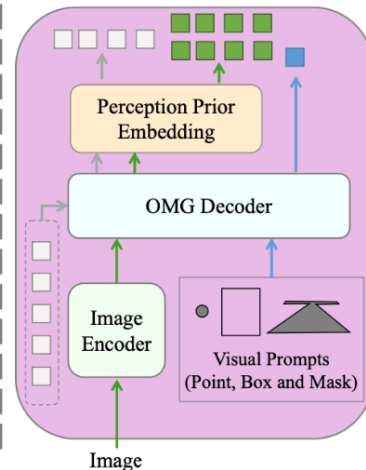
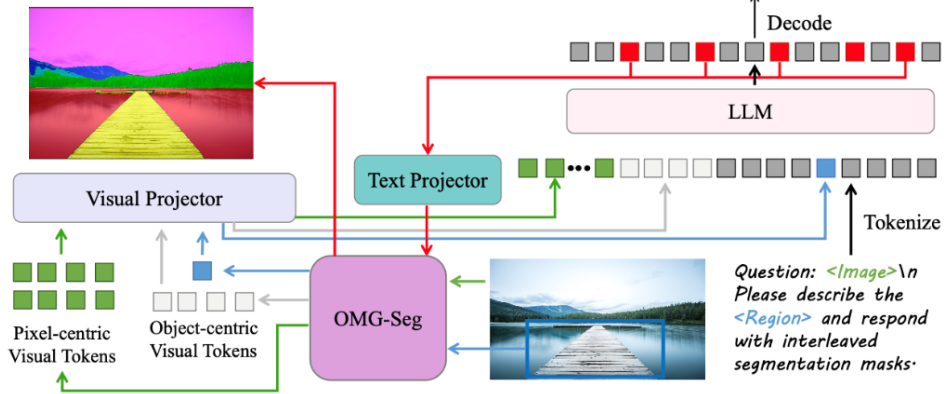
Method	Visual Encoder	Image-level Caption	Image-level Conversation	Object-level			Pixel-level		
				Visual Prompts	Caption	Conversation	Universal Seg	RES	GCG
LLaVA [69]	1	✓	✓						
MiniGPT4 [140]	1	✓	✓						
mPLUG-Owl [116]	1	✓	✓						
LLaMA-Adapter [130]	1	✓	✓						
Mini-Gemini [63]	2	✓	✓						
InternVL 1.5 [18]	1	✓	✓						
VisionLLM [95]	1	✓	✓						
Shikra [13]	1	✓	✓	Point & Box	✓	✓			✓
Kosmos-2 [80]	1	✓	✓	Box	✓	✓			
GPT4oL [131]	1	✓	✓	Box	✓	✓			
Ferret [117]	1	✓	✓	Point & Box & Mask	✓	✓			
Osprey [124]	1	✓	✓	Mask	✓	✓			
SPHINX-V [65]	1	✓	✓	Point & Box & Mask	✓	✓			
LISA [47]	2	✓	✓						✓
GLAMM [85]	2	✓	✓	Box	✓	✓			✓
Groundhog [132]	4	✓	✓	Point & Box & Mask	✓	✓			✓
AnyRef [33]	2	✓	✓	Box	✓	✓			✓
PixelLM [86]	1	✓	✓						✓
GSVA [107]	2	✓	✓						✓
Groma [76]	1	✓	✓						✓
VIP-LLaVA [8]	1	✓	✓	Box	✓	✓			✓
PSALM [133]	1	✓	✓	Point & Box & Mask	✓	✓			✓
LaSagna [100]	2	✓	✓	Point & Box & Mask	✓	✓			✓
OMG-Seg [56]	1	✓	✓	Point	✓	✓			✓
OMG-LLaVA	1	✓	✓	Point & Box & Mask	✓	✓	✓	✓	✓

OMG-LLaVA : Bridging Image-level, Object-level, Pixel-level Reasoning and Understanding, arxiv-2024.

# \* Recent Advanced MLLM Designs

## Unified Architecture

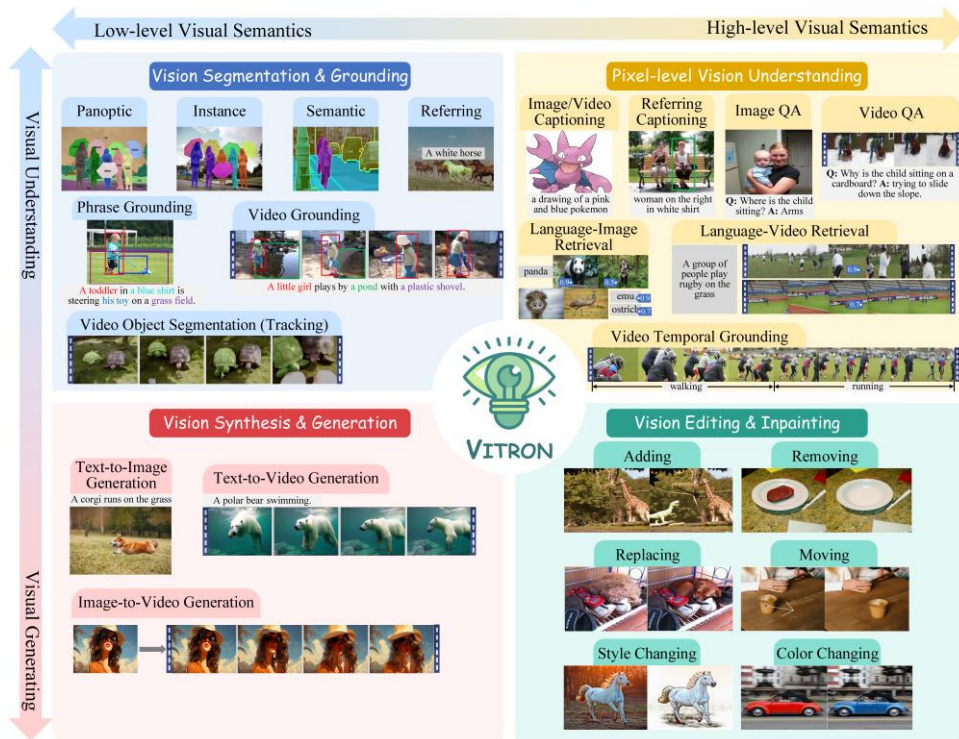
Answer: It's a wooden <p> pier </p> [SEG], surrounded by the <p> lake </p> [SEG] and the <p> forest </p> [SEG]. A blue <p> sky </p> [SEG] is above the pier and the <p> mountains </p> [SEG].



# \* Recent Advanced MLLM Designs

- Unified Pixel-wise MLLM

+ Vitron



# \* Recent Advanced MLLM Designs

Model	Vision Supporting		Pixel/Regional Understanding	Segmenting/ Grounding	Generating	Editing
	Image	Video				
Flamingo [1]	✓	✗	✗	✗	✗	✗
BLIP-2 [45]	✓	✗	✗	✗	✗	✗
MiniGPT-4 [126]	✓	✗	✗	✗	✗	✗
LLaVA [57]	✓	✗	✗	✗	✗	✗
GILL [39]	✓	✗	✗	✗	✓	✗
Emu [90]	✓	✗	✗	✗	✓	✗
MiniGPT-5 [125]	✓	✗	✗	✗	✓	✗
DreamLLM [23]	✓	✗	✗	✗	✓	✗
GPT4RoI [122]	✓	✗	✓	✓	✗	✗
NExT-Chat [118]	✓	✗	✓	✓	✗	✗
MiniGPT-v2 [13]	✓	✗	✓	✓	✗	✗
Shikra [14]	✓	✗	✓	✓	✗	✗
Kosmos-2 [72]	✓	✗	✓	✓	✗	✗
GLaMM [78]	✓	✗	✓	✓	✗	✗
Osprey [117]	✓	✗	✓	✓	✗	✗
PixelLM [79]	✓	✗	✓	✓	✗	✗
LLaVA-Plus [58]	✓	✗	✗	✓	✓	✓
VideoChat [46]	✗	✓	✗	✗	✗	✗
Video-LLaMA [120]	✗	✓	✗	✗	✗	✗
Video-LLaVA [52]	✓	✓	✗	✗	✗	✗
Video-ChatGPT [61]	✗	✓	✗	✗	✗	✗
GPT4Video [99]	✗	✓	✗	✗	✓	✗
PG-Video-LLaVA [67]	✗	✓	✓	✓	✗	✗
NExT-GPT [104]	✓	✓	✗	✗	✓	✗
VITRON (Ours)	✓	✓	✓	✓	✓	✓

# \* Recent Advanced MLLM Designs

- Unified Pixel-wise MLLM

Vitron

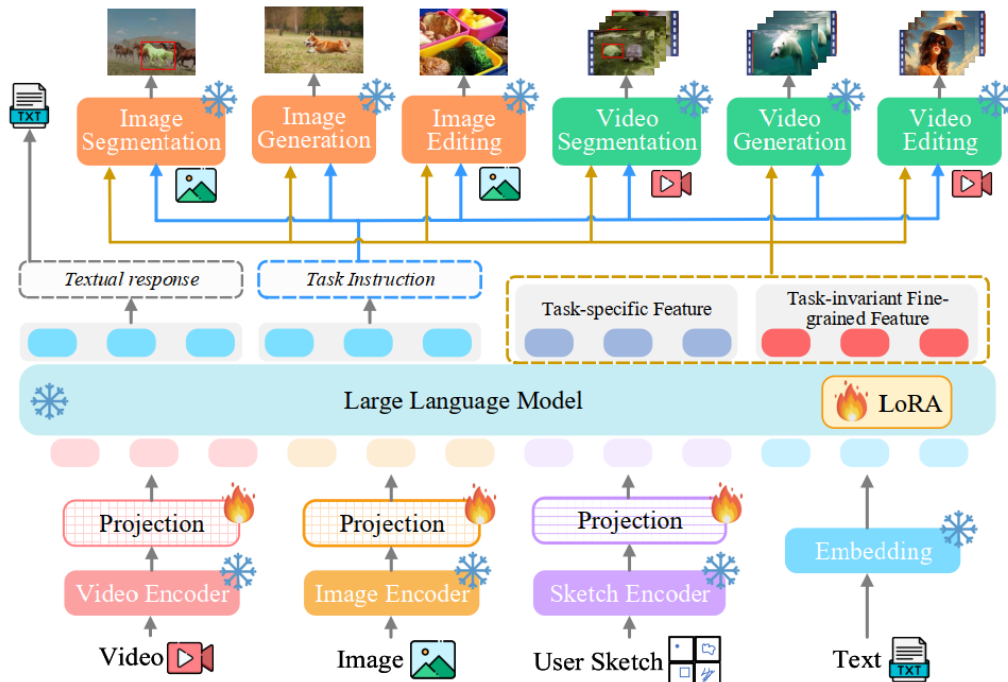


Figure 2: Technical overview of the VITRON framework.



□ Image Segmentation

□ Video Segmentation

□ Video Understanding

□ Video Editing

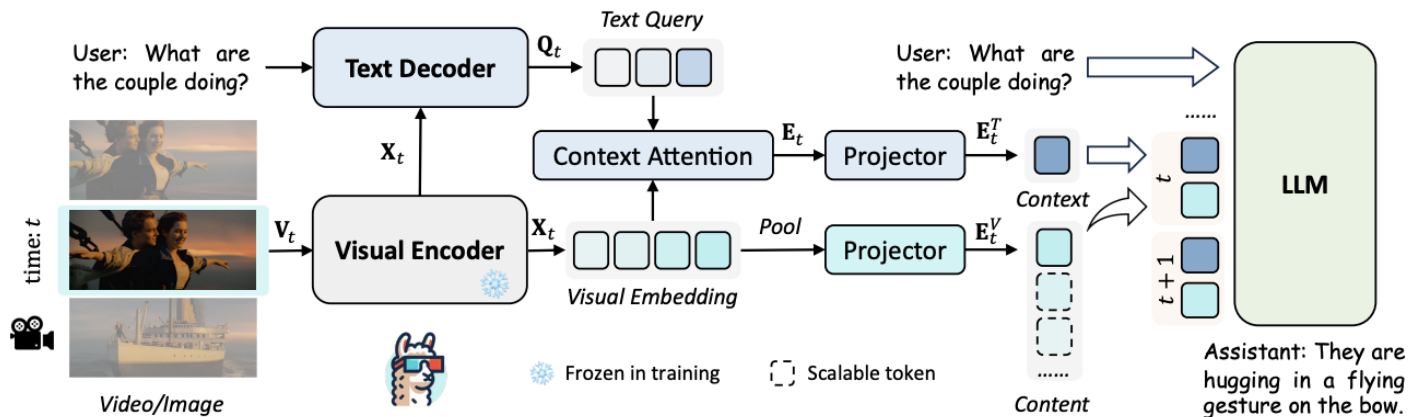
# \* Recent Advanced MLLM Designs

---

## + Advanced MLLM Design

- × Overview
- × Unified Architecture Designs.
- × **MLLM For Long Video Analysis.**
- × MLLM With MOE Design.

# \* Recent Advanced MLLM Designs



## Stage 1: Modality Alignment

232K 558K  
 User: <image-O>...<image-I>, Assistant: <caption>  
 User: <image>, Assistant: <caption>

## Stage 2: Instruction Tuning

98K 625K 40K  
 User: <prompt>\n<image-O>...<image-I>, Assistant: <answer>  
 User: <prompt>\n<image>, Assistant: <answer>.....(multi-turn)  
 User: <prompt>, Assistant: <answer>.....(multi-turn)

## Stage 3: Long Video Tuning (Optional)

9K 6K  
 User: <prompt>\n<image-O><subtitle-O>...<image-I><subtitle-I>,  
 Assistant: <answer>  
 User: <prompt>, Assistant: <answer>



User: Please describe this image in detail.

Assistant: The image features a young man in a blue outfit, likely from the game series "The Legend of Zelda," standing on top of a large rock formation. He appears to be looking out over the landscape, possibly contemplating his next move or enjoying the view. The rock formation is surrounded by a beautiful blue sky, with clouds scattered throughout the scene.....



User: Please describe this video in detail.

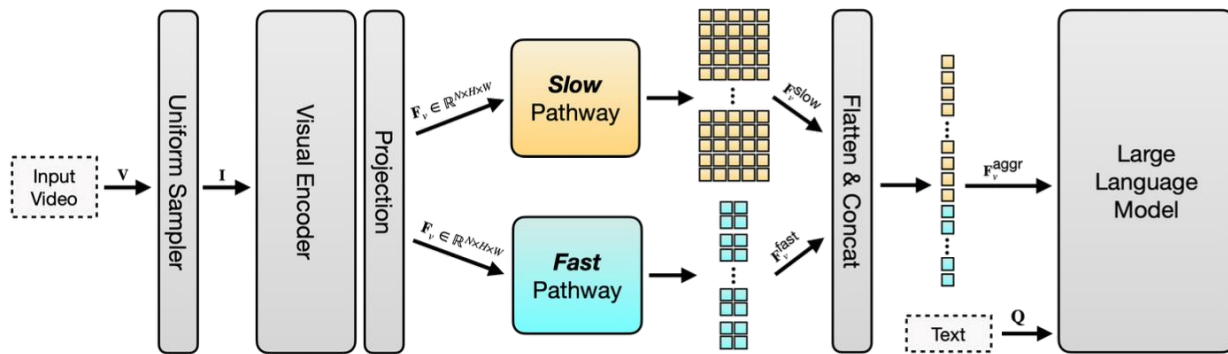
Assistant: The video starts with a cartoon character standing in front of a castle, which is surrounded by a beautiful landscape. The character is holding a bucket and appears to be observing the surroundings and enjoying the view. The scene then transitions to a train, where the character is seen sitting inside the train. The train is surrounded by lush greenery and appears to be moving through a scenic route.....



User: Summarize the plot of this movie.

Assistant: Avatar follows Jake Sully, a disabled former Marine who is recruited by the RDA to work on Pandora, a moon rich in a valuable mineral called unobtainium. Jake's twin brother, Tommy, died in a helmet-cam accident, and Jake is offered his place in the Avatar Program, which allows humans to link with the native Na'vi, a humanoid species. Jake is assigned to the avatar of his deceased brother.....

# \* Recent Advanced MLLM Designs



**Question 1:** What is happening in this video?

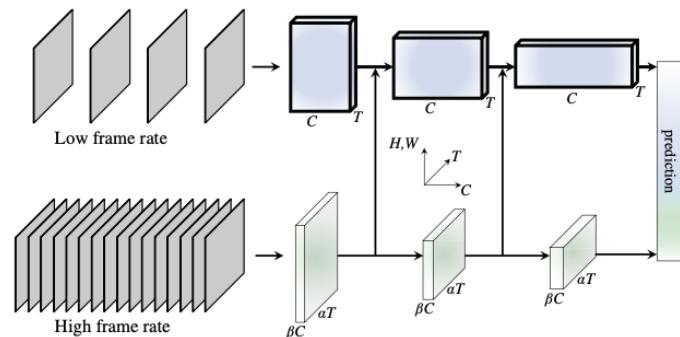
**SlowFast-LLaVA-7B:** The video depicts a person preparing a cup of coffee, which is a common activity in the morning or during breaks. The coffee machine is a modern, sleek design with a digital display. The kitchen is well-equipped with various appliances and utensils. There's a sink with a faucet, a dishwasher, and a refrigerator. On the countertop, there are several items including a bowl, a spoon, and a cup. The overall content of the video is a snapshot of everyday life, capturing the simple yet essential act of making coffee in a home setting.

**Question 2:** When does the person wash the portafilter?

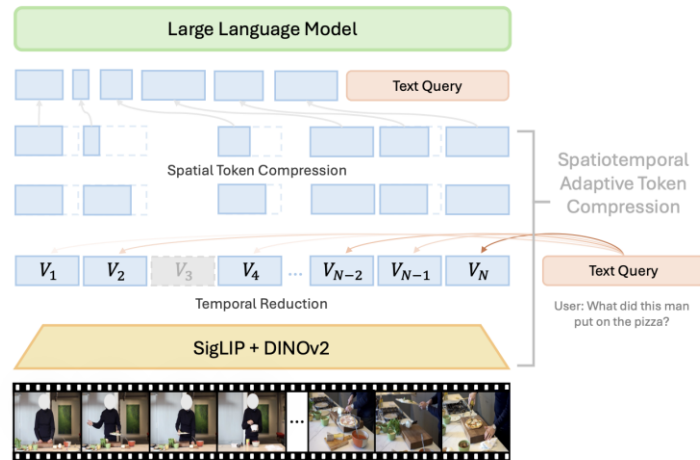
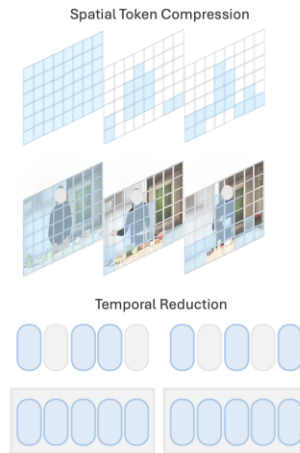
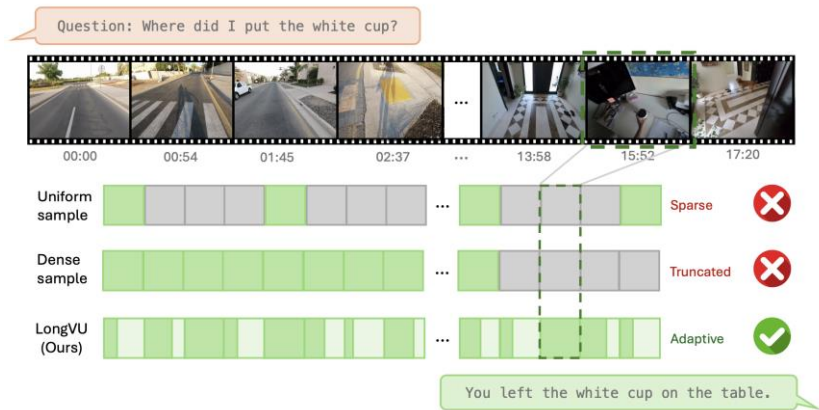
**SlowFast-LLaVA-7B:** The person washes the portafilter at the end of the video, after preparing the coffee.

**Question 3:** Does this person open the refrigerator?

**SlowFast-LLaVA-7B:** No, the person in the video does not open the refrigerator.



# \* Recent Advanced MLLM Designs



$$\mathbf{Top}_{N_h} \left( \frac{1}{H_h W_h L_q} \sum_{h,w,l} \mathcal{F}(V) Q^T \right), \quad N_h = \max \left( 0, \frac{L_{\max} - L_q - TH_l W_l}{H_h W_h - H_l W_l} \right),$$

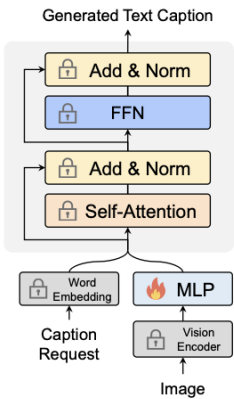
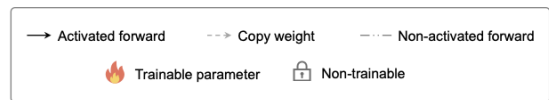
# \* Recent Advanced MLLM Designs

---

## + Advanced MLLM Design

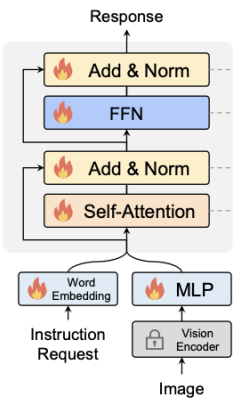
- × Overview
- × Unified Architecture Designs.
- × MLLM For Long Video Analysis.
- × **MLLM With MOE Design.**

# \* Recent Advanced MLLM Designs



(a) Stage I

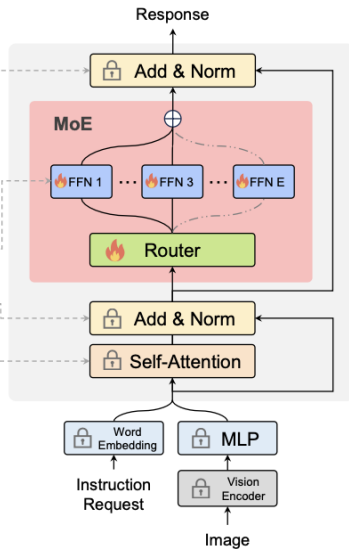
Copy weight



(b) Stage II

Copy weight

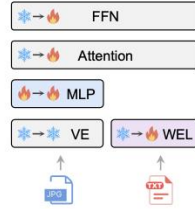
Copy weight



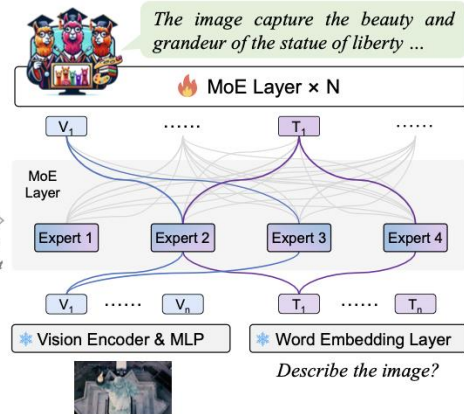
(c) Stage III



Convert LLM to LVLM



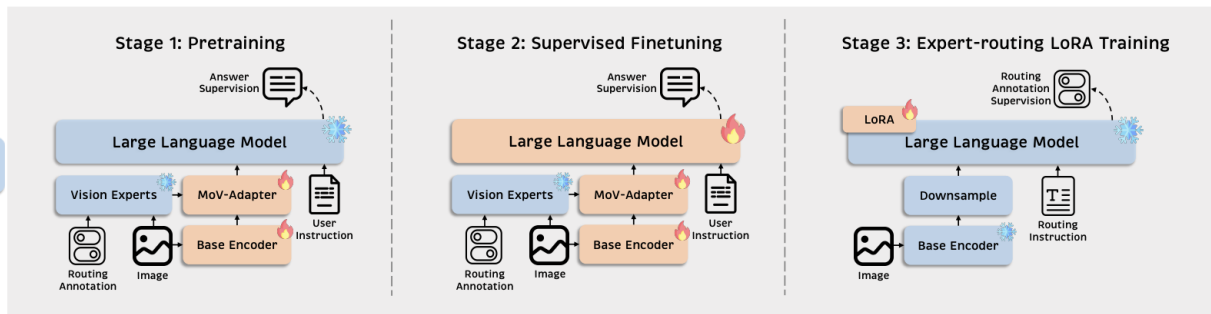
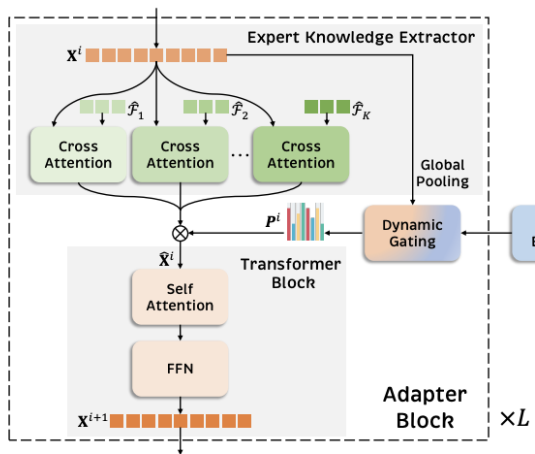
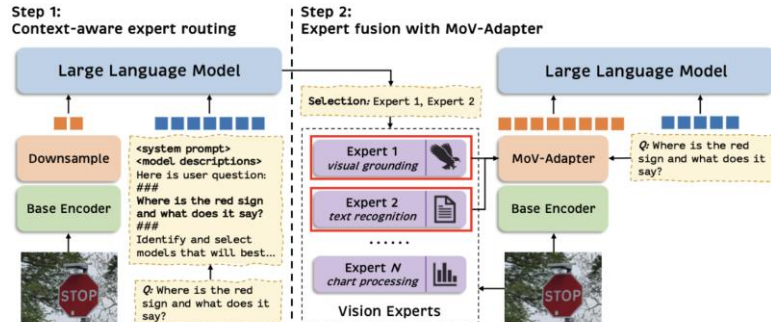
(a) Stage I and Stage II



(b) Stage III

# \* Recent Advanced MLLM Designs

Vision Encoder	Task	MMB	DocVQA	ChartQA	GQA	POPE	REC	RES	SLAKE
CLIP [60]	Image-text Contrastive	<b>64.9</b>	35.6	35.3	62.5	85.7	81.5	43.3	63.7
DINOv2 [57]	Visual Grounding	57.5	14.7	15.9	<b>63.9</b>	86.7	<b>86.1</b>	47.5	59.4
Co-DETR [86]	Object Detection	48.4	14.2	14.8	58.6	<b>88.0</b>	82.1	48.6	55.3
SAM [30]	Image Segmentation	40.7	13.9	15.0	54.0	82.0	79.2	<b>49.3</b>	57.7
Pix2Struct [35]	Text Recognition	41.9	<b>57.3</b>	<b>53.4</b>	51.0	78.1	59.2	32.2	44.0
Deplot [43]	Chart Understanding	36.2	40.2	<b>55.8</b>	48.1	75.6	51.1	27.0	44.5
Vary [75]	Document Chart Parsing	28.1	<b>47.8</b>	<b>41.8</b>	42.6	69.1	21.6	16.0	40.9
BiomedCLIP [84]	Biomedical Contrastive	40.0	15.3	16.8	50.8	76.9	57.8	27.4	<b>65.1</b>
Plain fusion	-	63.4	46.5	48.9	63.0	86.4	85.7	45.3	64.7
MoVA	-	<b>65.9</b>	<b>59.0</b>	<b>56.8</b>	<b>64.1</b>	<b>88.5</b>	<b>86.4</b>	<b>49.8</b>	<b>66.3</b>





# \* MLLM Functionality & Advance

---

## + Fine-Grained MLLM Design

- × With Visual Grounding.
- × With Visual Segmentation.
- × Video and 3D Fine-Grained MLLM.

Fine-Grained Understanding.

## + Advanced MLLM Design

- × Unified Architecture Designs.
- × MLLM For Long Video Analysis.
- × MLLM With MOE Design.

Stronger Features and Capacities.

# \* MLLM Functionality & Advance

---

## **Future Direction:**

- 1, Scaling MLLM features More.
- 2, Novel MoE operators designed for MLLMs.
- 3, Single Transformer Architecture. Eg: unify image generation and text generation in one model.
- 4, Long Video Grounding, Chat and Tracking in One Model.

# Thanks!

