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

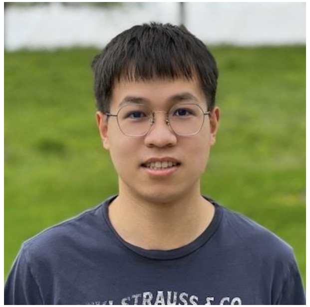
Melbourne, Australia

Hao Fei

National University of Singapore

Xiangtai Li

ByteDance/Tiktok

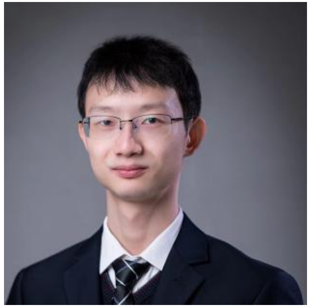


Haotian Liu

xAI

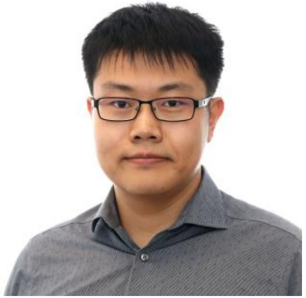
Fuxiao Liu

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

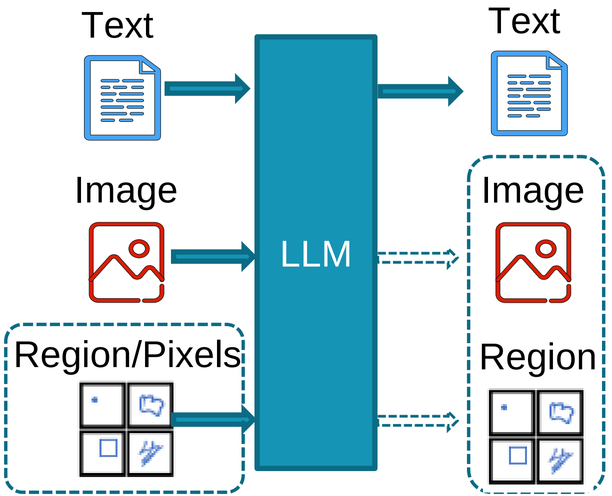
Overview and Concepts

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.

Overview of Fine-grained Models Before LLM

Visual Grounding

Referring Segmentation

Visual Segmentation

Video/3D Referring Segmentation

Visual Prompting.

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

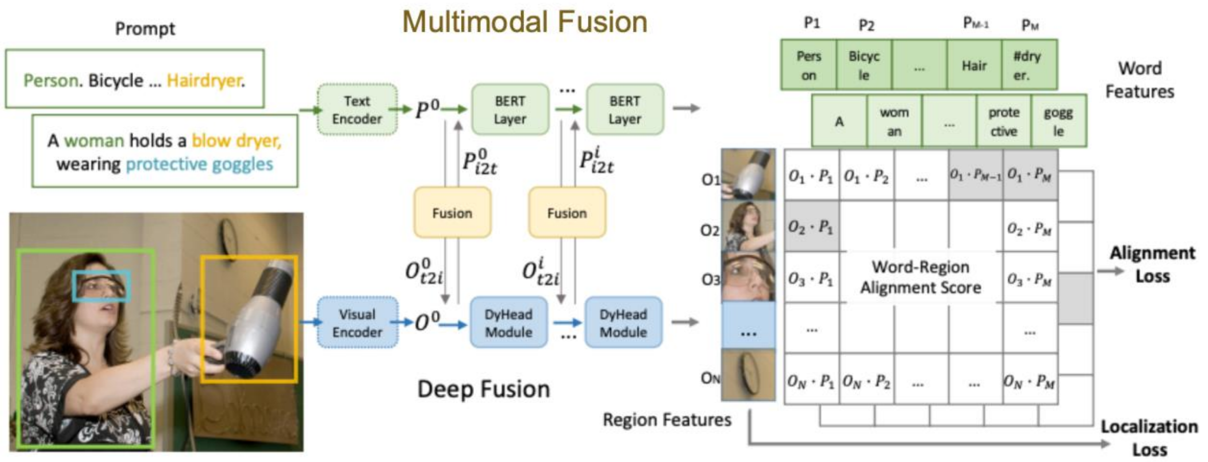
- Most works come from **vision** community.

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

Fine-grained Capability of MLLM

Overview of Fine-grained Models Before LLM

Dual Encoders

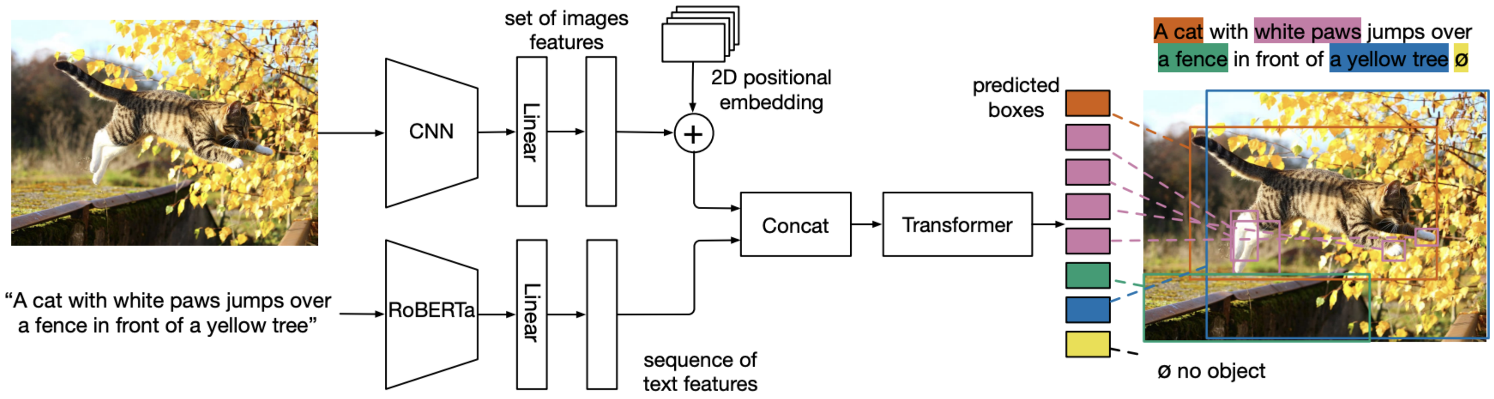


Contrastive Image-Text Alignment

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

Fine-grained Capability of MLLM

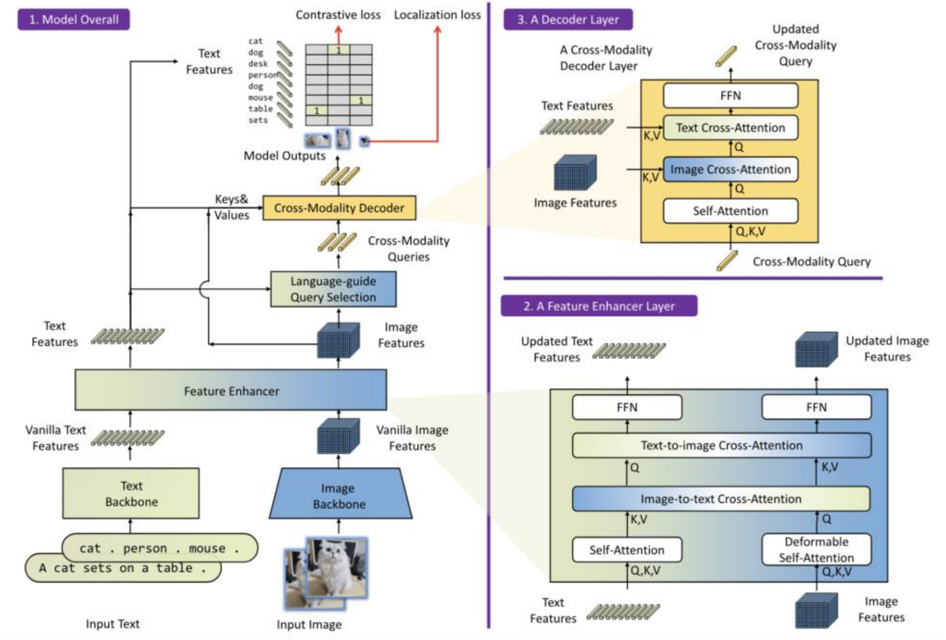
Overview of Fine-grained Models Before LLM

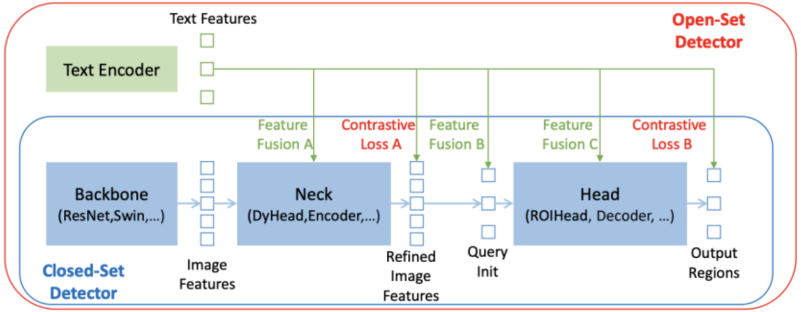


MDETR - Modulated Detection for End-to-End Multi-Modal Understanding-2021

Fine-grained Capability of MLLM

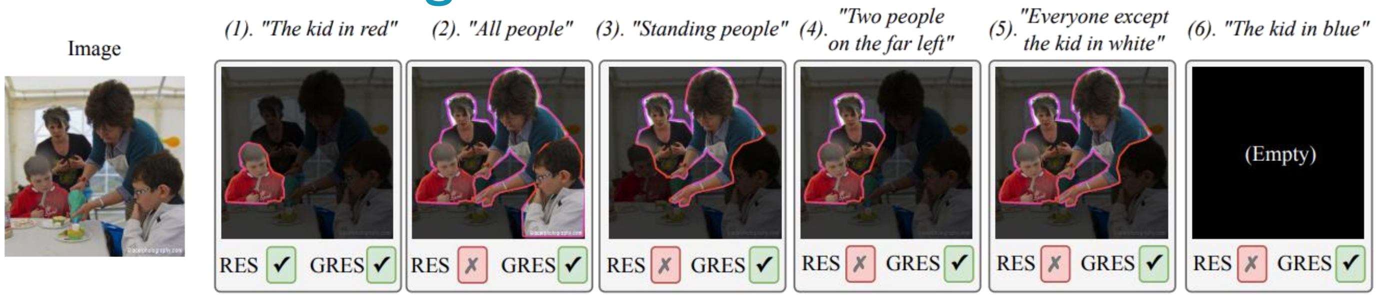
Overview of Fine-grained Models Before LLM

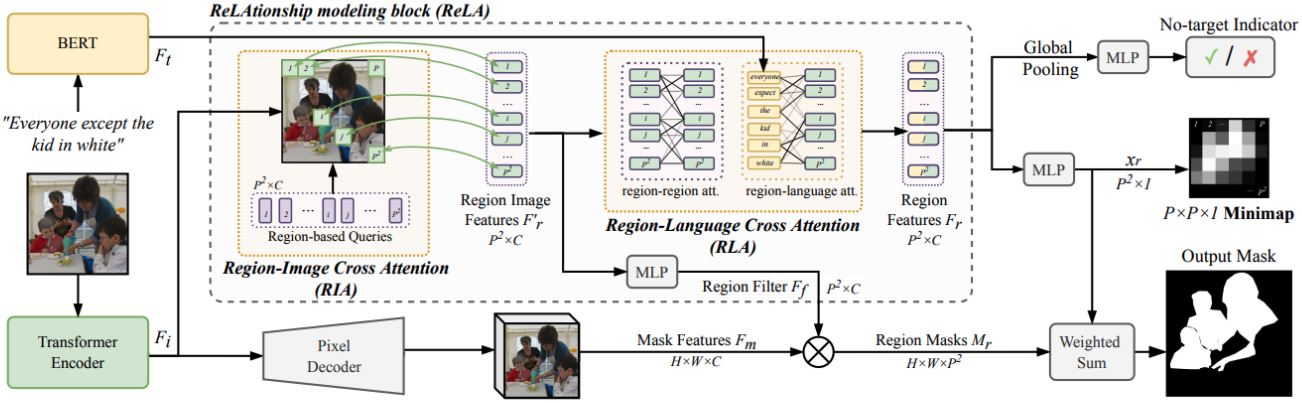




Fine-grained Capability of MLLM

Overview of Fine-grained Models Before LLM





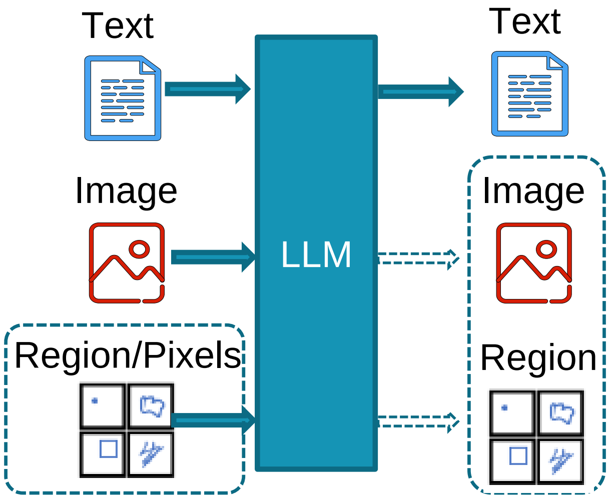
GRES: Generalized Referring Expression Segmentation, arxiv-2023.

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

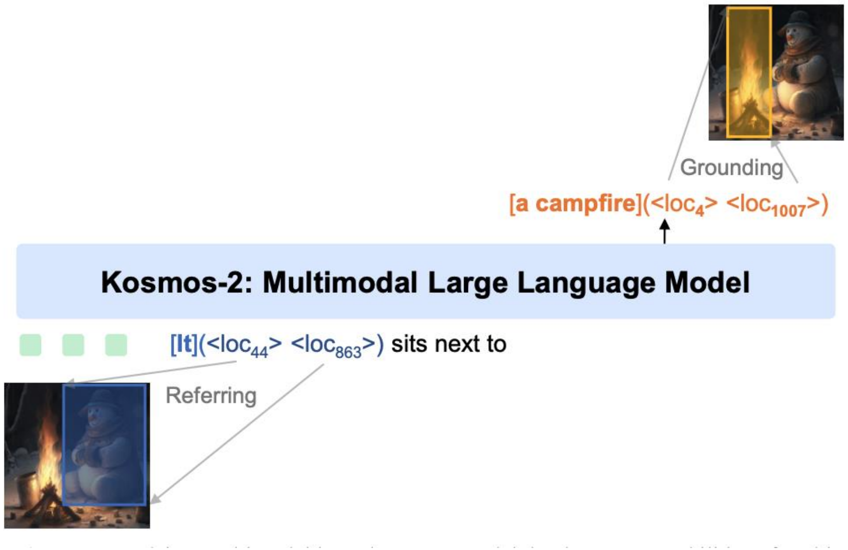
Fine-Grained MLLM Design

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

Fine-grained Capability of MLLM

With Visual Grounding.

Kosmos-2:



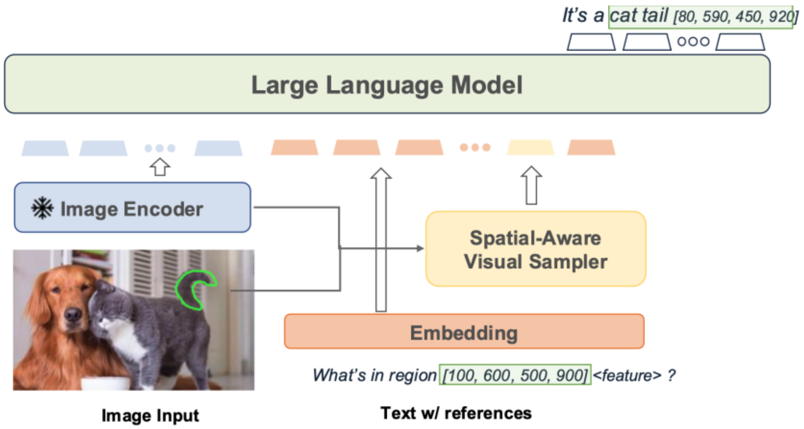
Shikra:

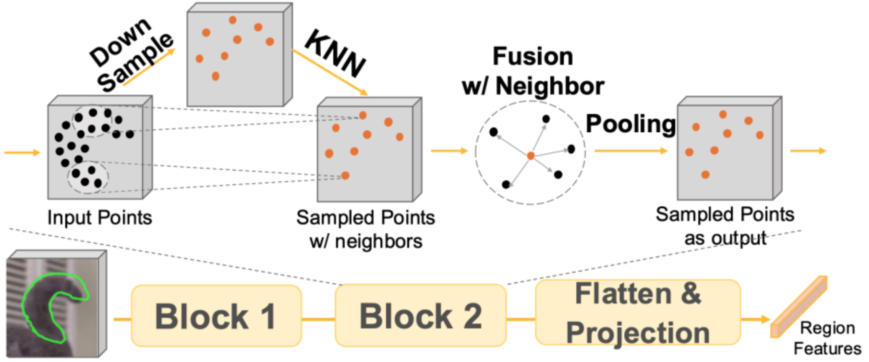
Kosmos-2: Grounding Multimodal Large Language Models to the World
Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic

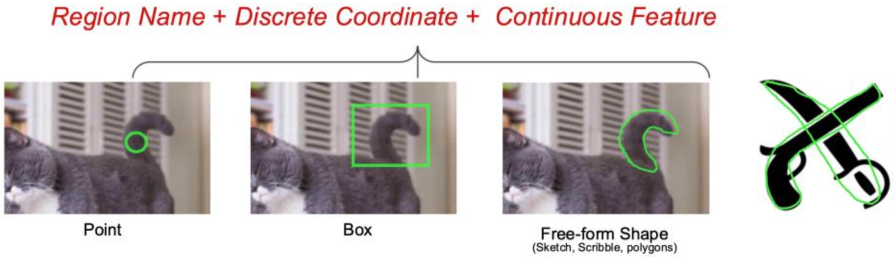
Fine-grained Capability of MLLM

With Visual Grounding.

Ferret





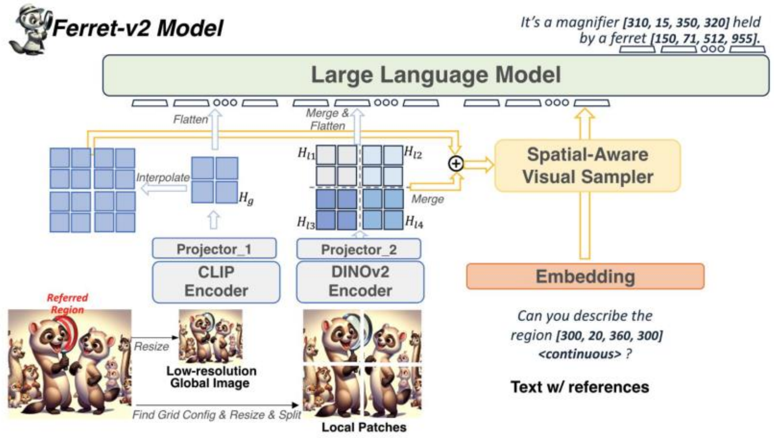


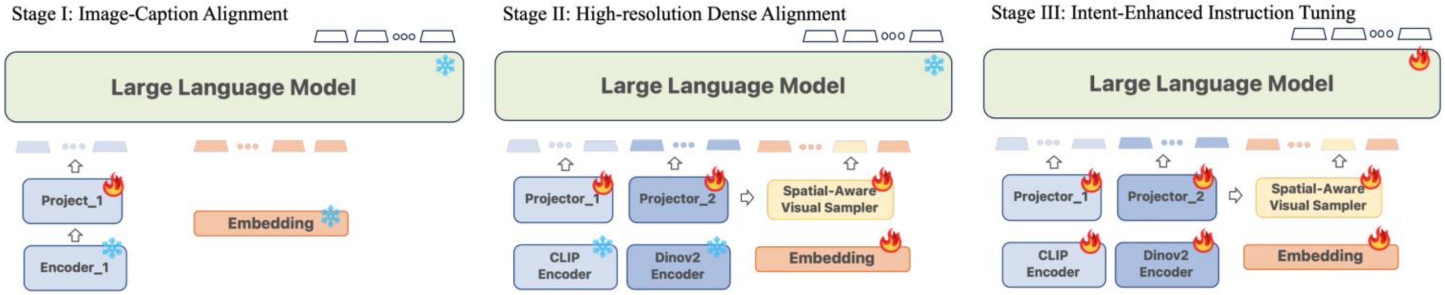
Ferret: Refer and Ground Anything Anywhere at Any Granularity, arxiv-2023.

Fine-grained Capability of MLLM

With Visual Grounding.

Ferret-v2



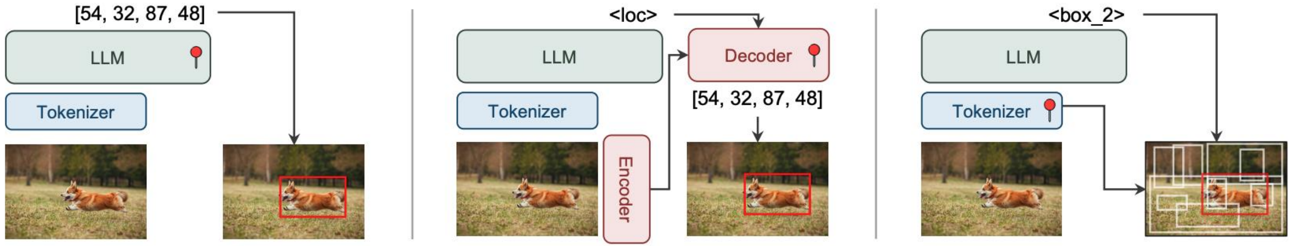


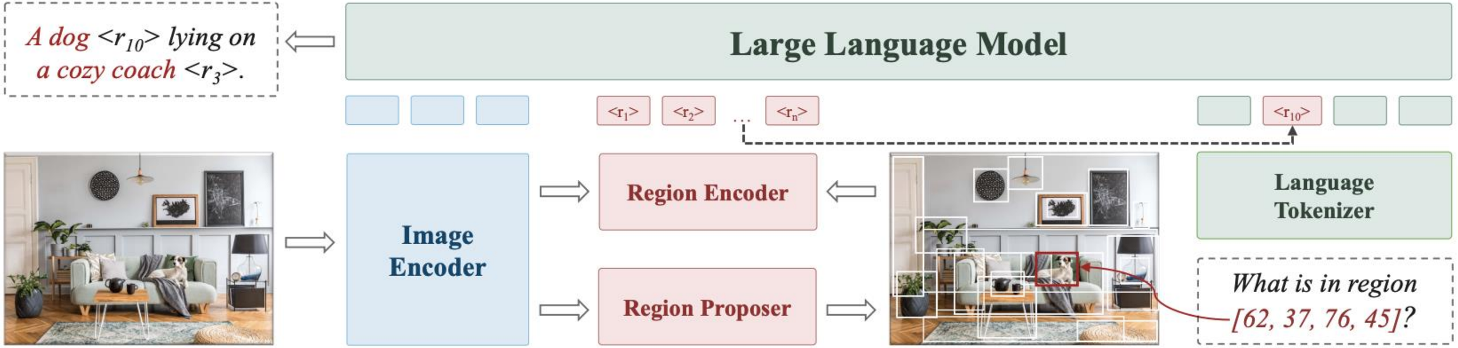
Ferret-v2: An Improved Baseline for Referring and Grounding with Large Language Models, arxiv-2024

Fine-grained Capability of MLLM

With Visual Grounding.

Groma:

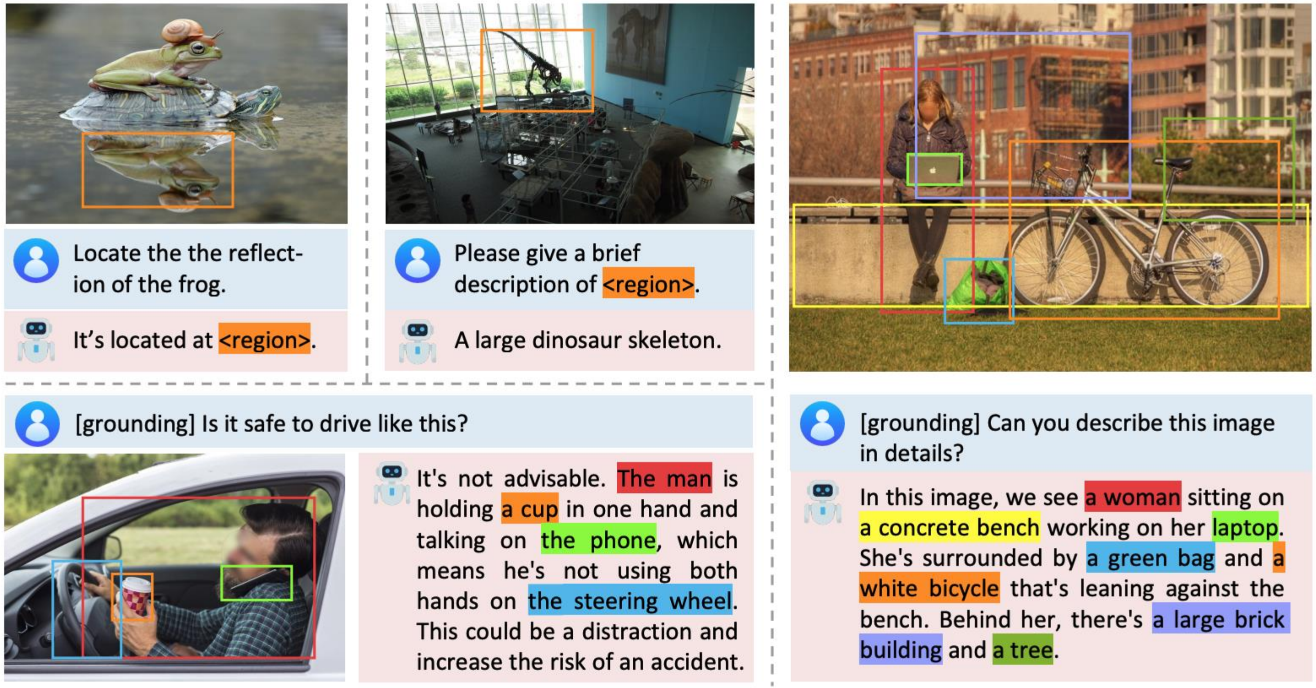




Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models, arxiv-2024

Fine-grained Capability of MLLM

With Visual Grounding.



Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models, arxiv-2024

Fine-grained Capability of MLLM

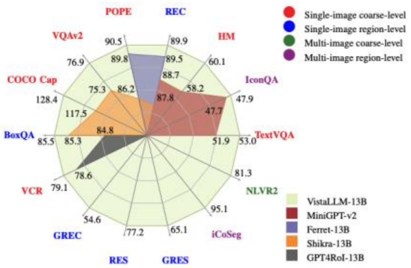
Fine-Grained MLLM Design

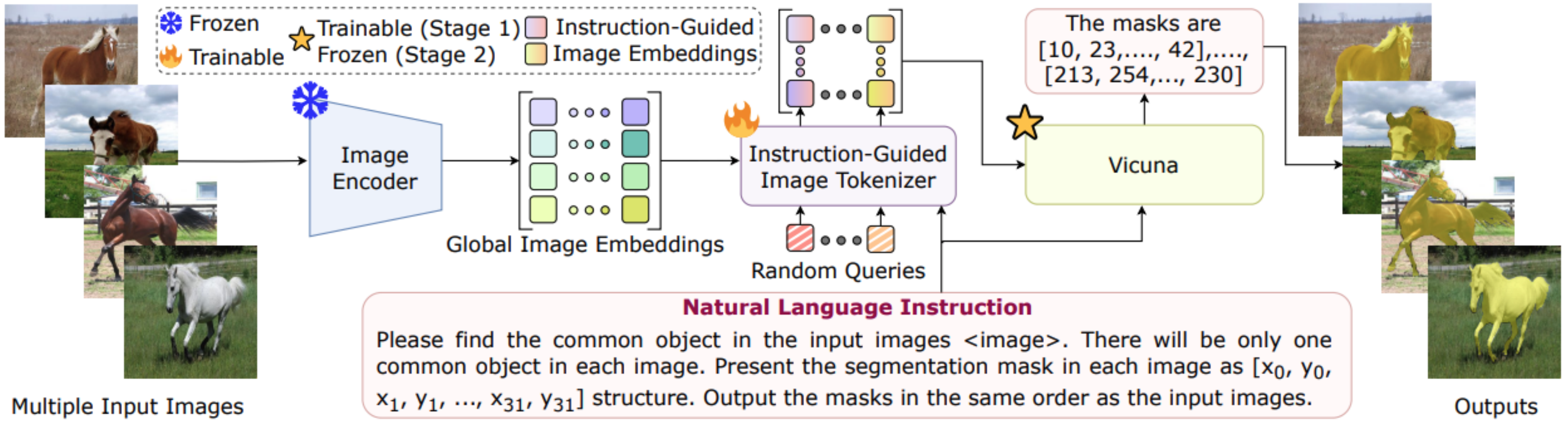
- Overview
- With Visual Grounding.
- **With Visual Segmentation.**
- Video and 3D Fine-Grained MLLM.

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, arixv-2023.

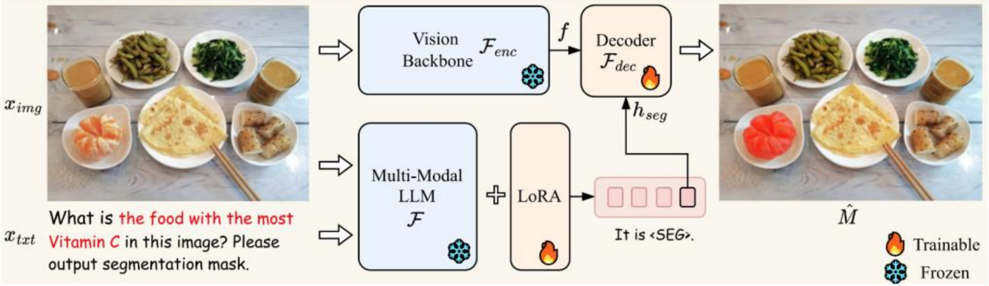
Fine-grained Capability of MLLM

With Visual Segmentation.

LISA: Large Language Instructed Segmentation Assistant

Paper | Models | Training | Inference | Local Deployment | Dataset | Online Demo

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



- **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>.

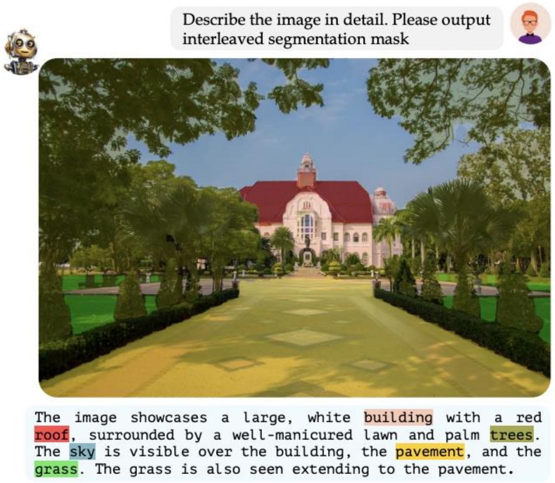
LISA: Reasoning Segmentation via Large Language Model, arxiv-2023.

Fine-grained Capability of MLLM

With Visual Segmentation.

Fine-grained Capability of MLLM

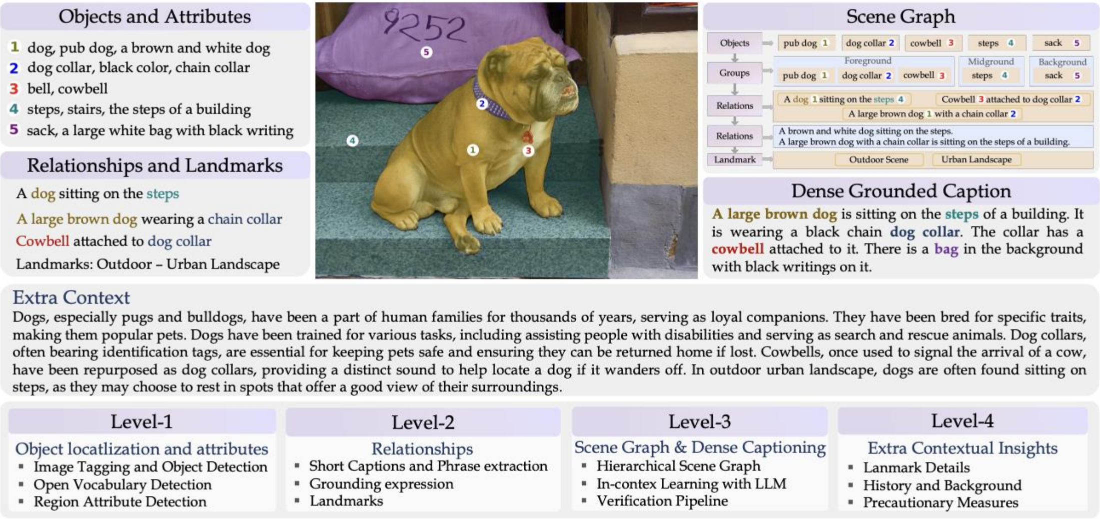
With Visual Segmentation.



Method	Image	Input / Output: Region	Input / Output: Multi-Region	Region Enc. / Dec.	Pixel-Wise Grounding	Multi-turn Conversation	End-End Model
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.



Model	Validation Set M	C	AP50	mIoU	Recall	Test Set 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**

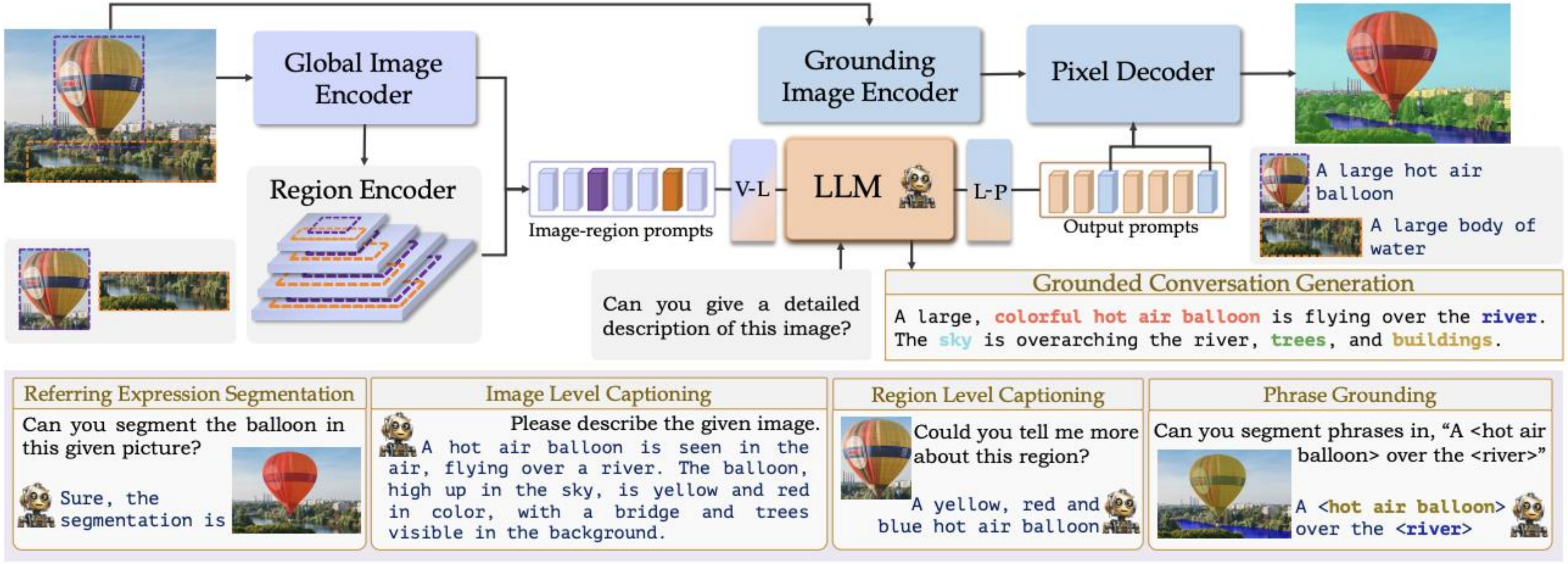
Table 3. **Performance on GCG Task**: Metrics include METEOR (M), CIDEr (C), AP50, mIoU, and Mask Recall. LISA* denotes LISA adapted for GCG. GLaMM† denotes training excluding 1K human annotated images. GLaMM shows better performance.

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.



GLaMM: Pixel Grounding Large Multimodal Model, arxiv-2023

Fine-grained Capability of MLLM

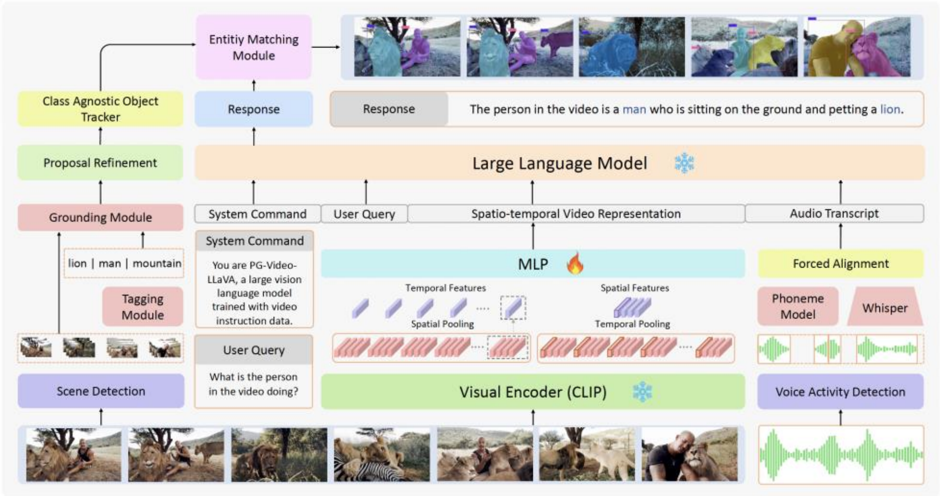
Fine-Grained MLLM Design

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

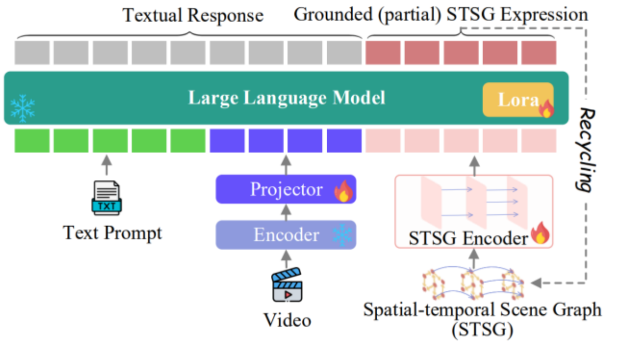
Fine-grained Capability of MLLM

Video and 3D Fine-Grained MLLM.

PG-Video-LLaVA



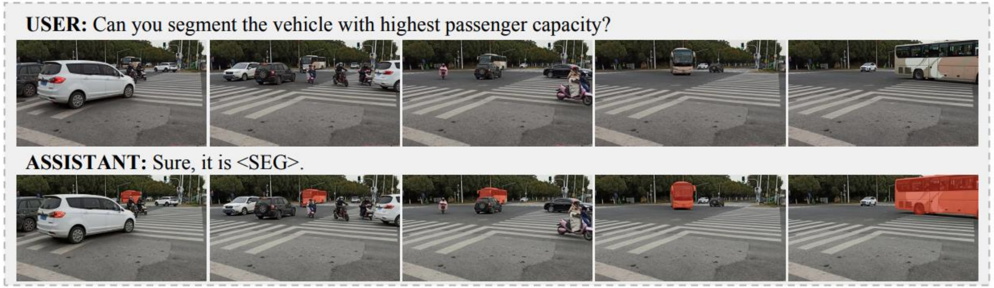
MotionEpic

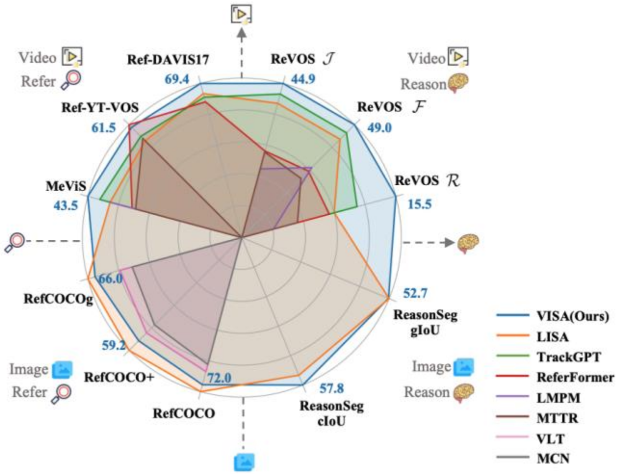


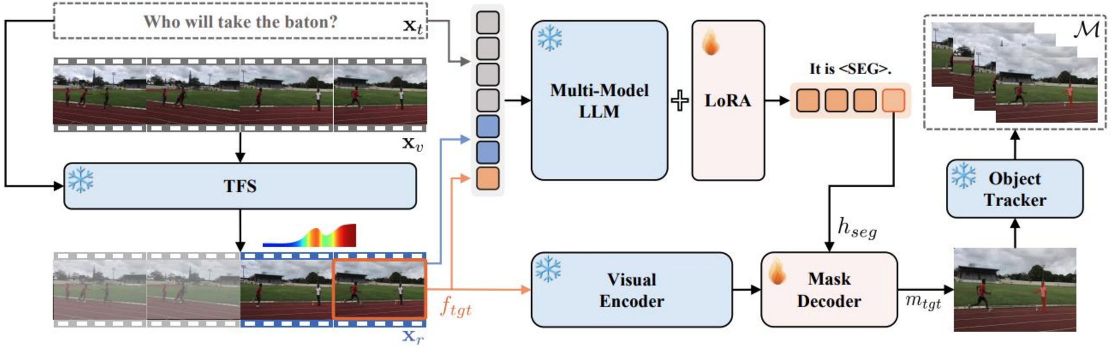
[1] PG-Video-LLaVA: Pixel Grounding in Large Multimodal Video Models. Arxiv-2023
[2] Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. Axiv-2024

Fine-grained Capability of MLLM

Video and 3D Fine-Grained MLLM.





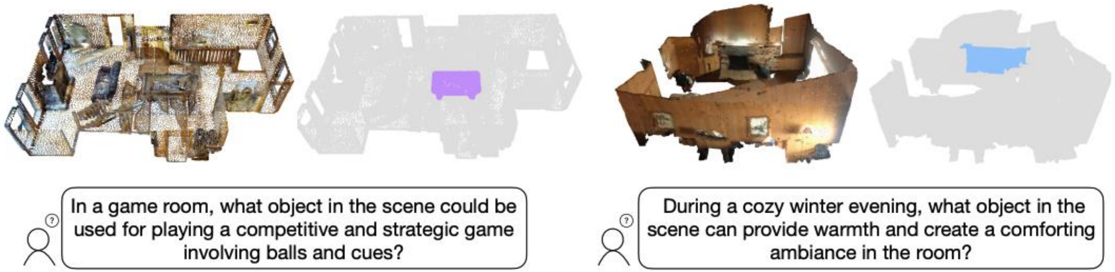


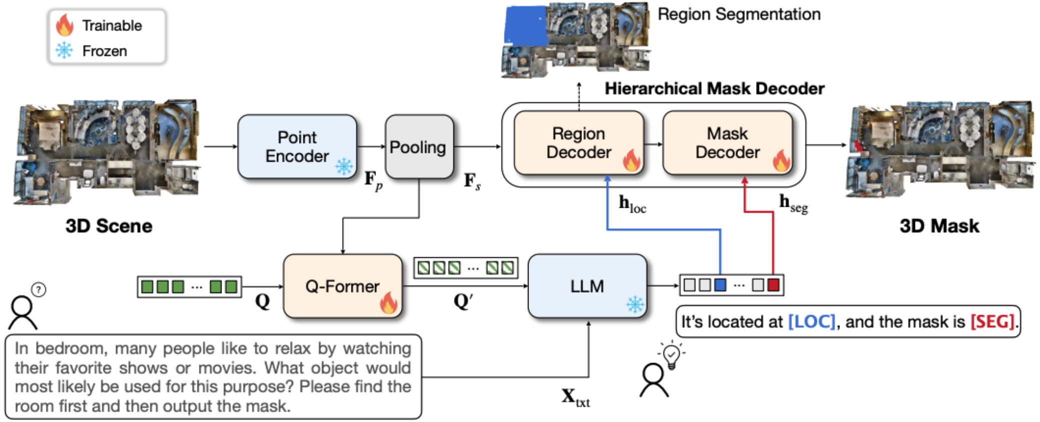
VISA: Reasoning Video Object Segmentation via Large Language Models, arxiv-2024

Fine-grained Capability of MLLM

Video and 3D Fine-Grained MLLM.

Reason3D

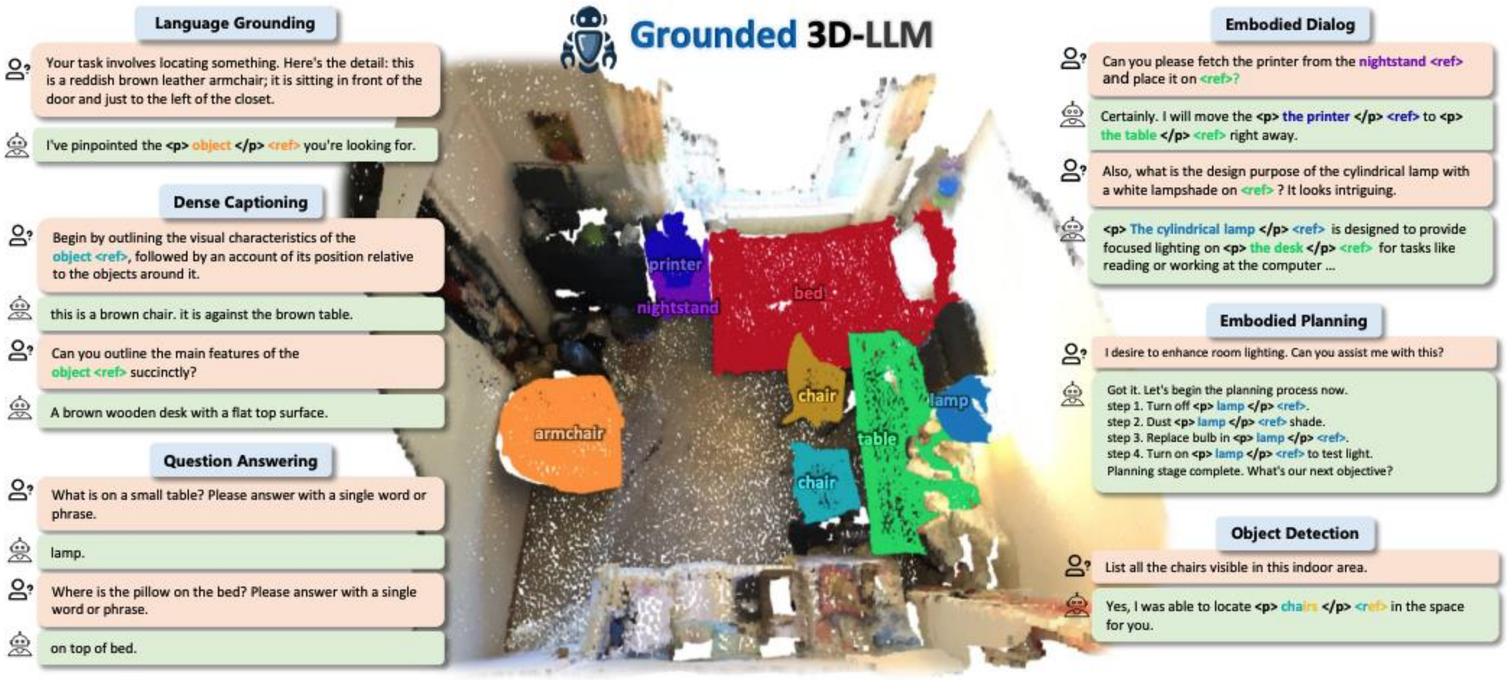




Reason3D: Searching and Reasoning 3D Segmentation via Large Language Model, arxiv-2024

Fine-grained Capability of MLLM

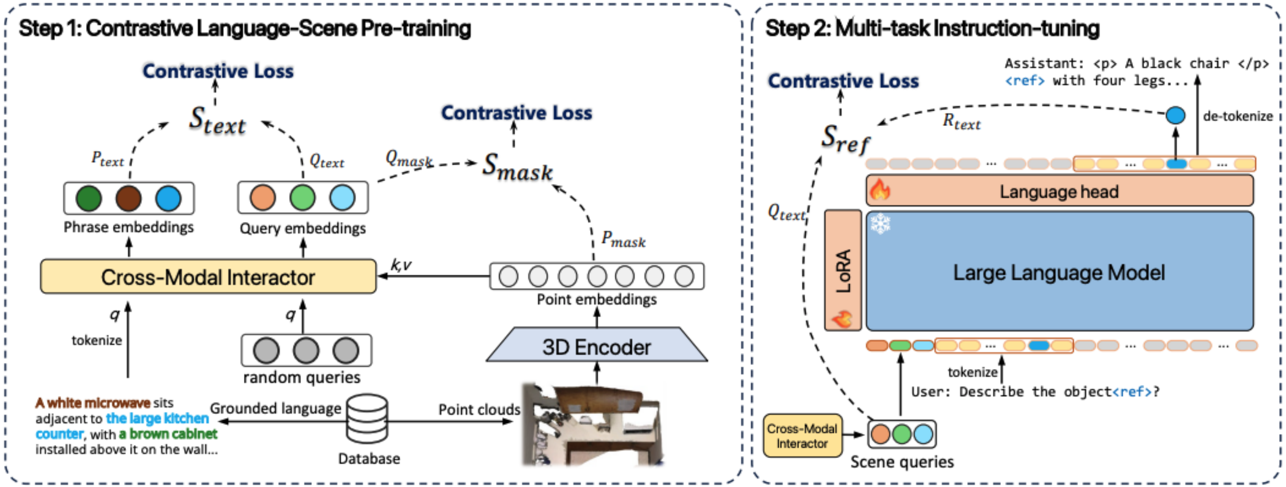
Video and 3D Fine-Grained MLLM.



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	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



Grounded 3D-LLM with Referent Tokens, arxiv-2024

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

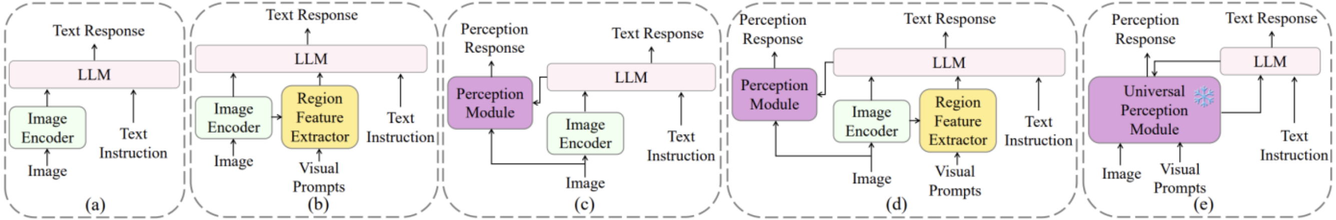
Advanced MLLM Design

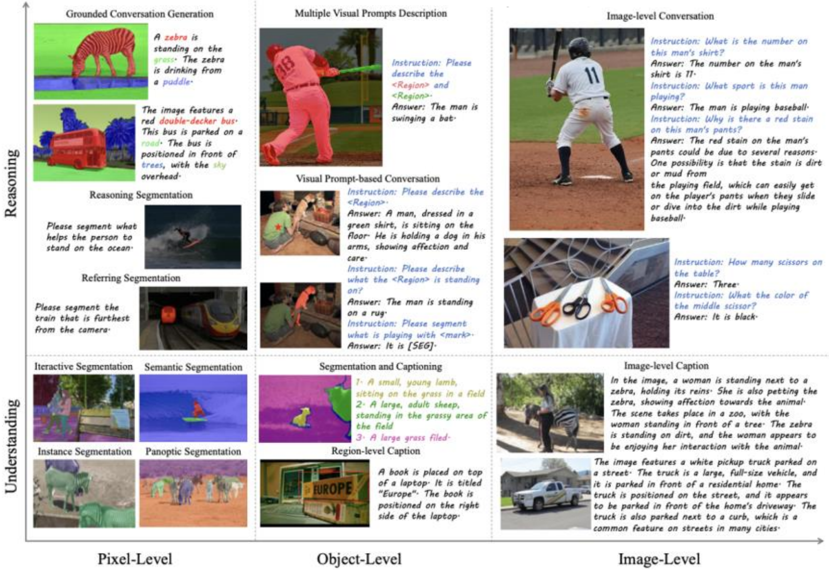
- Overview
- Unified Architecture Designs.
- MLLM For Long Video Analysis.
- MLLM With MOE Design.

Recent Advanced MLLM Designs

Unified Architecture

OMG-LLaVA



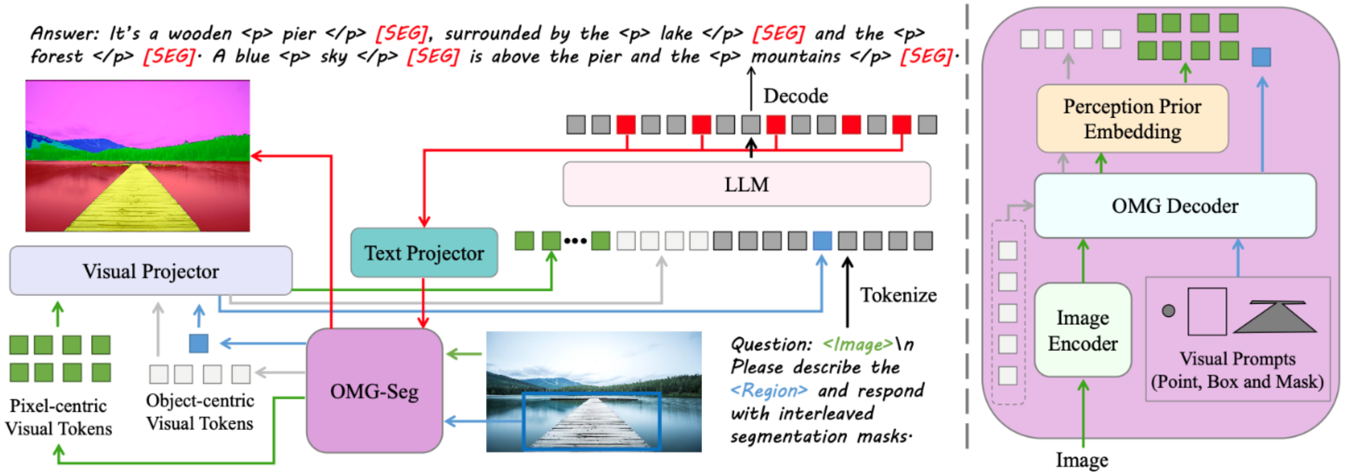


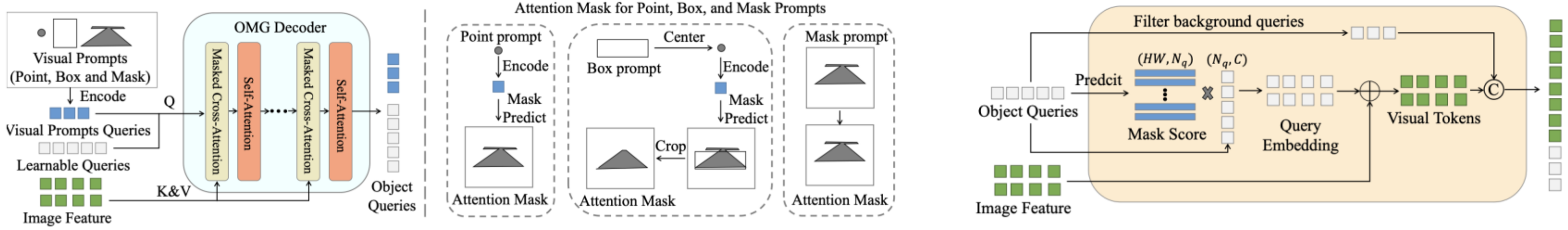
Method	Visual Encoder	Image-level Caption	Image-level Conversation	Object-level Visual Prompts	Object-level Caption	Object-level Conversation	Pixel-level Universal Seg	Pixel-level RES	Pixel-level GCG
LLAVA [69]	1	✓	✓						
MiniGPT4 [140]	1	✓	✓						
mPLUG-Owl [116]	1	✓	✓						
LLaMA-Adapter [130]	1	✓	✓						
Mini-Gemini [63]	2	✓	✓						
InternVL 1.5 [18]	1	✓	✓						
VisionLLM [95]	1	✓	✓					✓	
Shikra [13]	1	✓	✓	Point & Box	✓	✓			
Kosmos-2 [80]	1	✓	✓	Box	✓	✓			
GPT4RoI [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	✓	✓	Box	✓	✓			
VIP-LLaVA [8]	1	✓	✓	Point & Box & Mask	✓	✓			
PSALM [133]	1			Point & Box & Mask			✓	✓	
LaSagnA [100]	2							✓	
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

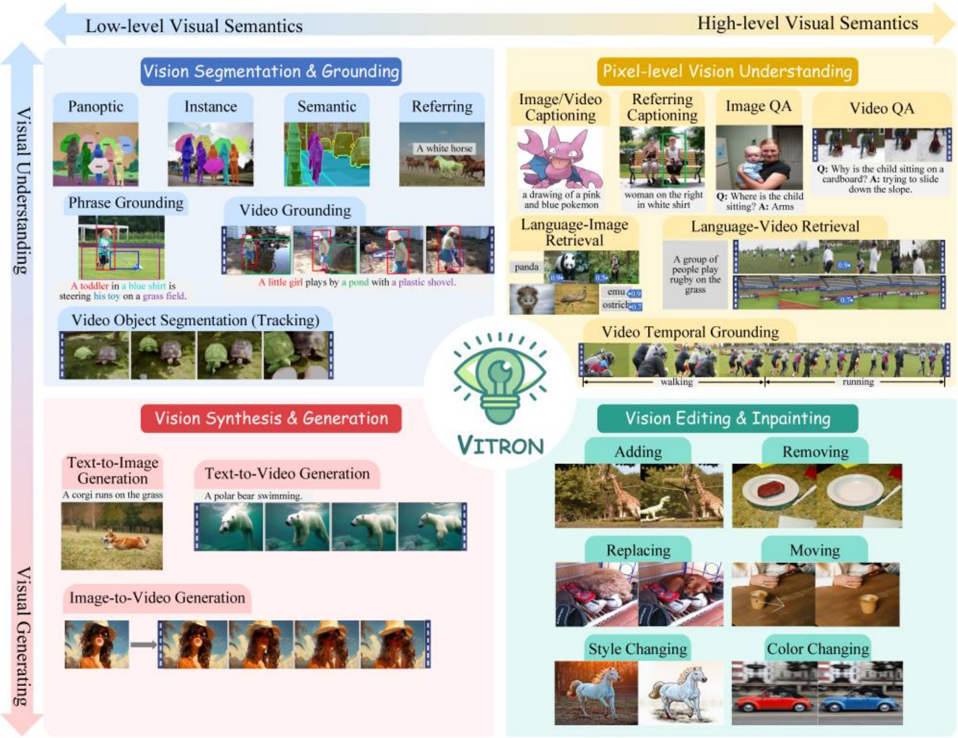




OMG-LLaVA : Bridging Image-level, Object-level, Pixel-level Reasoning and Understanding, arxiv-2024.

Recent Advanced MLLM Designs

- **Unified Pixel-wise MLLM**
  - Vitron



VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

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)	✓	✓	✓	✓	✓	✓

VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

Recent Advanced MLLM Designs

- ## Unified Pixel-wise MLLM

Vitron

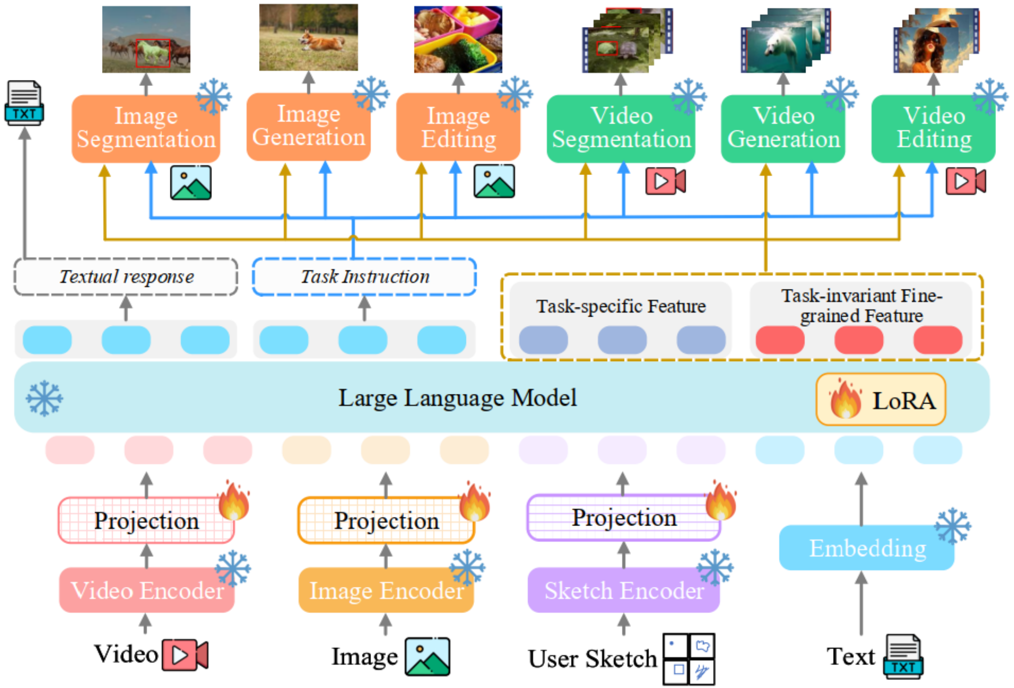


Figure 2: Technical overview of the VITRON framework.

VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

Image Segmentation

Video Segmentation

Video Understanding

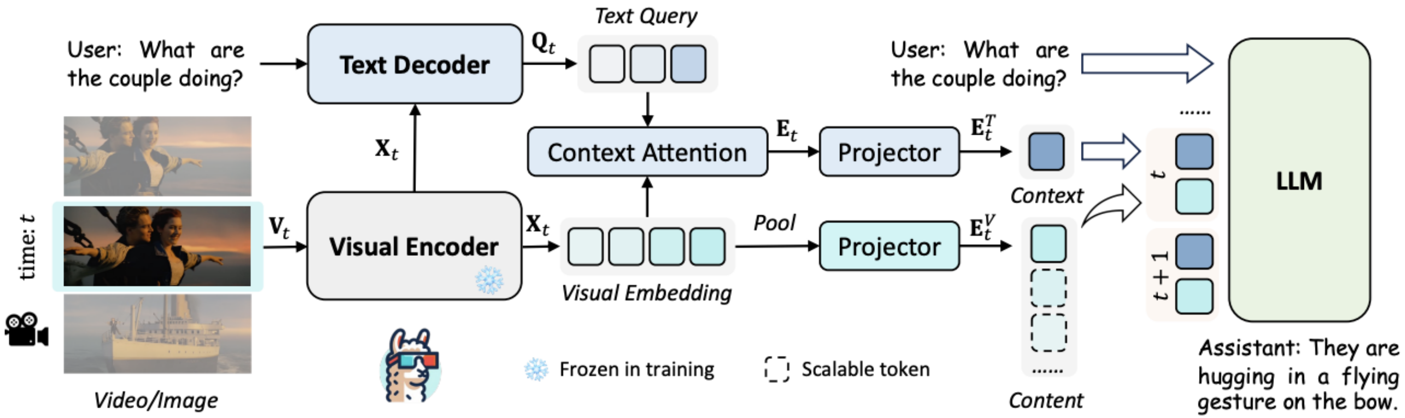
Video Editing

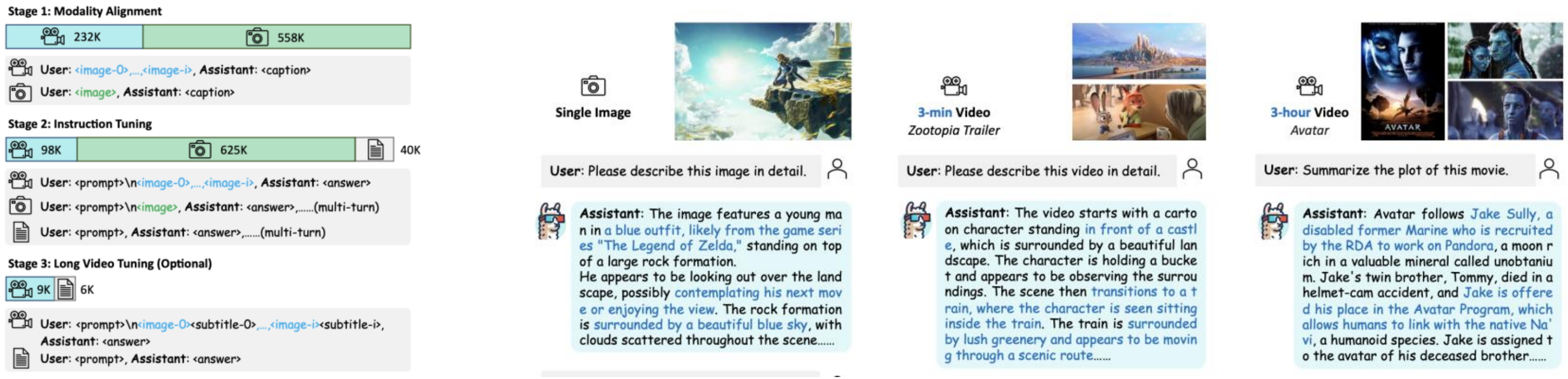
Recent Advanced MLLM Designs

Advanced MLLM Design

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

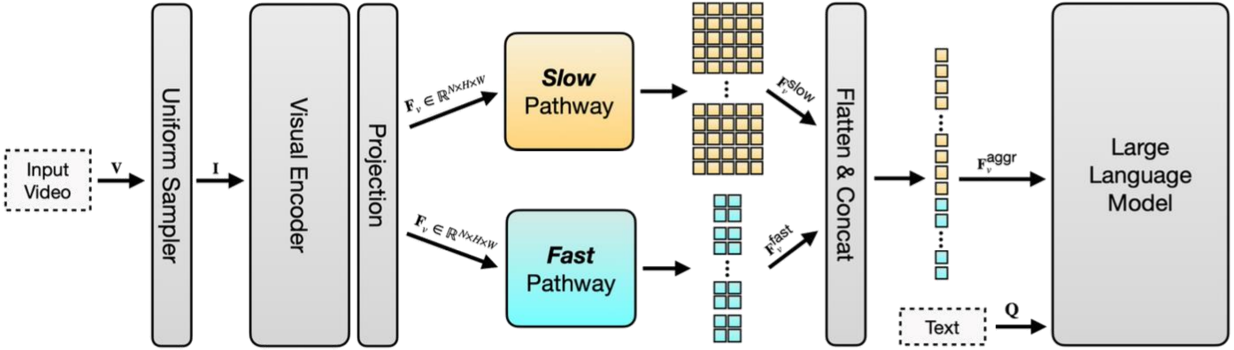
Recent Advanced MLLM Designs

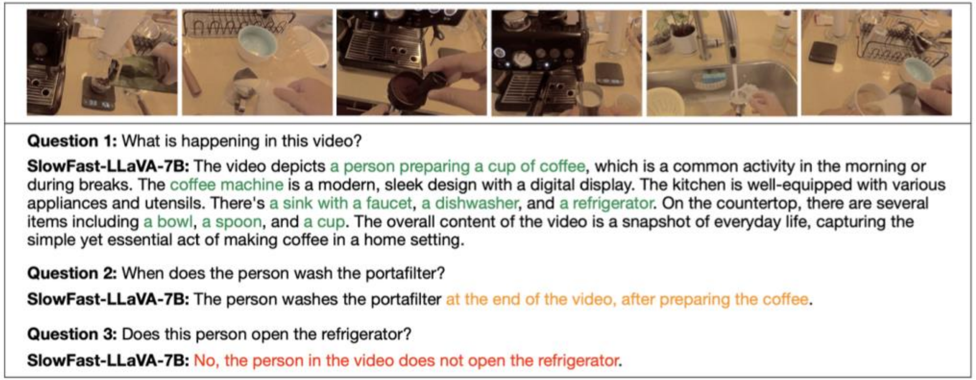


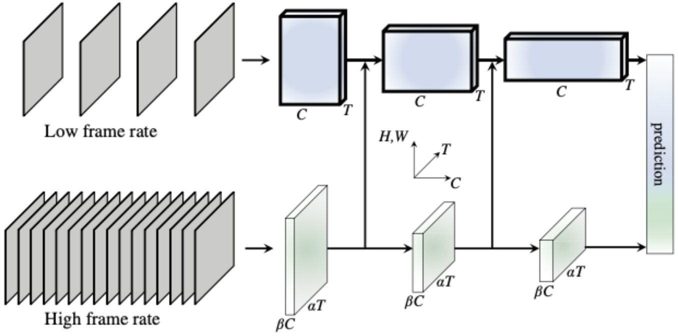


LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models, arxiv-2023

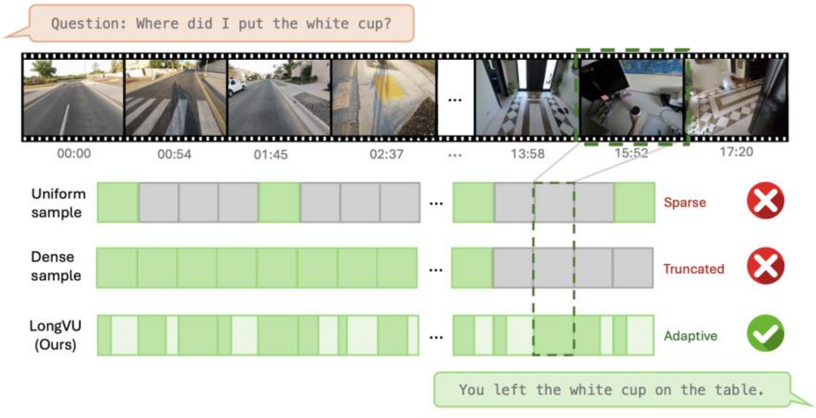
Recent Advanced MLLM Designs

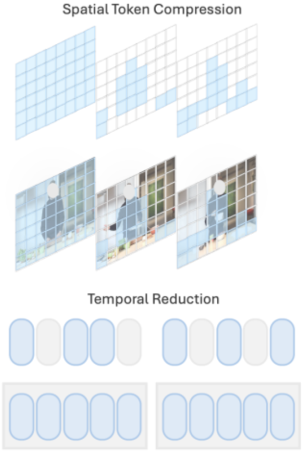


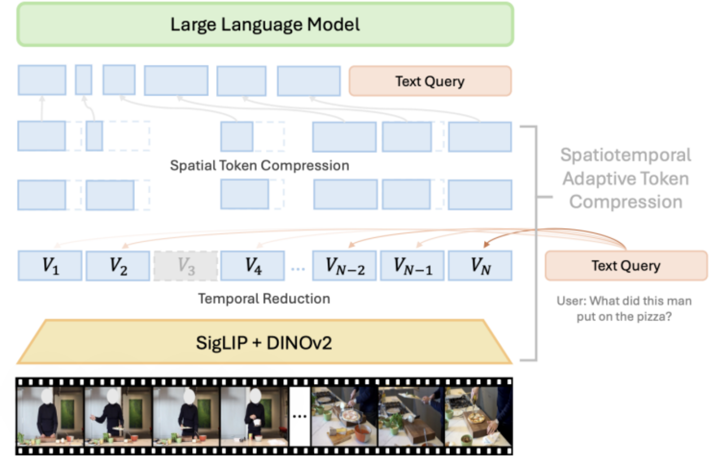




Recent Advanced MLLM Designs







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

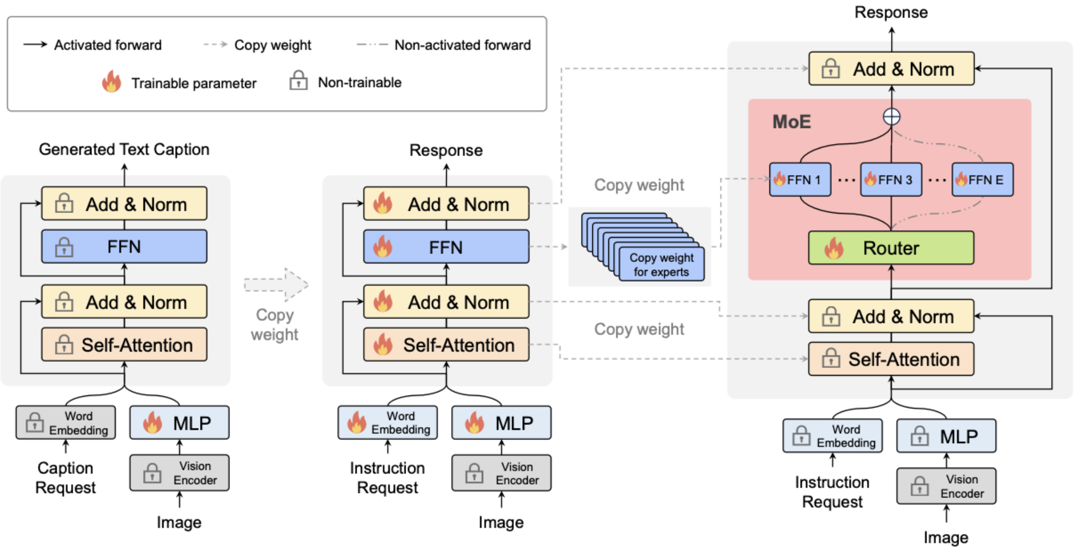
LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. Arixv-2024

Recent Advanced MLLM Designs

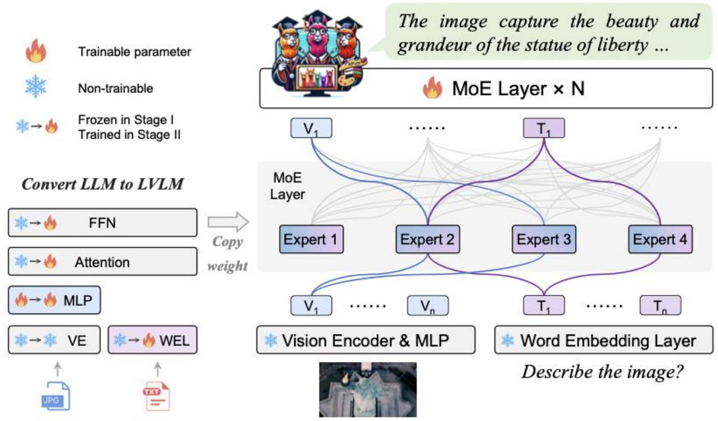
Advanced MLLM Design

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

Recent Advanced MLLM Designs



(a) Stage I (b) Stage II (c) Stage III

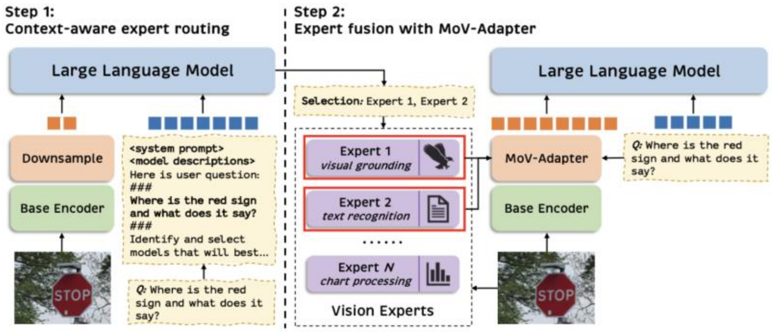


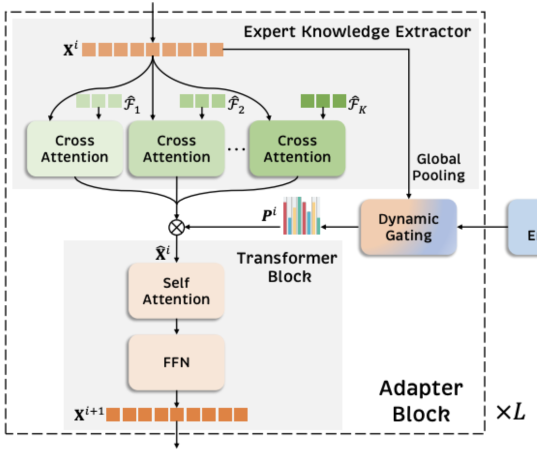
(a) Stage I and Stage II (b) Stage III

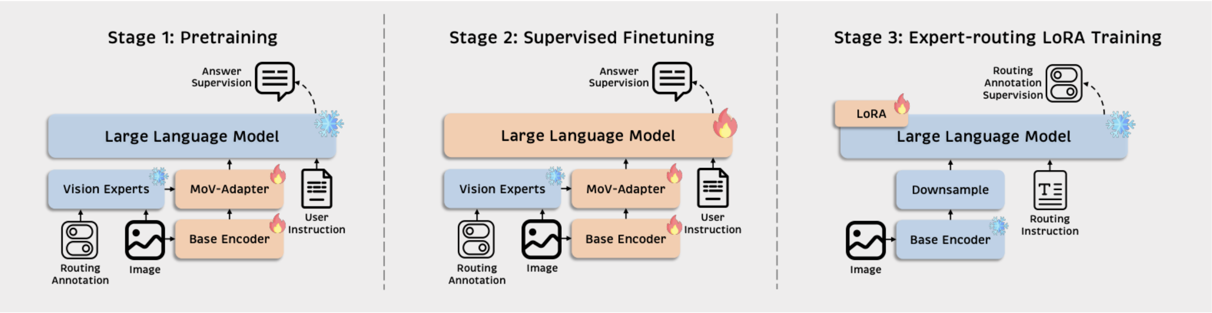
MoE-LLaVA: Mixture of Experts for Large Vision-Language Models, arxiv-2024

Recent Advanced MLLM Designs

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







MoVA: Adapting Mixture of Vision Experts to Multimodal Context, arxiv-2024.

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!