

# Fine-grained Pixel-level Understanding with VLMs

Wentong Li

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

2025-10-15

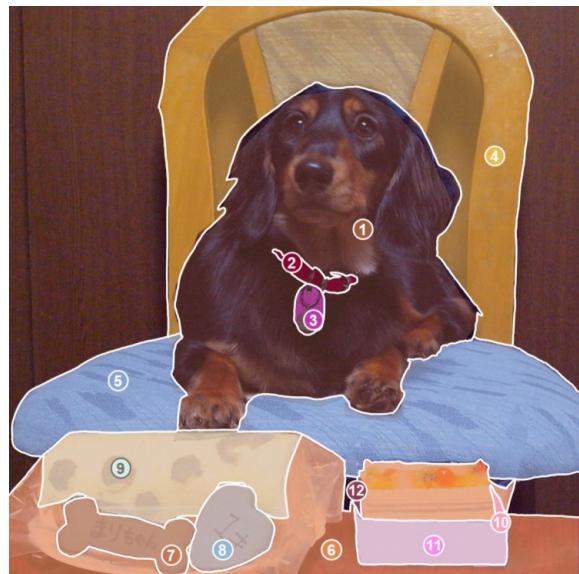
# Task Definitions

- **Fine-grained Spatial Understanding** can be reflected in two types of tasks:

## 1. Referring

Input image + text instruction + Region

Model is required to **understand** the referred regions and respond to the instruction.



### *Text prompt*

**Question:** Can you describe the pillow that the dog is resting on in [0,300,500,510]?

**Answer:** The pillow is a blue cushion or pillow that is part of the chair...

### *Fine-grained visual prompt*

**Question:** Can you describe the pillow that the dog is resting on in <region5>?

**Answer:** The pillow in region5 is a blue cushion or pillow that is part of the chair...

# Task Definitions

- **Fine-grained Spatial Understanding** can be reflected in two types of tasks:

## 2. Grounding

Output: text response + **Region**

Model is required to **localize** the objects in image when mentioning them in response.



### **Text output**

**Question:** Who was the president of the US in this image? Please output its box.

**Answer:** The president of the US is [600,150,800,500].

### **Fine-grained visual output**

**Question:** Who was the president of the US in this image? Please output segmentation mask.

**Answer:** Sure, the segmentation result is [SEG].



# Fine-grained Image-level Region Understanding

## Osprey



SAM "Segment Everything" Predictions

**No semantic information**

- Integrate images, target regions (masks), and textural data;
- Enable fine-grained semantic description of arbitrary regions or objects within images;
- Strong robustness and generalization.

Object Category: person

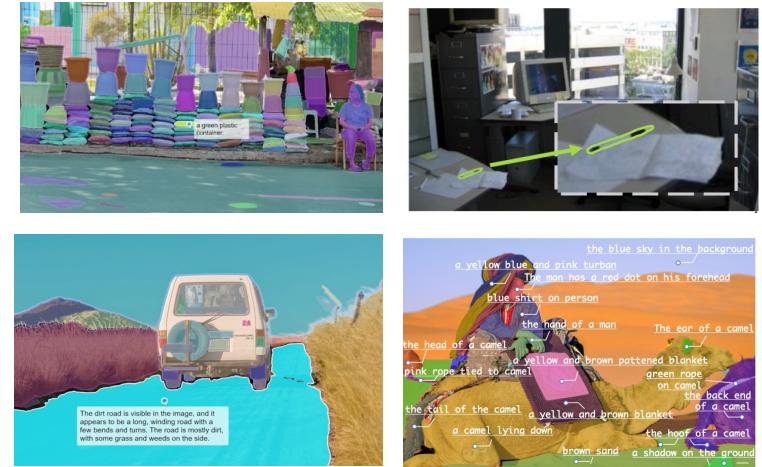
Part Taxonomy: body

Attribute: color, position ...

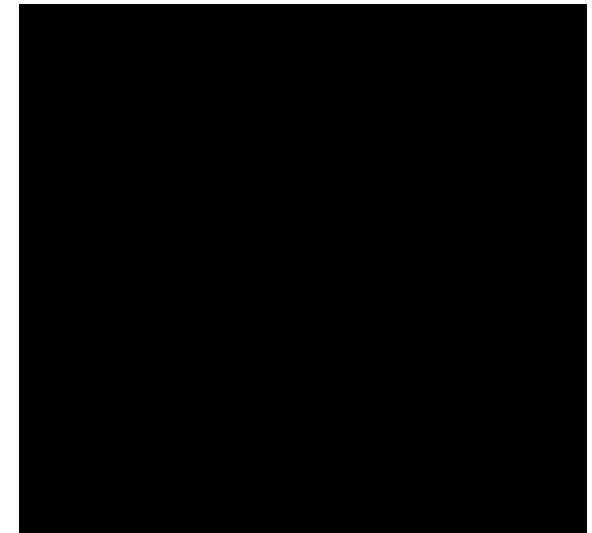
Caption: region short / detailed description

Fine-grained Region/Pixel Understanding

**Rich semantic information containing different granularities**

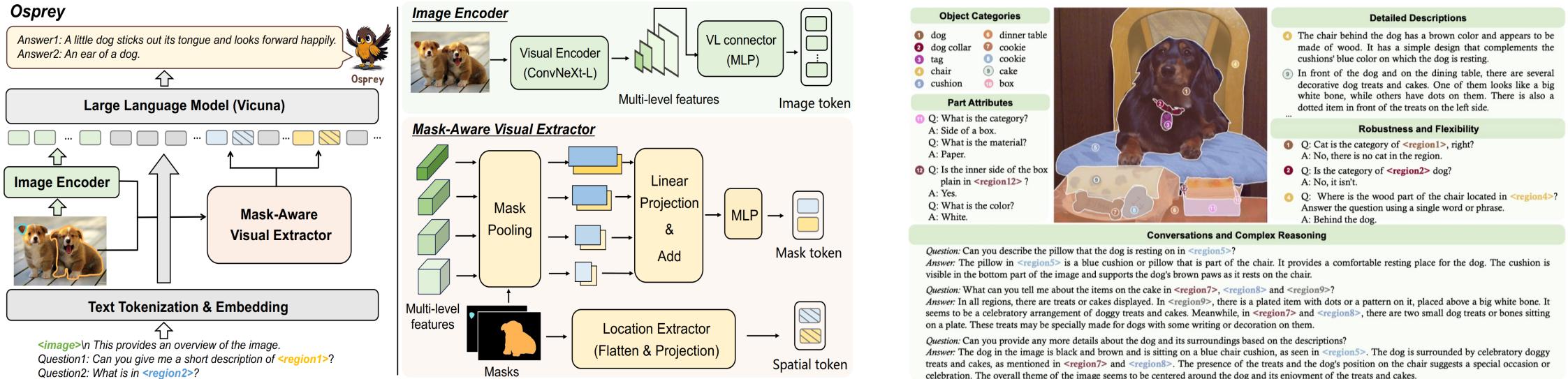


General scene



Out-of-domain Scene

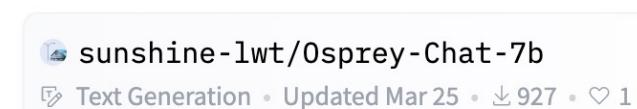
# Fine-grained Image-level Region Understanding



- Support high-resolution image
  - ConvNeXt (512x512@training, 800x800@inference)
- Pixel-level region feature extraction
  - Mask-Aware visual extractor (multi-level)

- 720K region-text pairs.
- Six types of object region-text data.

Open-source: <https://huggingface.co/sunshine-lwt>



# Fine-grained Image-level Region Understanding

## Online Demo

Segmentation

Description(Osprey)  
red light on top of police car.

point-prompt box-prompt segment everything

Input image

Generate segmentation and description

Clear Image

Segmentation

Description(Osprey)



Video Visualization

# Fine-grained Region Understanding for Spatial Cognition

- Spatial RGPT

Accurate Spatial Understanding & Reasoning on Objects in Image.



How wide is Region [1]?

The width of Region [1] is 7.73 feet.

Is Region [1] behind Region [4]?

No, it is in front of Region [4].

Between Region [0] and Region [2], which one has more height?  
Region [2] is taller.

- Relative Relation
- Metric Measurement
- Spatial Reasoning

**A Relative Relation**  
⑧ Can you confirm if ④ has a greater width compared to ⑥?  
In fact, ④ might be narrower than ⑥.  
⑧ Which of these four, ③, ④, ⑤, ⑥ is the tallest?  
⑥ is the tallest.  
⑧ From the viewer's perspective, is the position of ⑥ closer than that of ③?  
No, ③ is closer.

**B Metric Measurement**  
⑧ If you are at ③, where will you find ⑦?  
⑦ is at 1 o'clock from ③.  
⑧ What is the height of ④?  
④ is 1.38 meters tall.  
⑧ How far apart are ② and ③ horizontally?  
They are 15.33 meters apart horizontally.



**C Spatial Reasoning**  
⑧ Can you try to estimate how many floors are in ①? Think step by step.  
The height of ① is 204.54 feet. Assuming each floor is about 10 feet high, the total number of floors would be 20.454. Since you can't have a fraction of a floor, the total number of floors would be approximately 20.

⑧ I'm planning to build a bridge from ② to ①, is it possible?  
The distance between ② and ① is 14.82 meters, so it is possible to build a bridge between them.

⑧ If you are riding a motorcycle with 36 inches wide, do you think you can pass through the area between ③ and ④?  
The distance between ③ and ④ is 38.95 inches, so yes, you can pass through the area between ③ and ④ since the motorcycle is narrower than the distance between them.

# Fine-grained Region Understanding for Spatial Cognition

- Framework

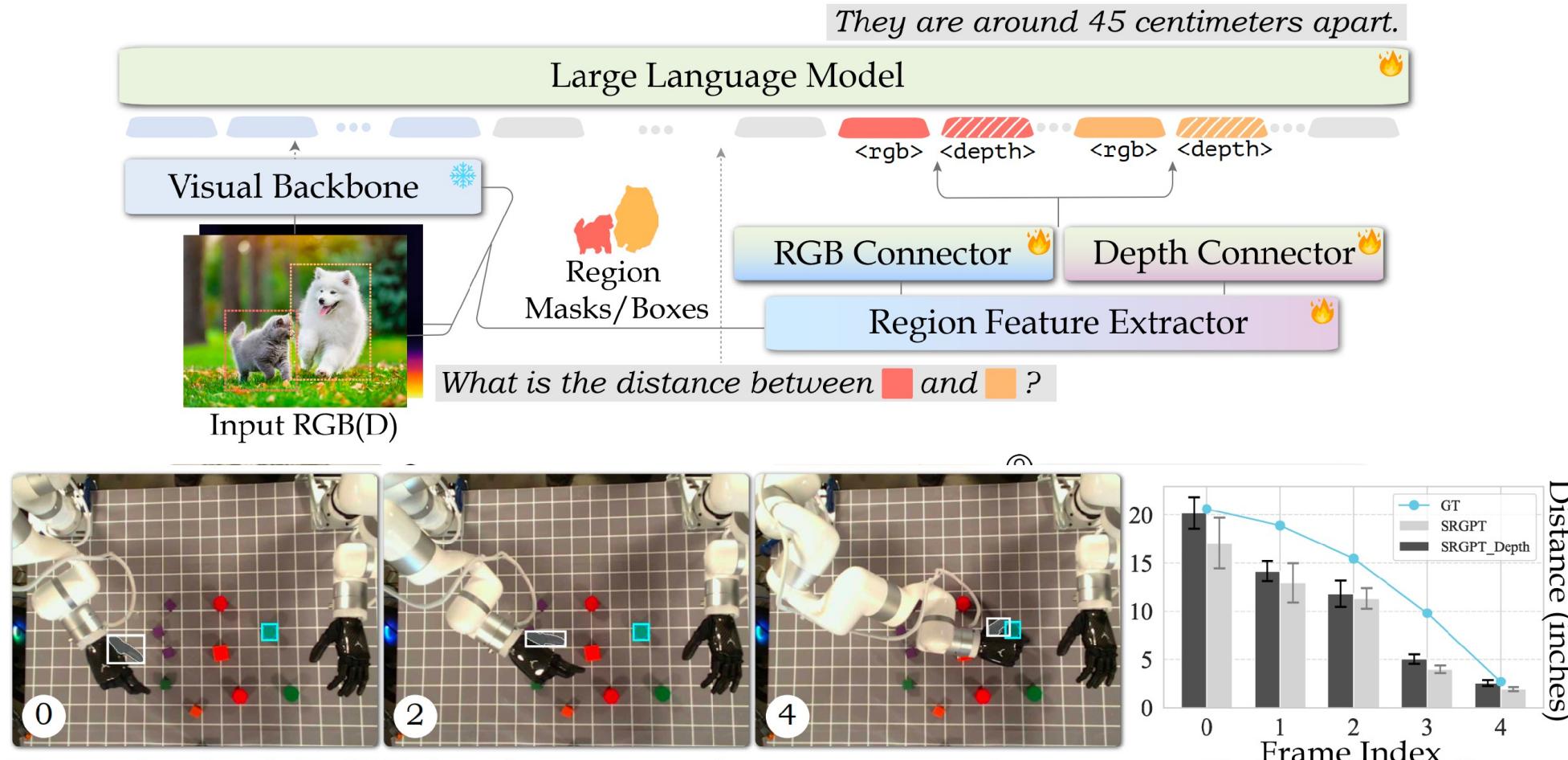
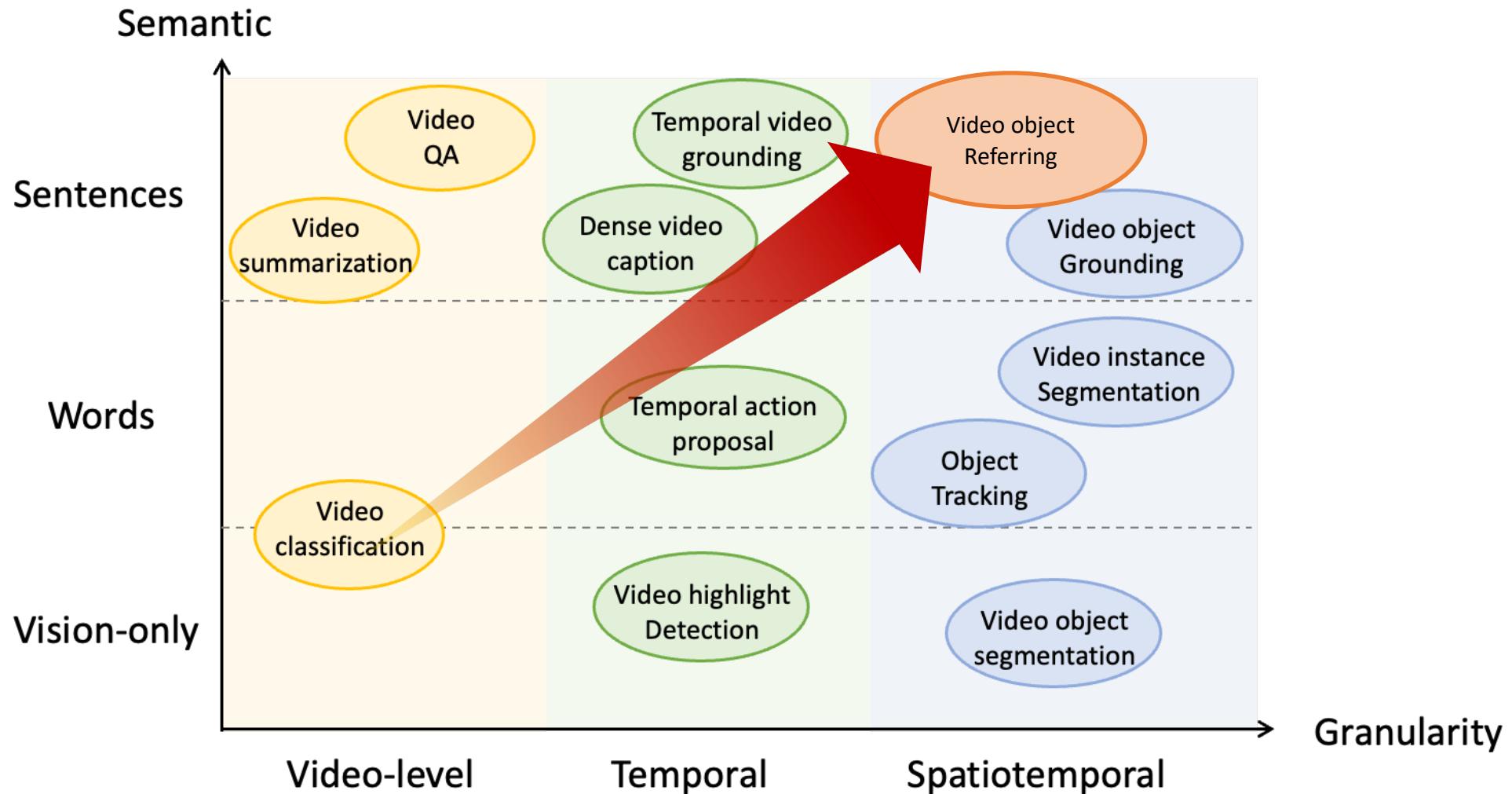


Figure 6: SpatialRGPT functions as a region-aware reward annotator. The estimated distance decreased monotonically as the fingertip moves towards the target.

# Fine-grained Spatiotemporal Understanding



# Fine-grained Spatiotemporal Understanding

## Video Object Referring



A man with a cocked hat and green robes, riding a horse, slowly riding from the left to the right.

## Future Reasoning



Q: What will <object1> probably do next?

A: <object1> will probably have to shoot or pass the ball to a teammate.

## Video Objects Relationship



The knife <object1> moves the spring onions from the chopping board <object2> to the pan.

## Video Object Retrieval



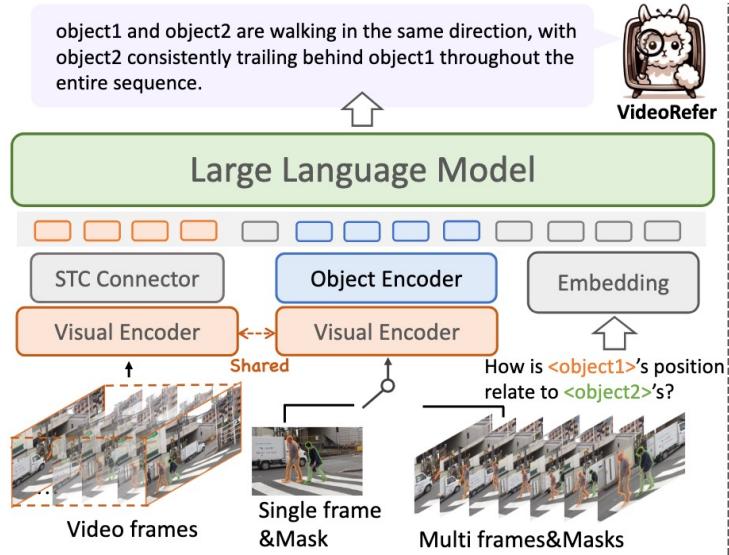
Input image



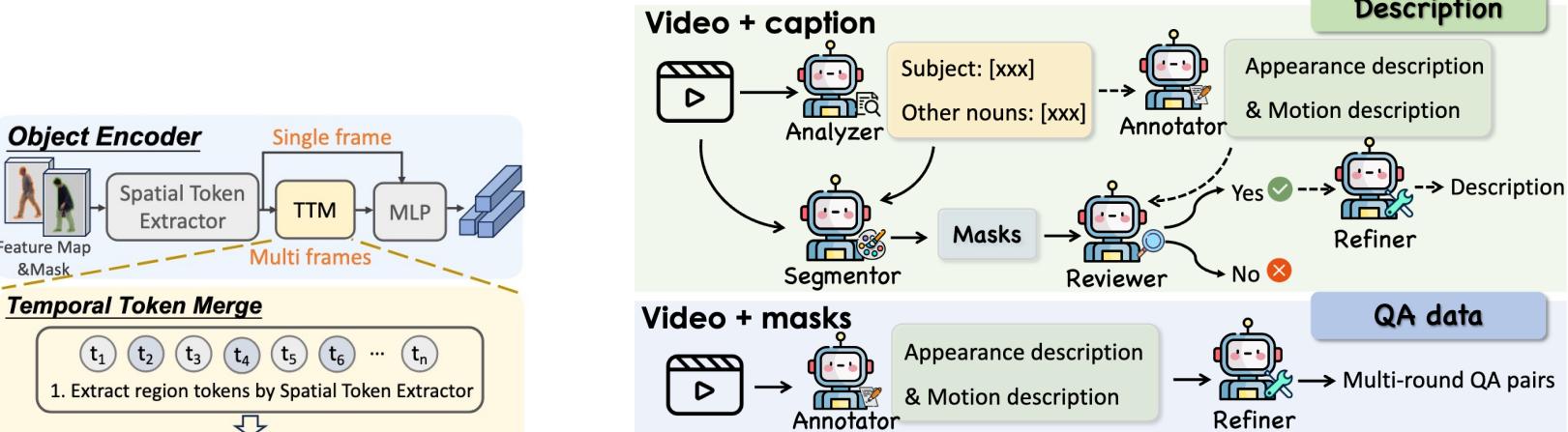
The man was Trump, who stood in the crowd waving and waving his fist to the left and right.

# Fine-grained Spatiotemporal Understanding

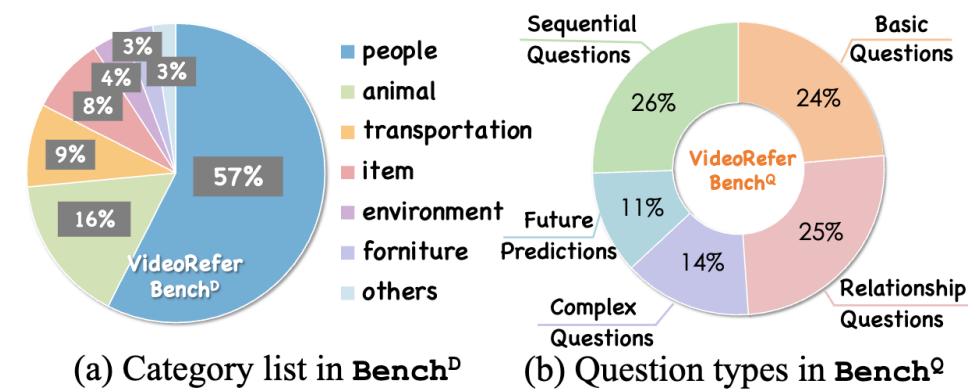
## VideoRefer Suite



- Spatiotemporal Region-level understanding Architecture;
- Constructing Large-scale Video Region Dataset;
- Evaluation Benchmarks for Video-based Object Understanding.



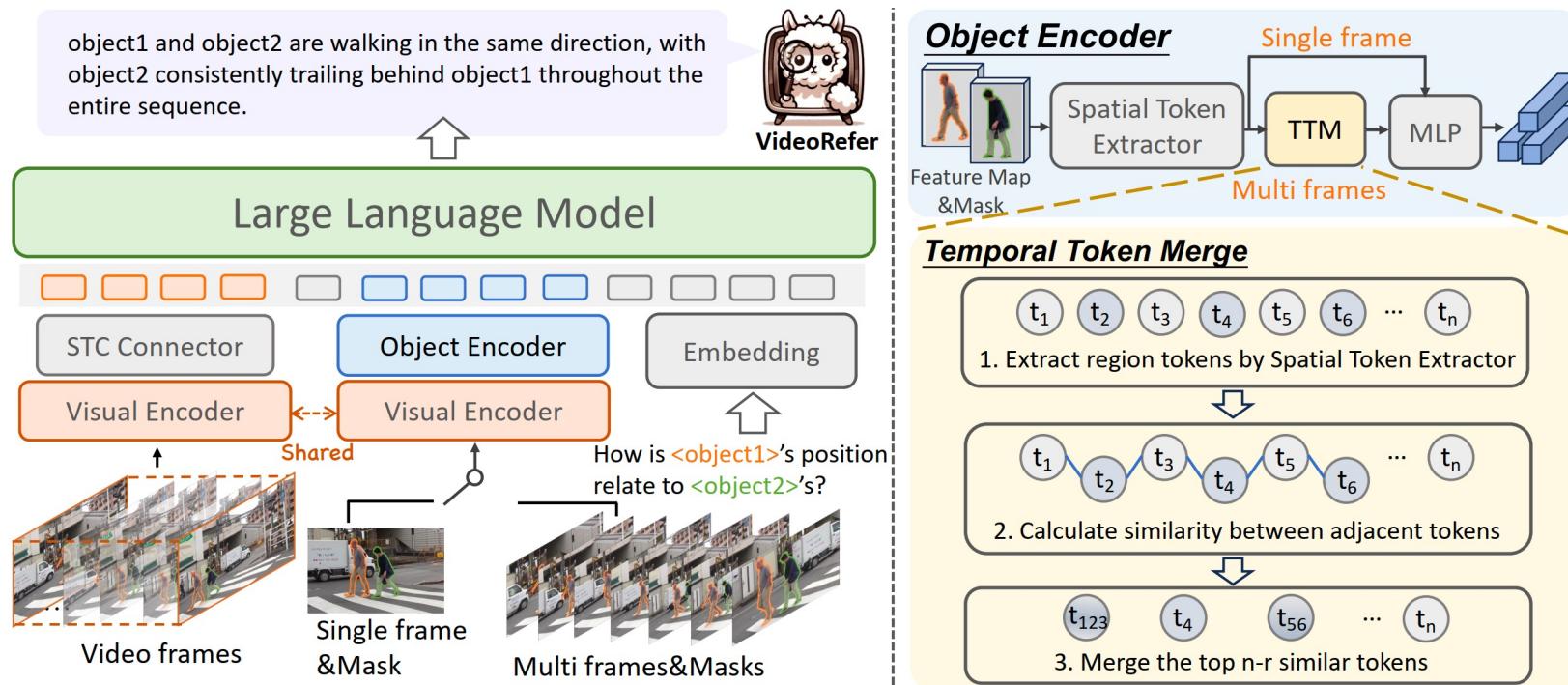
**VideoRefer-700K—Multi-agent Data Engine**



**VideoRefer-Bench**

# Fine-grained Spatiotemporal Understanding

## VideoRefer Model



A plug-and-play Spatial-Temporal Object Encoder:

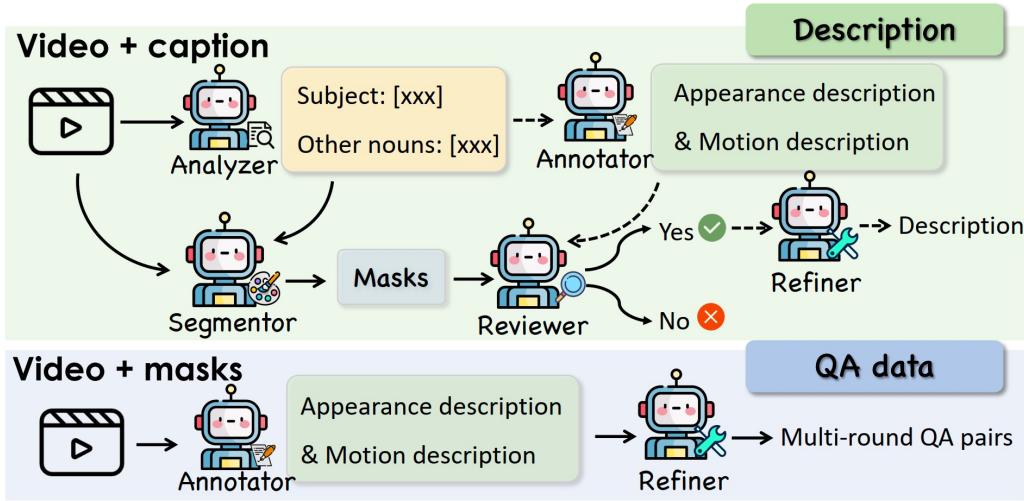
- Spatial Token Extractor (*Single-frame*)
- Temporal Token Merge Module (*Multi-frame*)
- Free-from input region (*Mask*)

Optimization Loss:

$$\mathcal{L} = \sum_{(V, R, x, y)} \log P(y | V, R_1, \dots, R_n, x)$$

# Fine-grained Spatiotemporal Understanding

## VideoRefer-700K



## Multi-agent Data Engine

Step1- Analyzer: Qwen2-Instruct-7B

Step2-Annotator: InternVL2-26B

Step3-Segmentor:Grounding DINO&SAM 2

Step4-Reviewer: Osprey&Qwen2-Instruct-7B

Step5-Refiner:GPT-4o

### Three types:

- Object-level Detailed Caption
- Object-level Short Capton
- Object-level QA

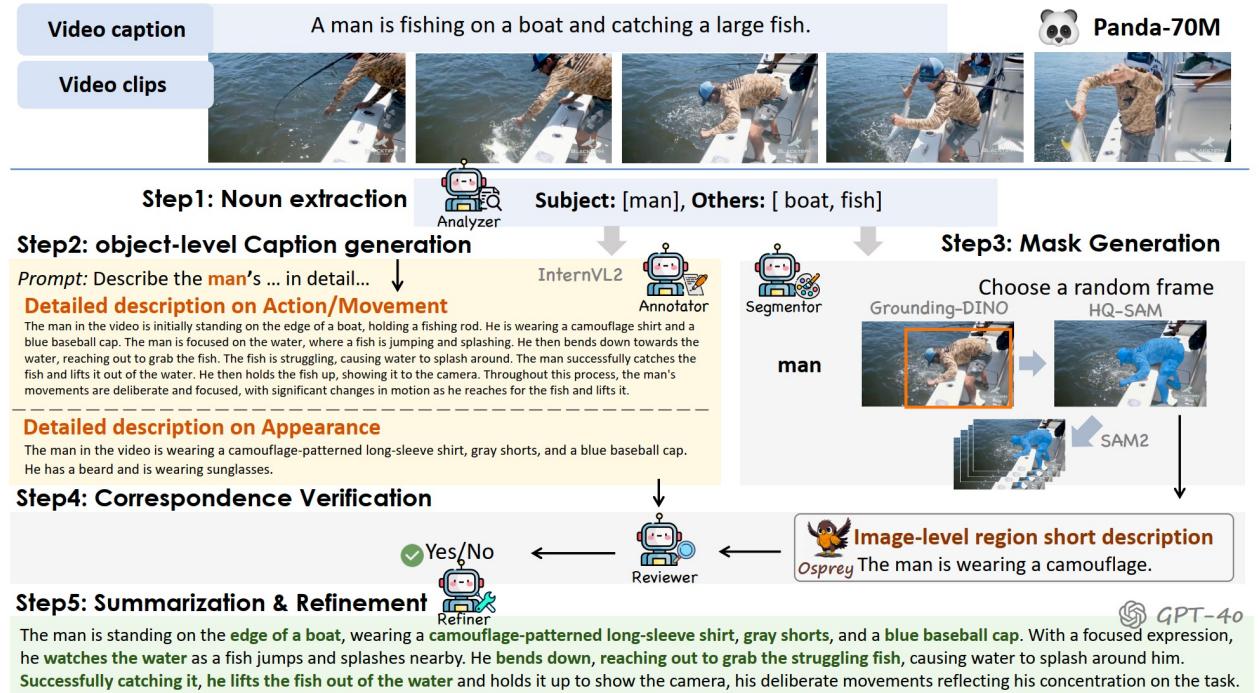


Table 8. Confusion matrix of the randomly sampled 100 items in the Reviewer evaluation.

	Manually True	Manually False
Reviewer True	88 (TP)	12 (FP)
Reviewer False	36 (FN)	64 (TN)

# Fine-grained Spatiotemporal Understanding

## VideoRefer-Bench

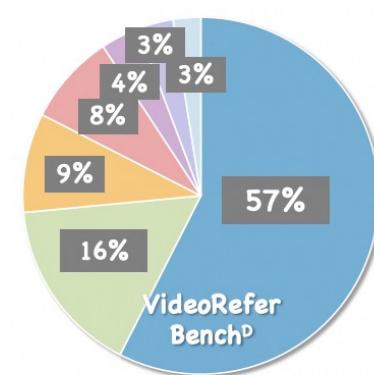
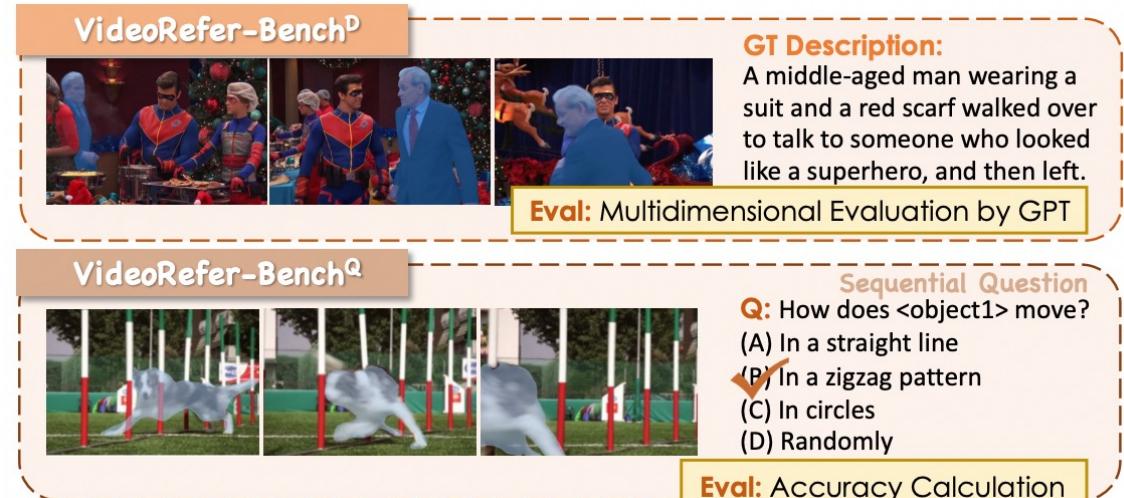
### VideoRefer-Bench<sup>D</sup> (*Description Generation*)

GPT assign scores from 0 to 5 across:

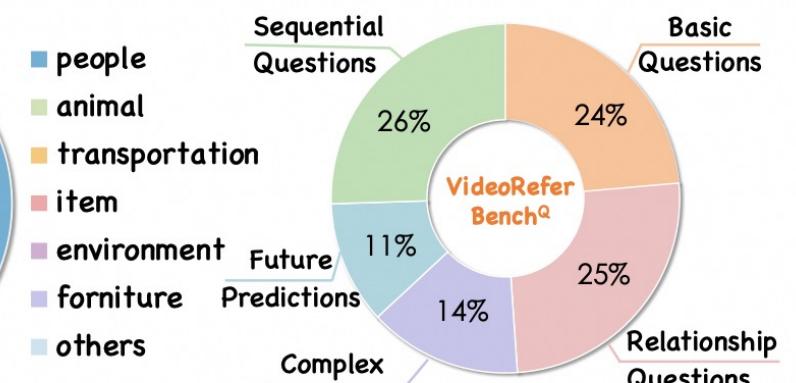
- Subject Correspondence
- Appearance Description
- Temporal Description
- Hallucination Detection

### VideoRefer-Bench<sup>Q</sup> (*Multi-choice QA*)

- Basic Questions
- Sequential Questions
- Relationship Questions
- Reasoning Questions
- Future Predictions



(a) Category list in **Bench<sup>D</sup>**



(b) Question types in **Bench<sup>Q</sup>**

# Fine-grained Spatiotemporal Understanding

## Describe Anything Model (DAM)



### Describe Anything: Detailed Localized Image and Video Captioning

Long Lian<sup>1,2</sup> Yifan Ding<sup>1</sup> Yunhao Ge<sup>1</sup> Sifei Liu<sup>1</sup> Hanzi Mao<sup>1</sup> Boyi Li<sup>1,2</sup> Marco Pavone<sup>1</sup>

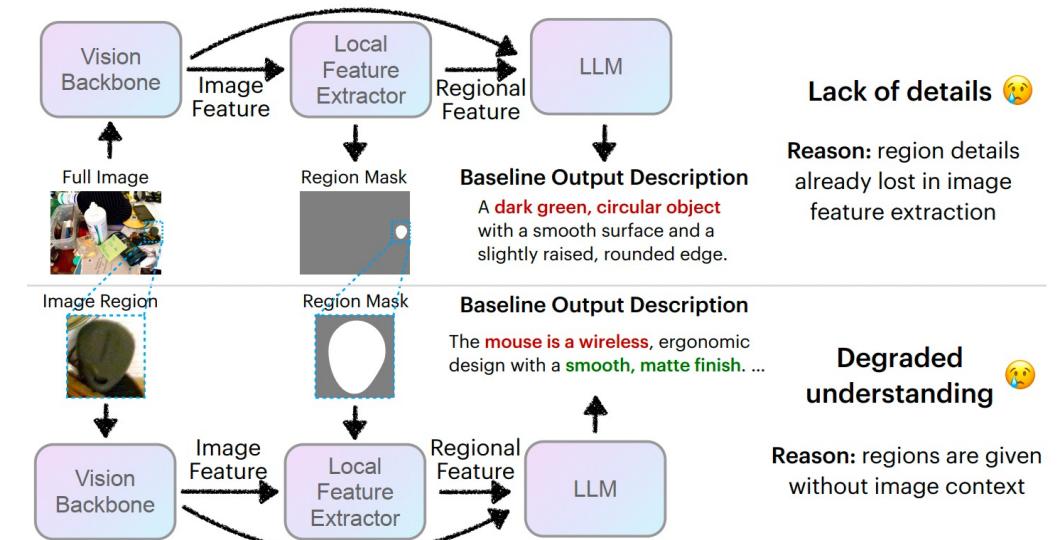
Ming-Yu Liu<sup>1</sup> Trevor Darrell<sup>2</sup> Adam Yala<sup>2,3</sup> Yin Cui<sup>1</sup>

<sup>1</sup>NVIDIA <sup>2</sup>UC Berkeley <sup>3</sup>UCSF



Figure 1: Describe Anything Model (DAM) generates **detailed localized captions** for user-specified regions within **images** (top) and **videos** (bottom). DAM accepts various region specifications, including clicks, scribbles, boxes, and masks. For videos, specifying the region in *any frame* suffices.

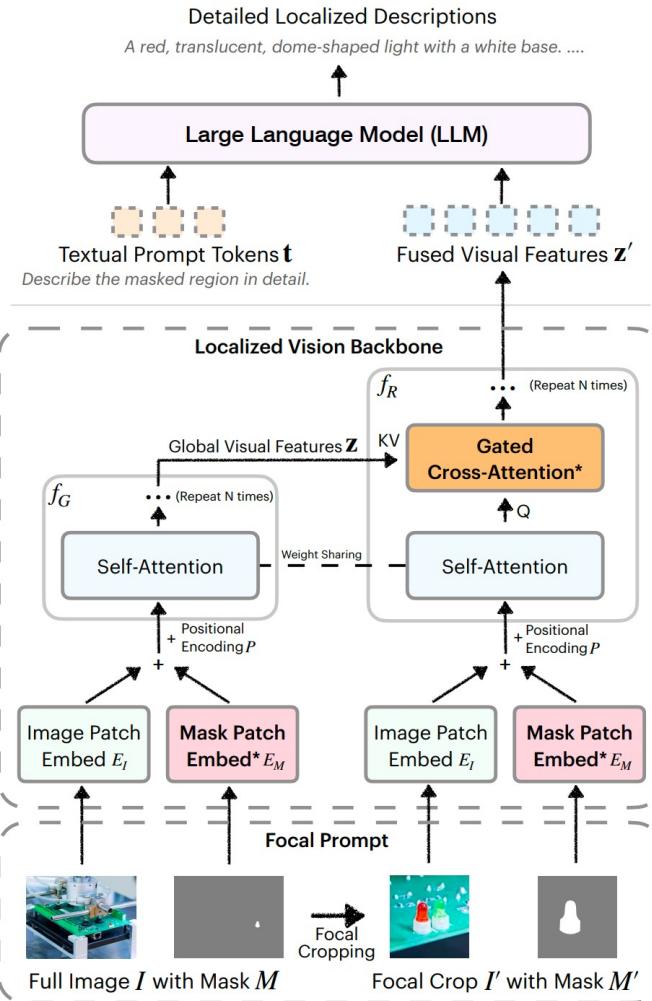
Adopting VideoRefer-Bench & Osprey Evaluation.



Typical two regional frameworks

# Fine-grained Spatiotemporal Understanding

## DAM



- **Focal Prompt**

**Full image and a zoomed-in region with corresponding mask**

$$x = E_I(I) + E_M(M) + P, \quad z = f_G(x)$$

$$x' = E_I(I') + E_M(M') + P, \quad z' = f_R(x', z)$$

- **Localized Vision Backbone**

**Inject global features into the encoding of local regions using Gated Cross-Attention Adaptor**

$$\mathbf{h}^{(l)'} = \mathbf{h}^{(l)} + \tanh(\gamma^{(l)}) \cdot \text{CrossAttn}(\mathbf{h}^{(l)}, \mathbf{z}),$$

$$\mathbf{h}_{\text{Adapter}}^{(l)} = \mathbf{h}^{(l)'} + \tanh(\beta^{(l)}) \cdot \text{FFN}(\mathbf{h}^{(l)'}, \mathbf{z}),$$

# Fine-grained Spatiotemporal Understanding

## Perceive Anything Model (PAM)

### Perceive Anything: Recognize, Explain, Caption, and Segment Anything in Images and Videos

Weifeng Lin<sup>1\*</sup> Xinyu Wei<sup>3\*</sup> Ruichuan An<sup>4\*</sup> Tianhe Ren<sup>2\*</sup> Tingwei Chen<sup>1</sup>  
Renrui Zhang<sup>1</sup> Ziyu Guo<sup>1</sup> Wentao Zhang<sup>4</sup> Lei Zhang<sup>3</sup> Hongsheng Li<sup>1†</sup>

<sup>1</sup>CUHK <sup>2</sup>HKU <sup>3</sup>PolyU <sup>4</sup>Peking University

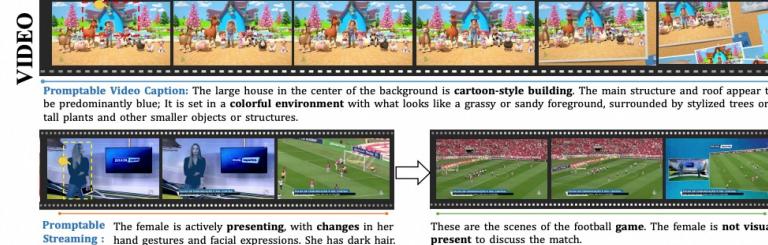
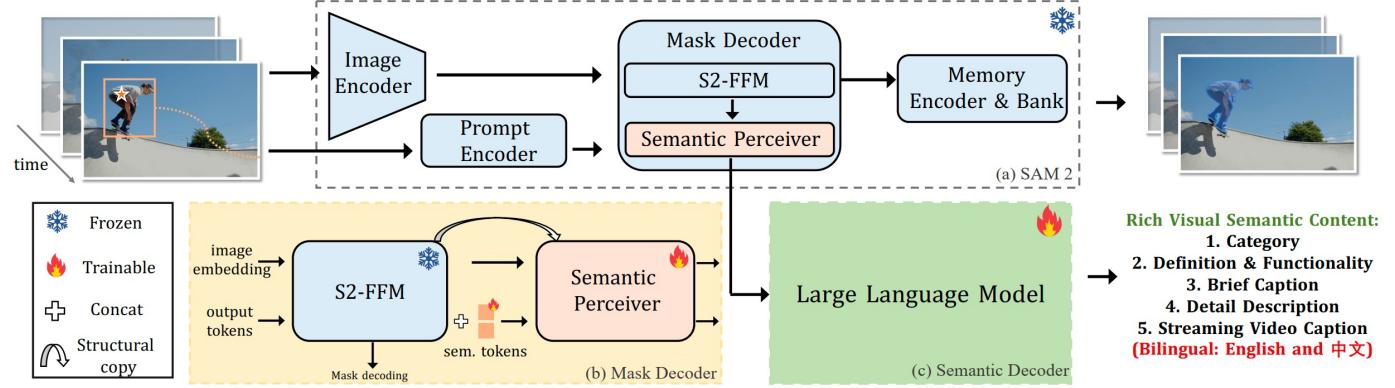


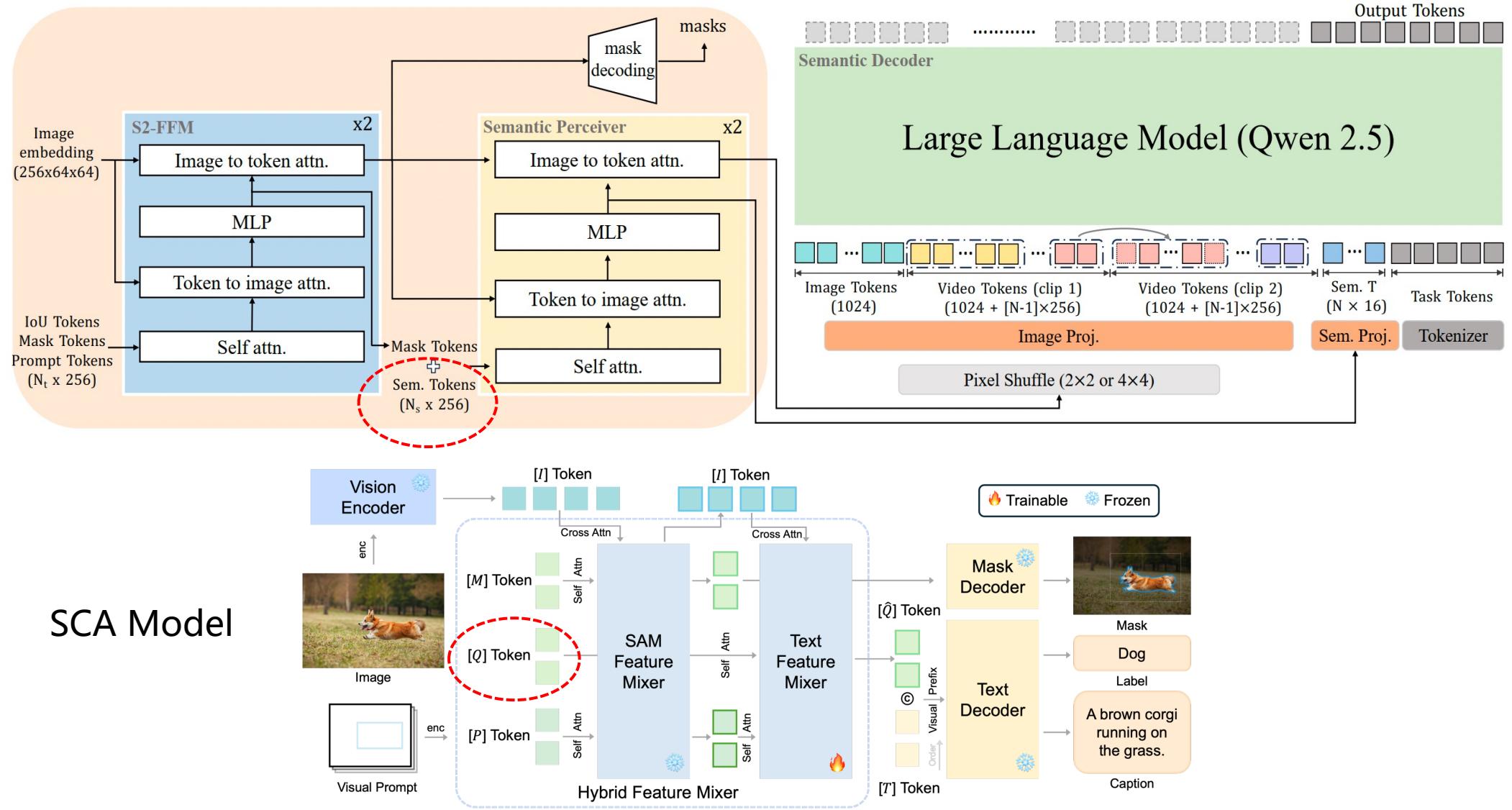
Figure 1: **Perceive Anything Model (PAM):** PAM accepts various visual prompts (such as clicks, boxes, and masks) to produce region-specific information for images and videos, including masks, category, label definition, contextual function, and detailed captions. The model also handles demanding region-level streaming video captioning.



- Extends **SAM 2** by extracting its intermediate visual features and transforming them into **LLM-compatible tokens**.
- Enables segmentation mask decoding and semantic content decoding simultaneously.

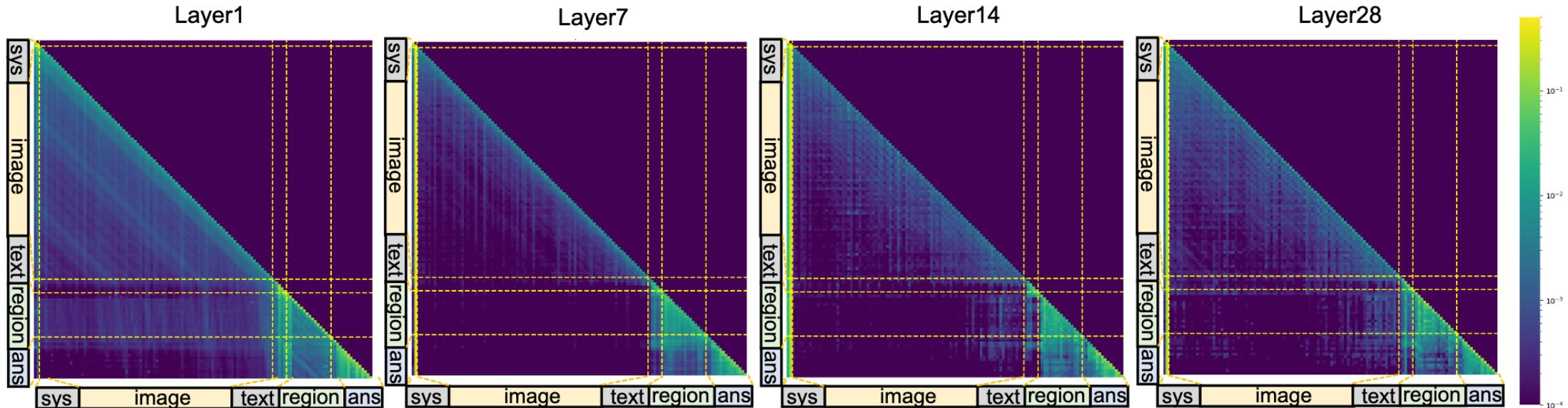
CUHK & HK PolyU

# Fine-grained Spatiotemporal Understanding



# Fine-grained Spatiotemporal Understanding

## PixelRefer

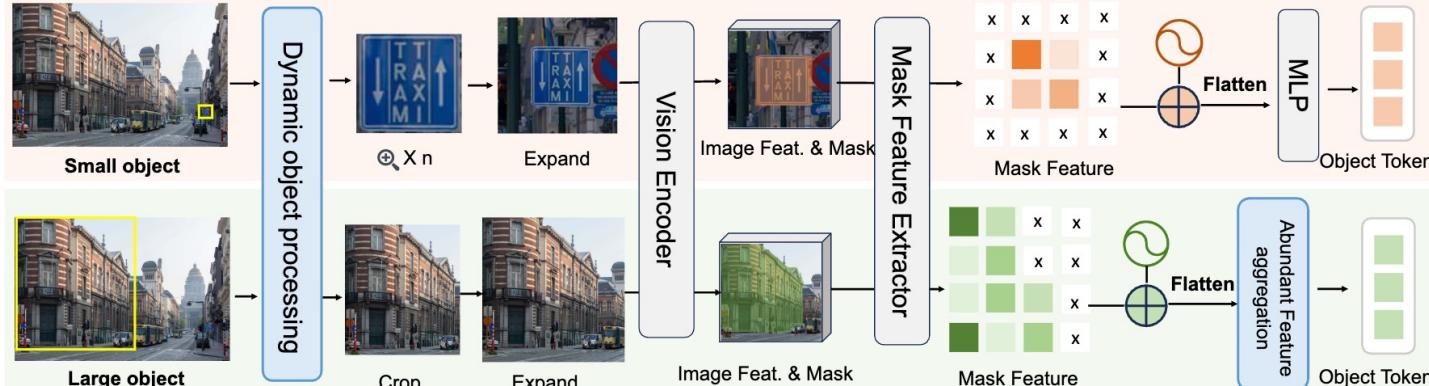


- Answer tokens prioritize object tokens → Construct robust region representation
- The attention between answer and image tokens are sparse → Vision-Object Framework
- Early fusion of object and image tokens

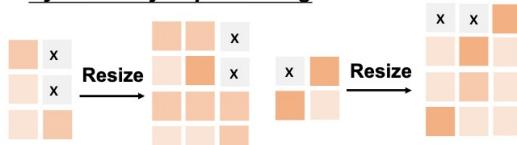
# Fine-grained Spatiotemporal Understanding

## PixelRefer

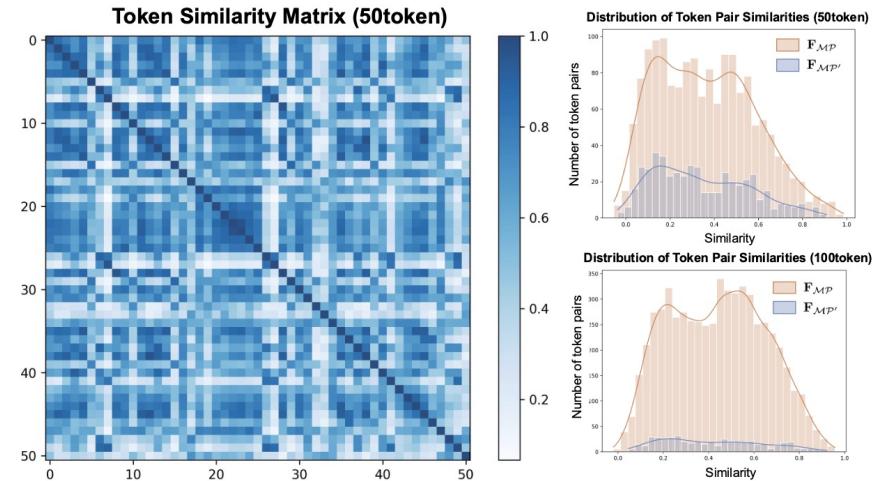
### Scale-adaptive Object Tokenizer



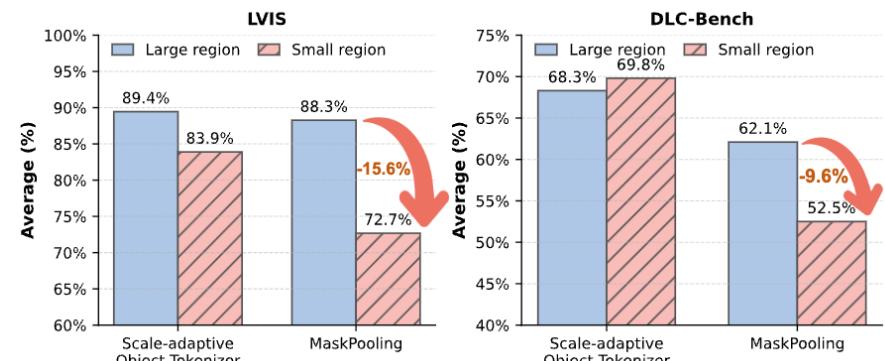
### Dynamic object processing



$$s = \begin{cases} \sqrt{\frac{\Omega \cdot 100}{|\mathcal{M}|}}, & \text{if } |\mathcal{M}| > 100 \cdot \Omega \\ \sqrt{\frac{\Omega \cdot n}{|\mathcal{M}|}}, & \text{elseif } |\mathcal{M}| < n \cdot \Omega \\ 1, & \text{otherwise.} \end{cases}$$



Token similarity

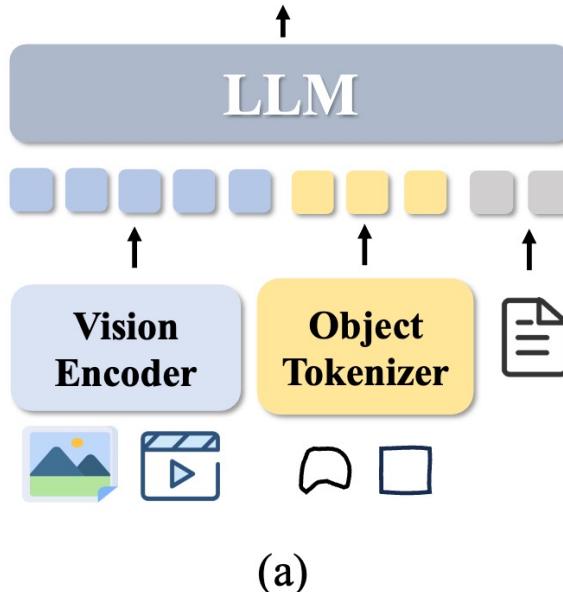


Accurate understanding of  
extremely small objects.

# Fine-grained Spatiotemporal Understanding

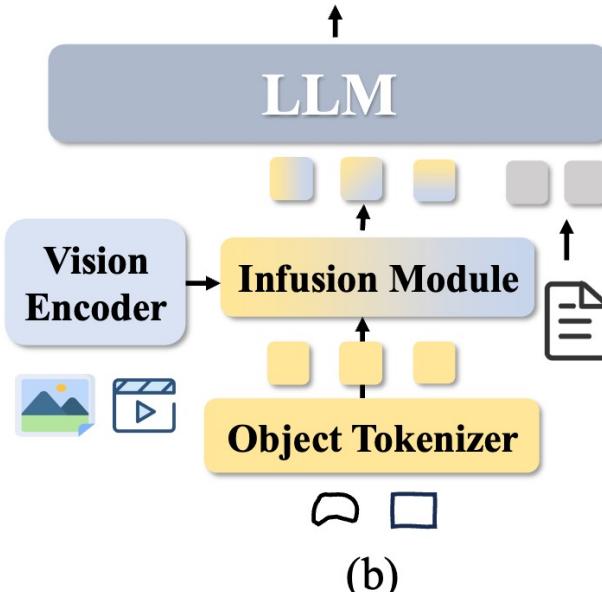
## PixelRefer

Vision-Object Framework



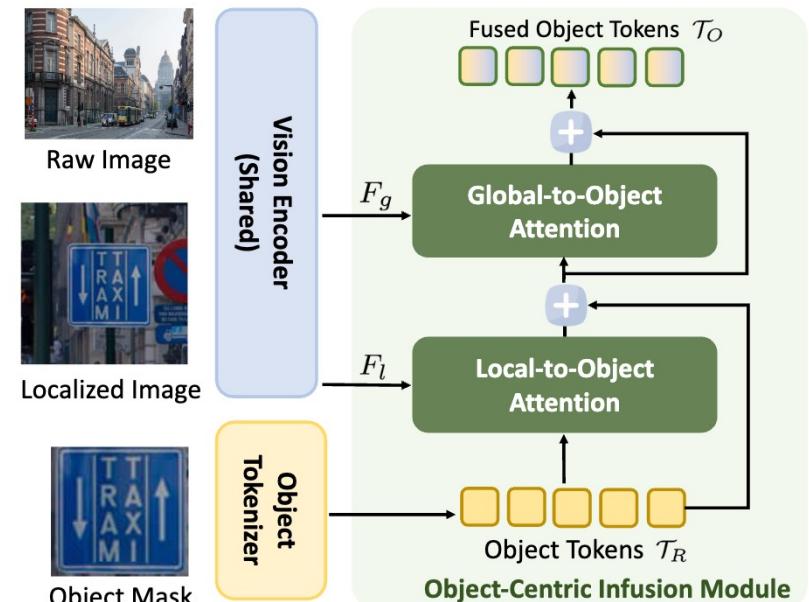
(a)

Object-Only Framework



(b)

Object-Centric Infusion Module

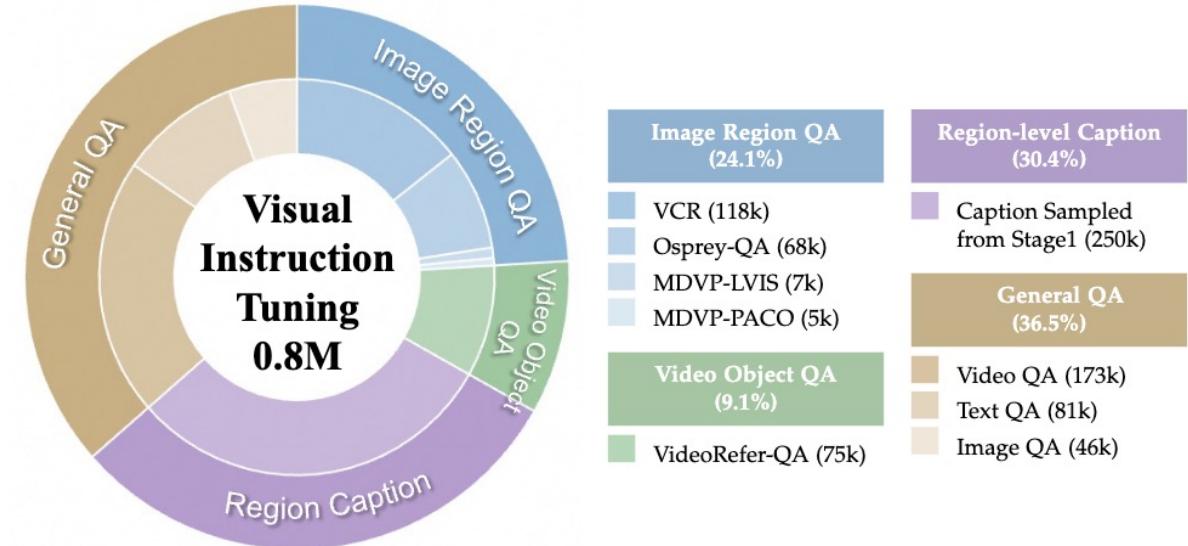
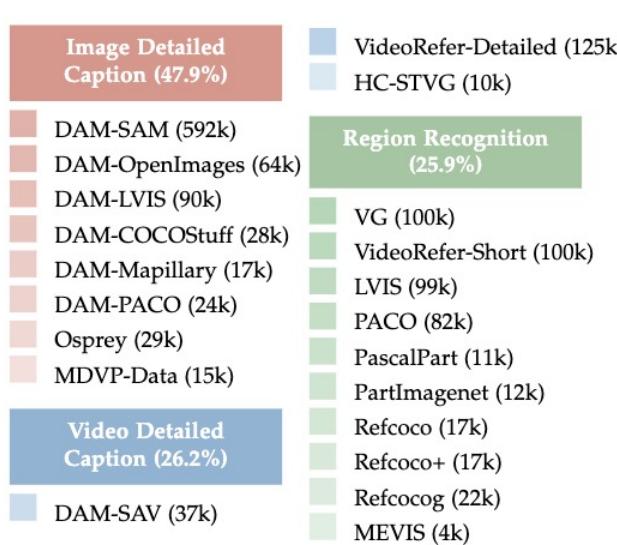
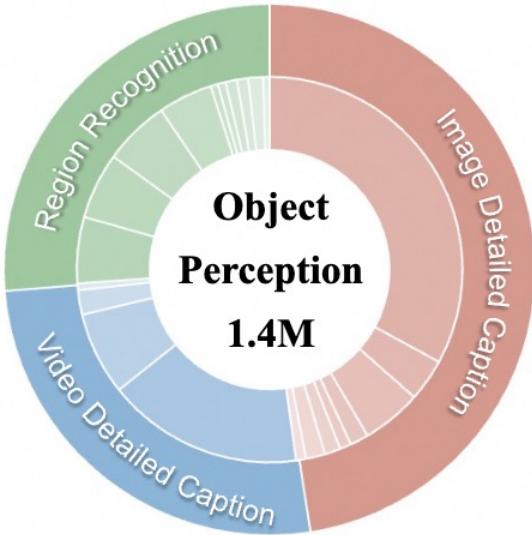


(a) Within LLM (inside LLM: vision and object tokens are fused)

(b) Before LLM (token fusion before feeding into LLM)

# Fine-grained Spatiotemporal Understanding

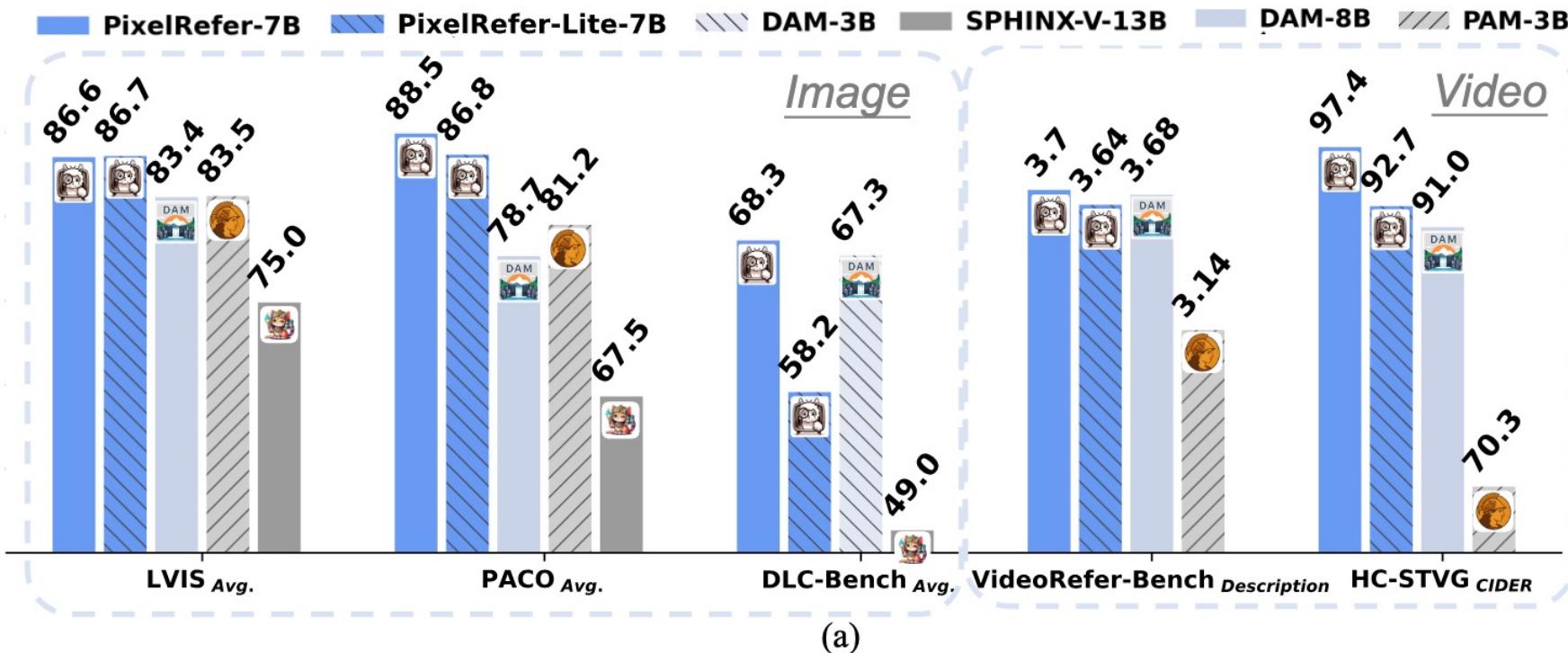
## PixelRefer



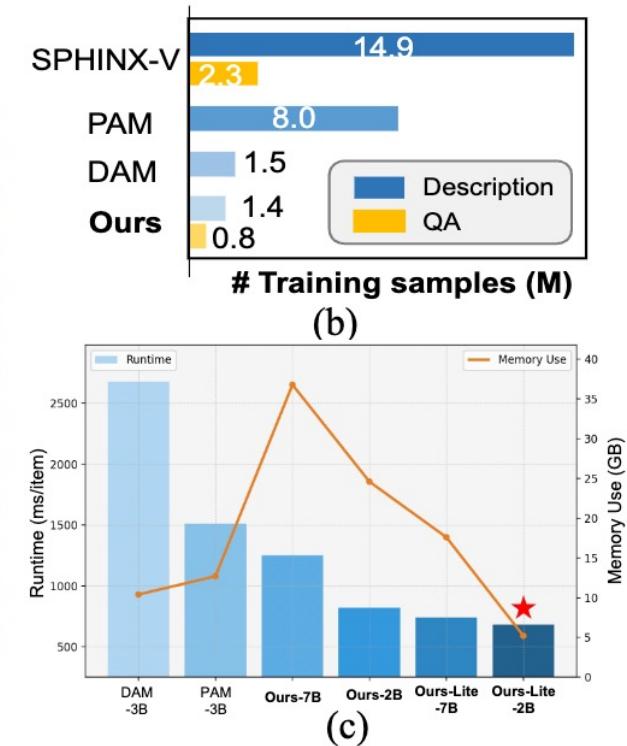
Data	#Samples	Image-Region-Bench		Video-Region-Bench			General-Bench	
		LVIS	DLC-Bench	HC-STVG	VideoRefer-D	VideoRefer-Q	POPE	MVBench
Region Recognition	390K	89.6	61.2	11.9	2.94	72.3	87.3	60.3
+ Image Detailed Cap.	860K	89.7	66.4	13.0	2.97	71.9	88.2	58.7
+ Video Detailed Cap.	180K	89.7	66.0	19.1	<b>3.69</b>	74.8	88.0	61.9
+ Region QA	560K	89.7	<b>66.6</b>	<b>19.6</b>	3.62	75.8	83.9	61.6
+ General QA	300K	<b>89.8</b>	66.1	19.5	3.58	<b>76.5</b>	<b>88.7</b>	<b>63.4</b>

# Fine-grained Spatiotemporal Understanding

**PixelRefer-Lite: Only 32 object tokens for each object without image tokens**



(a)



# Fine-grained Spatiotemporal Understanding

**PixelRefer-Lite: Only 32 object tokens for each object without image tokens**

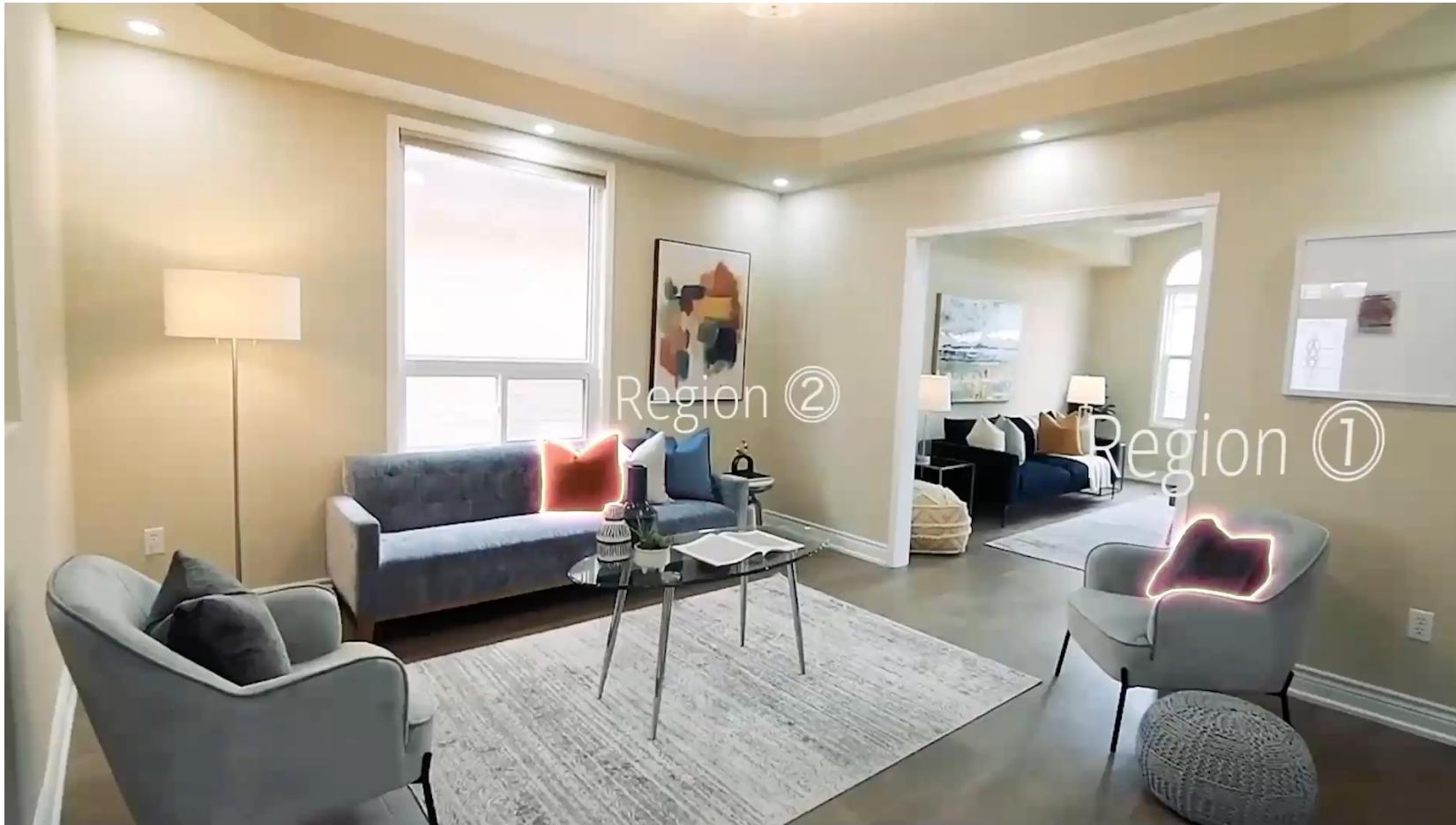
**FLOPs and memory consumption**

Method	L <sub>R</sub>	L <sub>Z</sub>	L <sub>Z<sub>G</sub></sub>	L <sub>Z<sub>L</sub></sub>	FLOPs(T)	Memory
<b>Image</b>						
PixelRefer-2B	32	~ 1408	–	–	1.51	13.2GB
PixelRefer-2B-Lite	32	0	576	256	0.03	4.9GB
PixelRefer-7B	32	~ 1408	–	–	7.08	25.1GB
PixelRefer-7B-Lite	32	0	576	256	0.17	15.8GB
<b>Video</b>						
PixelRefer-2B	32	~ 7185	–	–	11.15	24.6GB
PixelRefer-2B-Lite	32	0	576	256	0.11	5.1GB
PixelRefer-7B	32	~ 7185	–	–	43.83	36.9GB
PixelRefer-7B-Lite	32	0	576	256	0.61	17.6GB

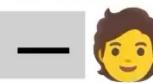
**Inference time and memory usage**

Model	DLC-Bench		HC-STVG	
	Infer Time	Memory	Infer time	Memory
DAM-3B	1.29s	7.8GB	5.64s	10.4GB
PAM-3B	1.09s	9.4GB	1.51s	12.7GB
<b>PixelRefer-2B</b>	1.04s	13.2GB	0.82s	24.6GB
<b>PixelRefer-Lite-2B</b>	0.88s	4.86GB	0.68s	5.2GB
<b>PixelRefer-7B</b>	1.44s	25.1GB	1.25s	36.9GB
<b>PixelRefer-Lite-7B</b>	1.10s	15.8GB	0.74s	17.6GB

# Fine-grained spatial reasoning

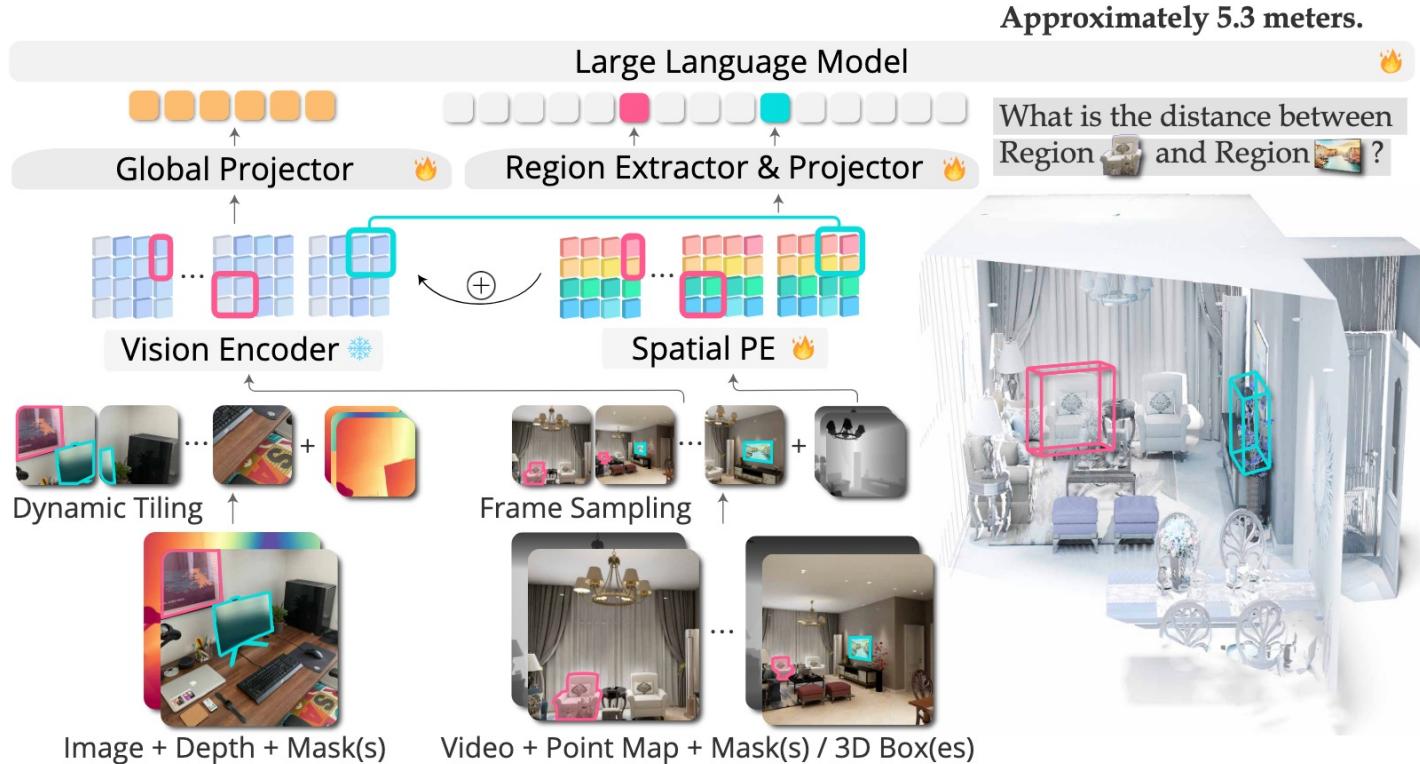


SR-3D



# Fine-grained spatial reasoning

## Architecture

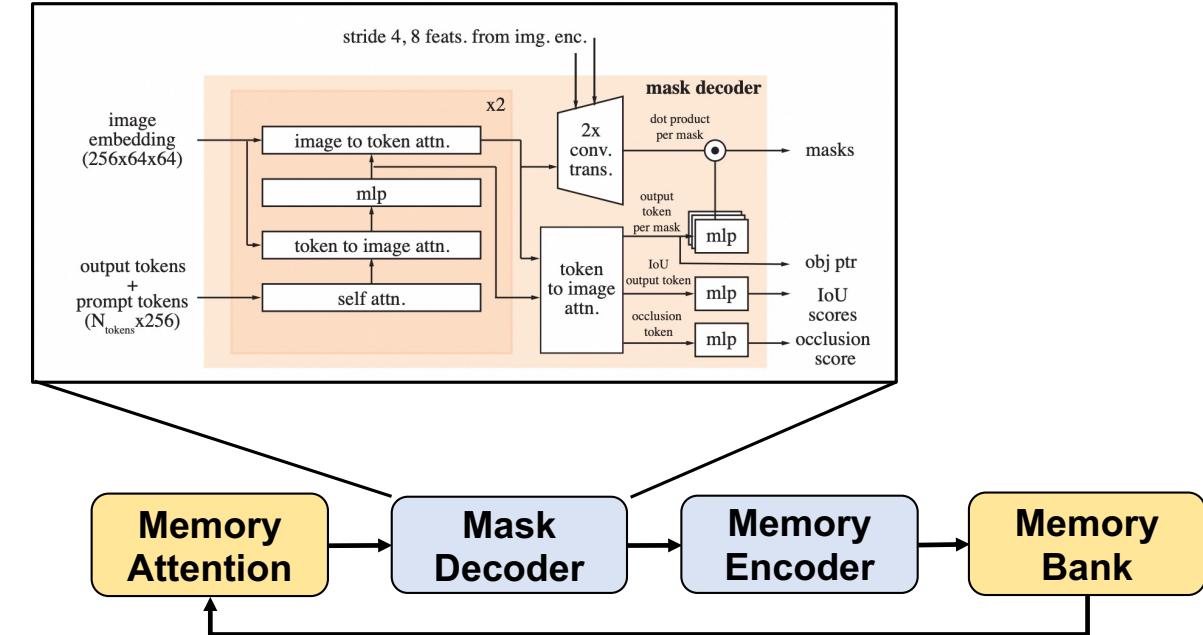
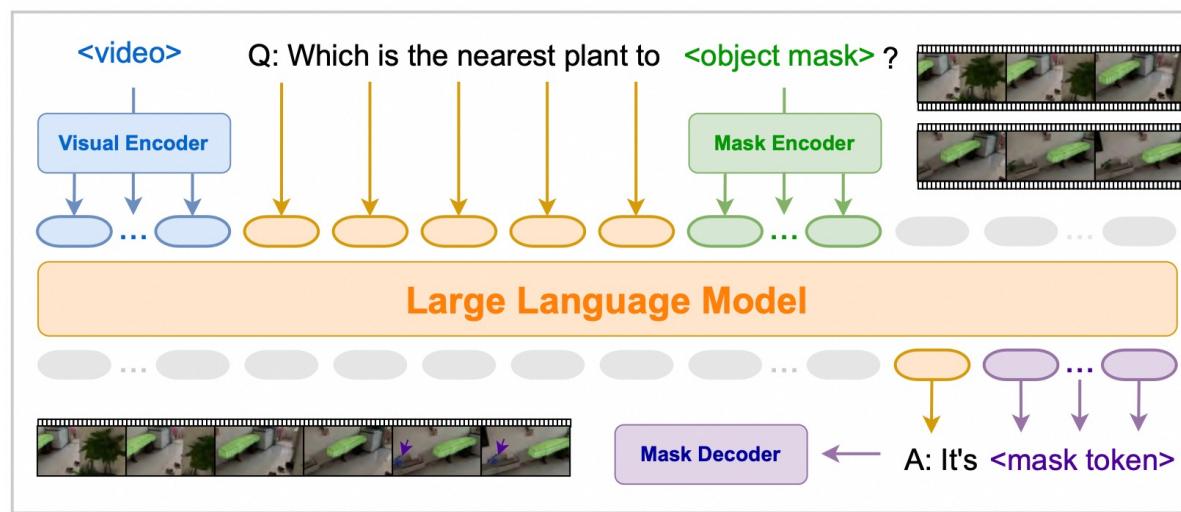


Methods	Quantitative			Qualitative	
	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Diff.
Random	-	-	-	25.0	36.1
Human Level <sup>†</sup>	47.0	60.4	45.9	94.7	95.8
<b>Proprietary Models (API)</b>					
GPT-4o [1]	5.3	43.8	38.2	37.0	41.3
Gemini-1.5 Flash [100]	30.8	53.5	54.4	37.7	41.0
Gemini-1.5 Pro [100]	30.9	64.1	43.6	51.3	46.3
<b>Open-source Models</b>					
InternVL2-2B [101]	24.9	22.0	35.0	33.8	44.2
InternVL2-8B [101]	28.7	48.2	39.8	36.7	30.7
InternVL2-40B [101]	26.9	46.5	31.8	42.1	32.2
LongVILA-8B [102]	9.1	16.7	0.0	29.6	30.7
VILA-1.5-8B [103]	21.8	50.3	18.8	32.1	34.8
VILA-1.5-40B [103]	24.8	48.7	22.7	40.5	25.7
LongVA-7B [104]	16.6	38.9	22.2	33.1	43.3
LLaVA-NeXT-Video-7B [71]	14.0	47.8	24.2	43.5	42.4
LLaVA-NeXT-Video-72B [71]	22.8	57.4	35.3	42.4	36.7
LLaVA-OneVision-0.5B [105]	28.4	15.4	28.3	28.9	36.9
LLaVA-OneVision-7B [105]	20.2	47.4	12.3	42.5	35.2
LLaVA-OneVision-72B [105]	23.9	57.6	37.5	42.5	39.9
<b>SR-3D-8B</b>	<b>52.8</b>	<b>75.5</b>	<b>41.9</b>	<b>57.3</b>	<b>82.3</b>

Results on VSI-Bench

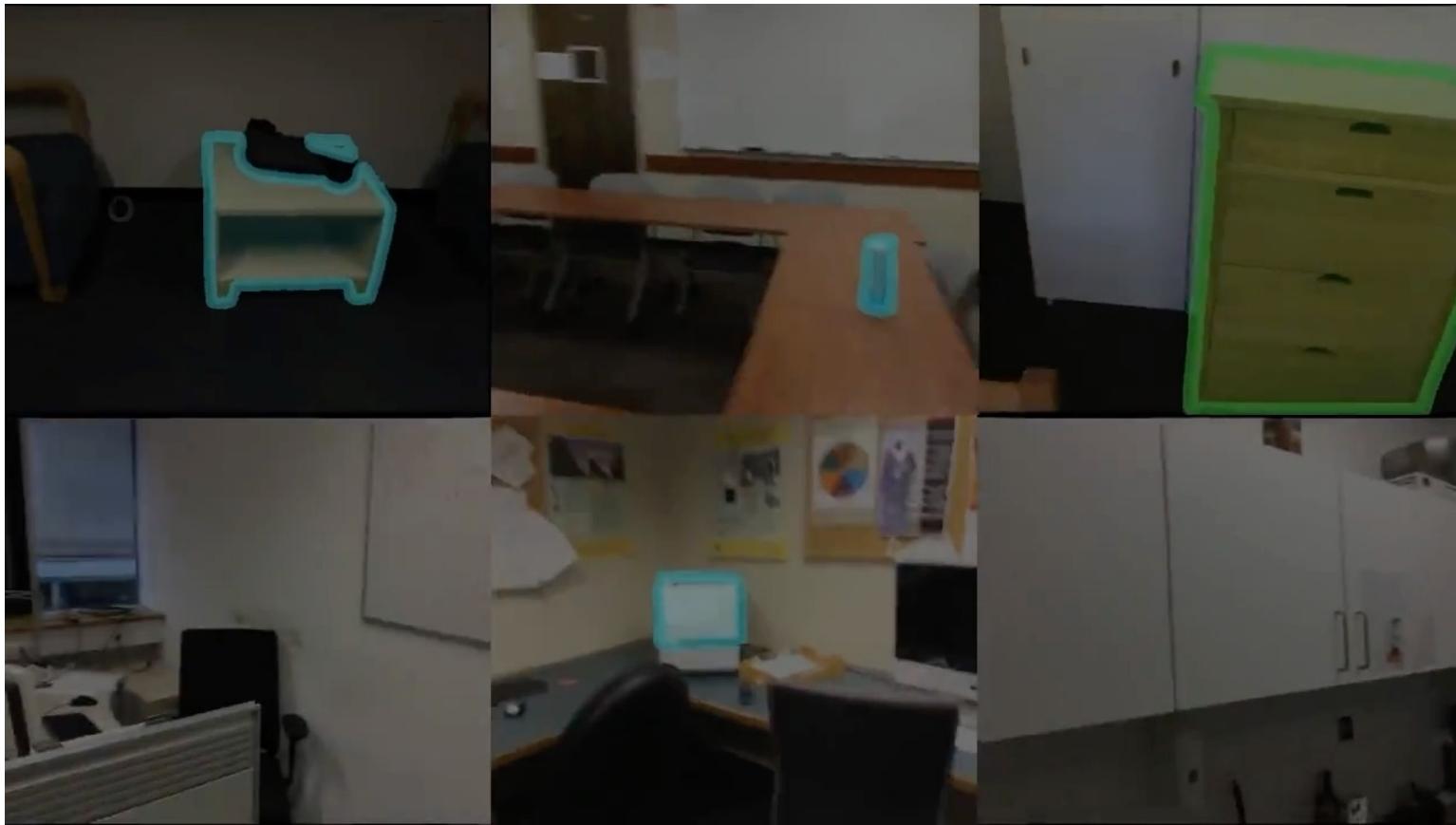
# RynnEC: Bringing MLLMs into Embodied World

## Model



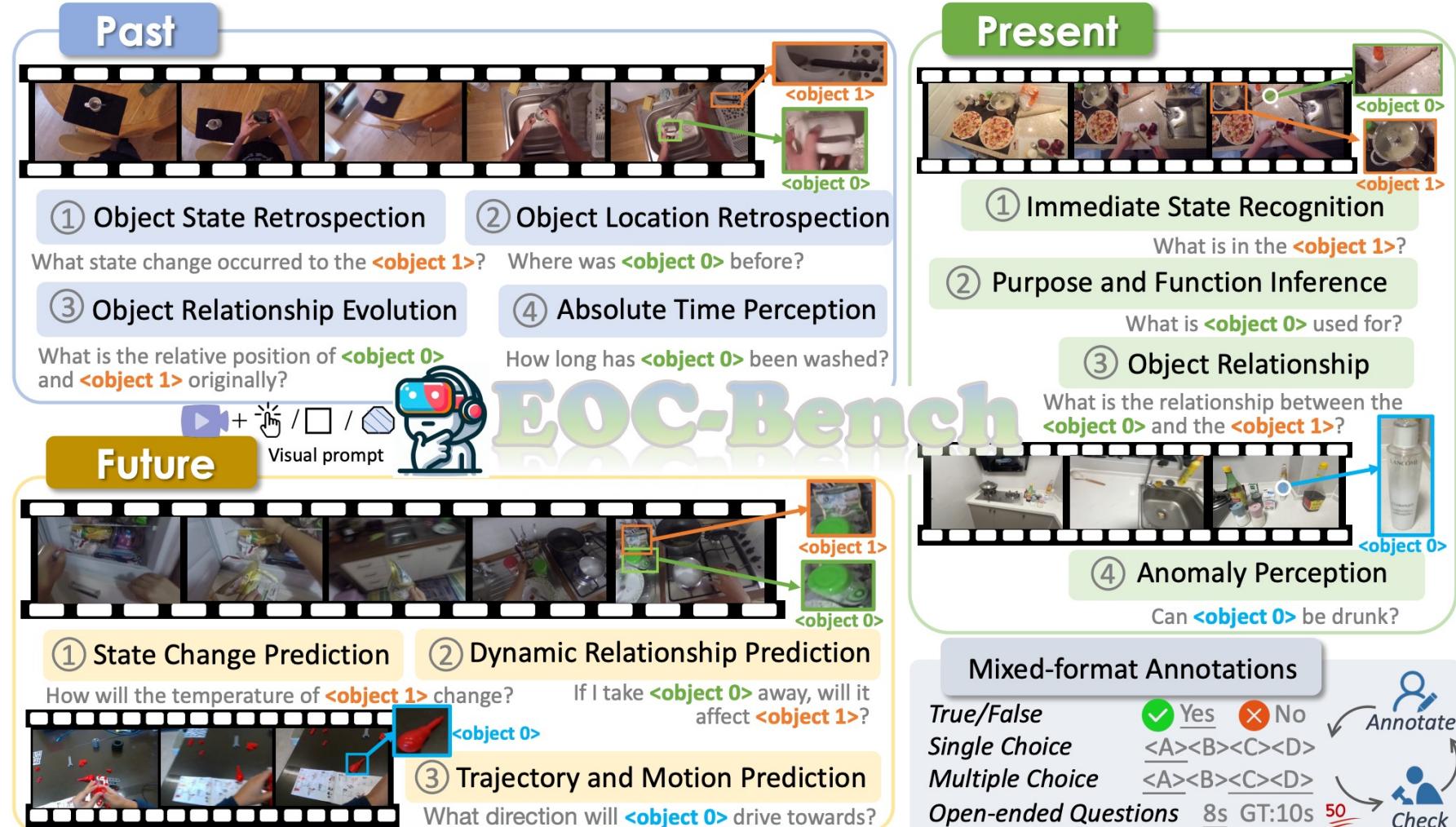
- Based on VideoLLaMA3
- **Mask Encoder** for object-text alignment
- **Mask Decoder** for grounding and segmentation (SAM2)

# Fine-grained Spatiotemporal Understanding in Embodied Recognition

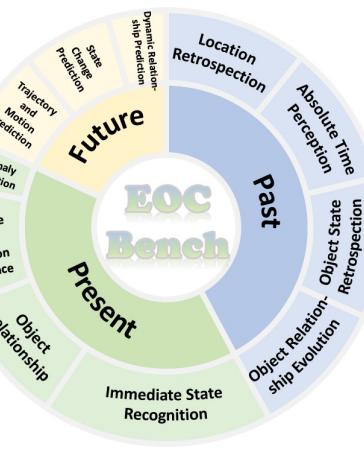


# Fine-grained Spatiotemporal Understanding in Dynamic Embodied World?

- Temporal dimensions: Past, Present and Future



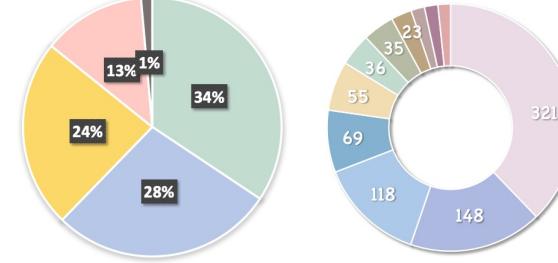
# Fine-grained Spatiotemporal Understanding in Dynamic Embodied World?



(a) Overview of EOC-Bench dimensions



EPIC-KITCHENS (239) Ego4D (201) Charades-ego (294) MECCANO (12) Self-recording (110)



(b) Video source distribution (c) Number of various scenario categories



# Fine-grained Spatiotemporal Understanding in Dynamic Embodied World?

EOC-Bench Leaderboard																	
#	Method	Input	Mean	Past					Present					Future			
				OSR	OLR	ORE	ATP	Mean	ISR	OR	PFI	AP	Mean	TMP	SCP	DRP	Mean
1	GPT-4o 🥇	32f	61.83	66.04	71.93	46.56	34.46	54.91	71.46	52.85	78.18	62.75	67.32	69.61	68.69	68.97	69.11
2	Gemini-2.0-flash 🏆	32f	57.38	63.46	65.10	32.56	28.60	47.87	68.84	57.52	69.68	65.69	65.95	58.54	64.02	57.95	60.75
3	InternVL2.5-78B 🥈	32f	52.33	53.46	63.96	33.15	12.01	41.35	66.67	50.74	67.10	52.94	61.72	67.80	50.47	54.55	58.19
4	InternVL2.5-38B	32f	52.31	55.40	59.62	30.92	10.89	39.89	64.15	54.28	71.29	64.71	63.35	60.98	54.67	57.95	57.79
5	Qwen2.5-VL-72B	1fps	49.87	51.25	51.22	40.11	8.48	38.41	61.31	47.79	67.10	57.84	58.98	56.10	60.65	54.55	57.76
6	LLaVA-Video-72B	32f	49.59	49.03	56.91	26.74	24.02	39.59	63.32	47.20	63.87	50.00	58.38	56.10	55.14	47.73	54.24
7	GPT-4o-mini	32f	49.47	53.26	52.35	29.68	21.10	39.47	58.46	49.26	67.74	58.82	58.31	56.59	50.00	54.55	53.45
8	LLaVA-OV-72B	32f	47.88	46.81	50.95	26.46	12.91	34.81	64.15	51.33	64.52	49.02	59.87	58.05	46.73	54.55	52.66
9	VideoLLaMA3-7B	1fps	46.04	45.15	52.85	24.51	15.54	35.00	57.96	48.67	62.58	49.02	56.01	52.20	49.54	48.86	50.49
10	InternVL2.5-8B	32f	45.15	45.71	54.47	39.00	9.76	37.87	55.44	48.97	54.84	41.18	52.60	49.76	38.79	53.41	45.76
11	Qwen2.5-VL-7B	1fps	43.13	47.37	46.34	21.45	8.18	31.38	57.29	44.54	59.35	49.02	53.93	48.78	46.30	46.59	47.35
12	LLaVA-Video-7B	32f	41.82	44.32	48.51	22.56	9.76	31.82	54.27	43.66	55.81	49.02	51.56	45.85	40.65	47.73	43.98
13	VideoLLaMA2-72B	16f	41.55	43.77	51.22	24.23	6.46	32.03	50.08	37.46	58.06	45.10	48.37	49.27	50.47	51.14	50.10
14	LLaVA-OV-7B	32f	40.46	40.72	45.53	22.84	9.53	30.15	54.10	43.07	52.58	46.08	50.37	47.32	37.38	46.59	43.00
15	VideoRefer-7B	16f	40.44	47.37	55.01	23.40	10.59	34.69	48.91	39.82	53.55	38.24	46.88	41.95	35.51	43.18	39.45
16	VideoLLaMA3-2B	1fps	38.41	37.12	46.88	21.17	11.26	29.57	49.92	43.36	48.39	38.24	47.03	43.41	36.11	43.18	40.28
17	Qwen2.5-VL-3B	1fps	38.17	38.78	48.78	23.96	7.66	30.34	49.92	38.94	45.16	38.24	45.18	42.93	36.57	50.00	41.45
18	VideoLLaMA2.1-7B	16f	37.74	44.88	42.82	19.22	11.64	30.08	47.24	37.17	51.94	39.22	45.18	40.00	36.92	44.32	39.45
19	NVILA-8B	32f	37.69	37.40	46.61	20.89	12.09	29.69	44.39	41.59	49.03	46.08	44.88	42.44	38.32	44.32	41.03
20	LongVA-7B	32f	35.34	36.84	43.36	17.83	15.32	28.69	38.19	36.58	48.06	42.16	40.36	39.02	42.06	40.91	40.63
21	VideoLLaVA-7B	8f	34.11	31.86	37.94	27.58	13.14	27.97	41.04	35.10	40.97	37.25	39.24	40.98	31.78	44.32	37.67



**PRCV 2025**  
智汇沪上，海纳视界



*Thanks !*

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

wentong\_li@nuaa.edu.cn