北京大学 PEKING UNIVERSITY

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

Weifeng Lin[1*] Xinyu Wei[3*] Ruichuan An[4*] Tianhe Ren[2*] Tingwei Chen[1]
Renrui Zhang[1] Ziyu Guo[1] Wentao Zhang[4] Lei Zhang[3] Hongsheng Li[1]

[1]CUHK [2]HKU [3]PolyU [4]PKU

arXiv 2025.06

Presenter: Shaofan Sun
2025.7.21

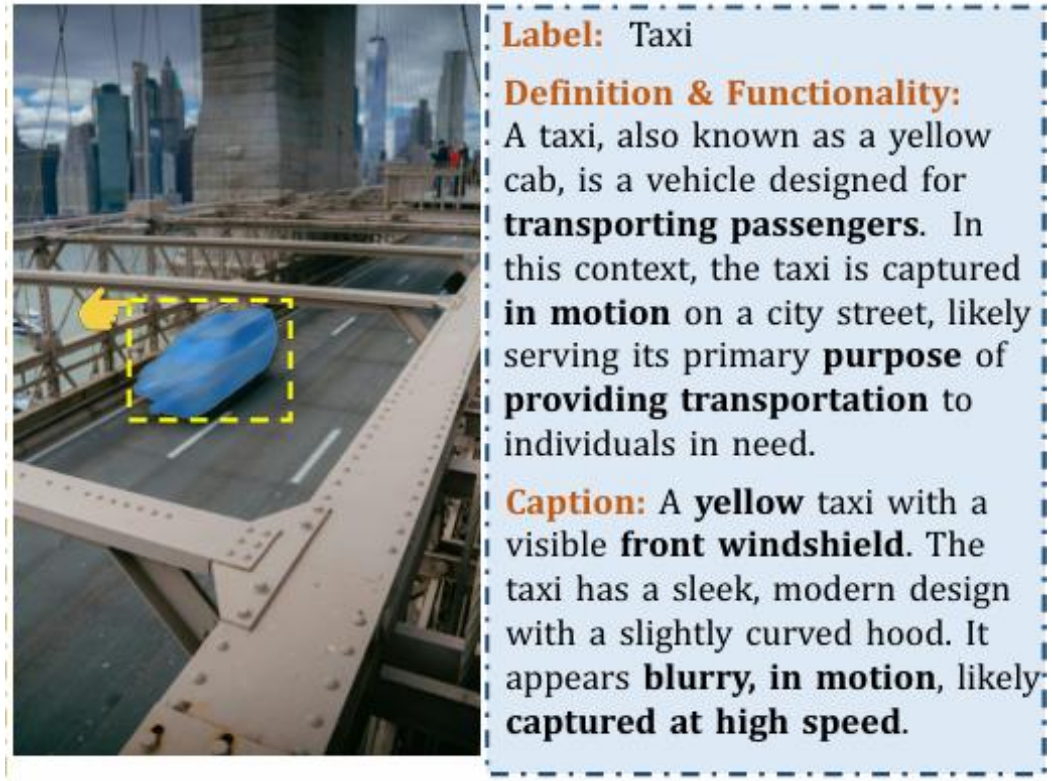
Directory

北京大学 PEKING UNIVERSITY

Background: VLMs in region-level understanding

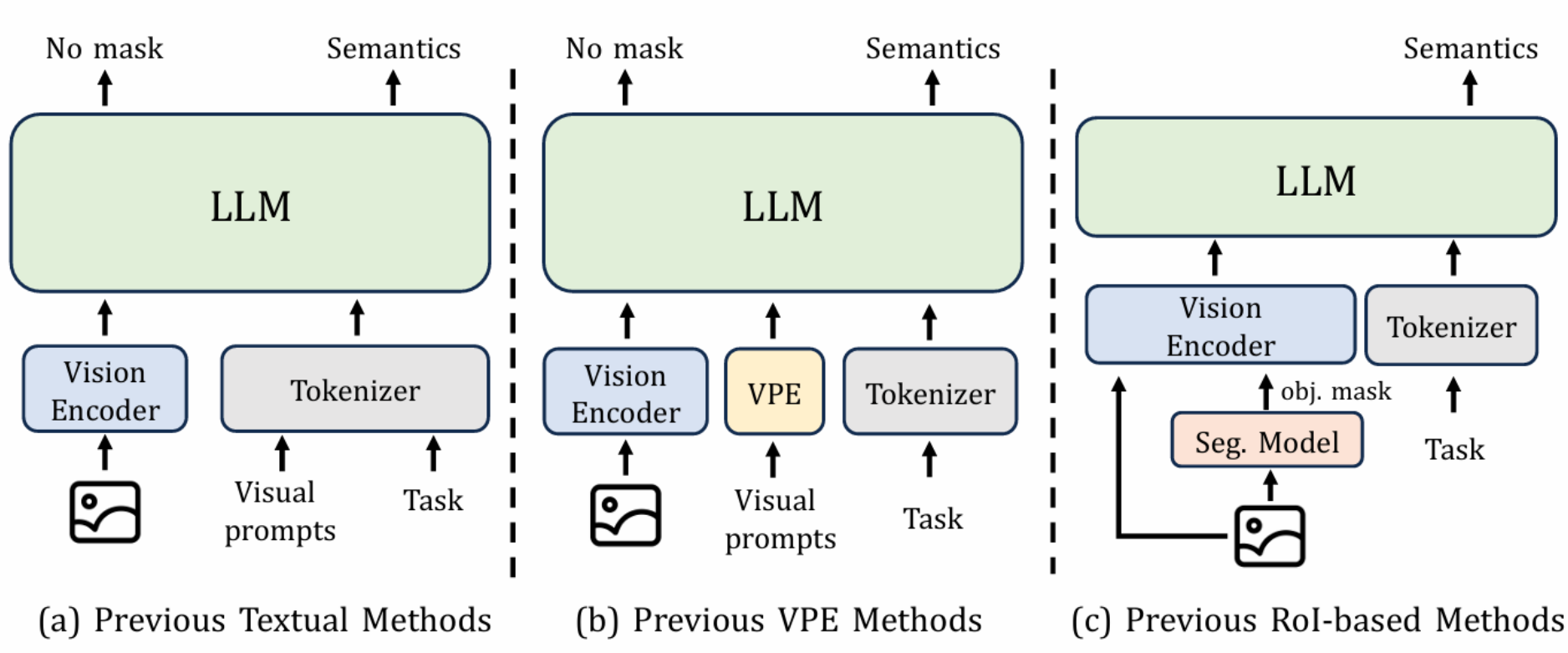
- Region-level understanding focuses on fine-grained parsing and understanding of semantics, attributes, and relationships in specific regions of images or videos.

- Example:

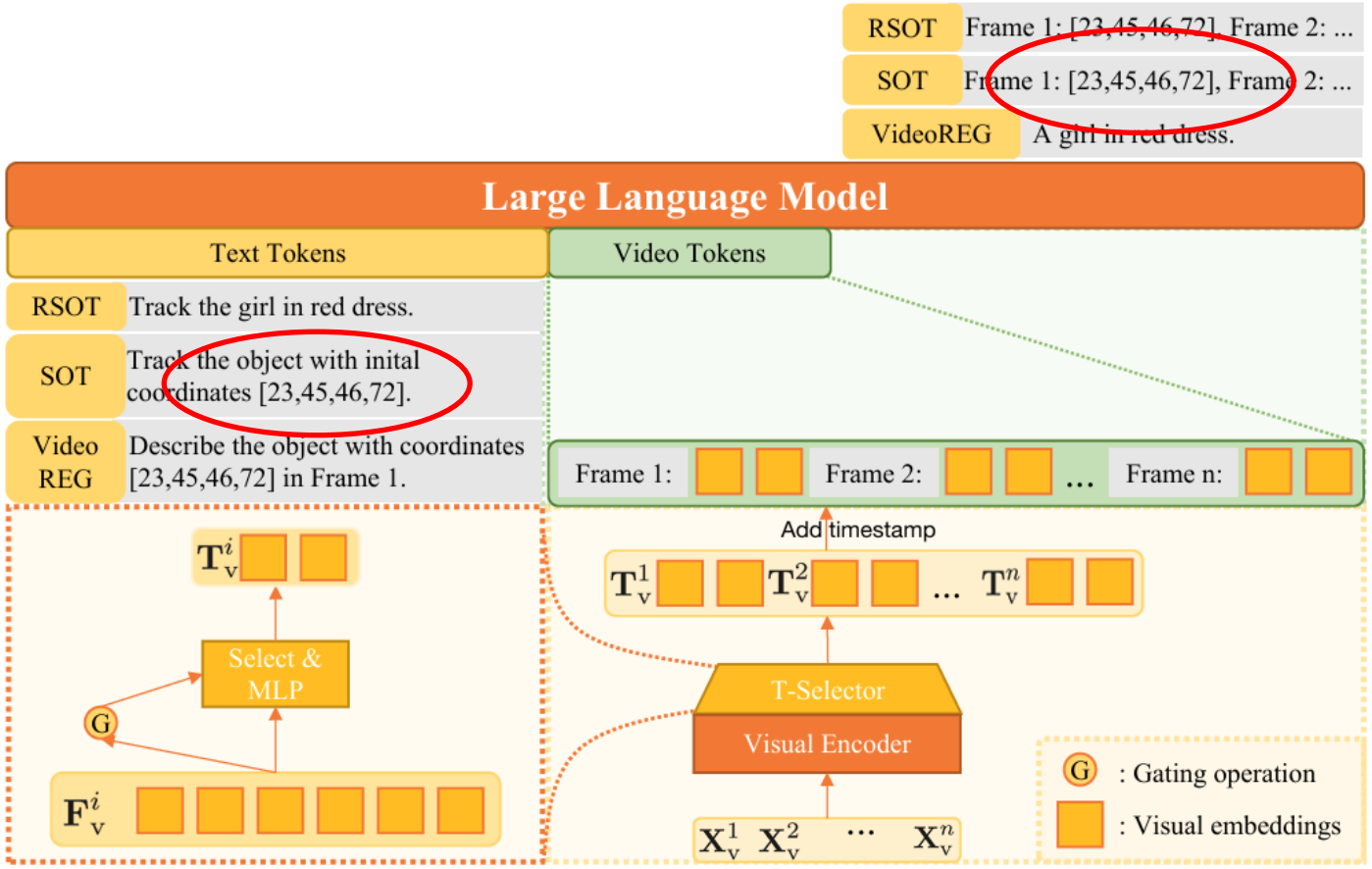


Background: VLMs in region-level understanding

- Paradigms:



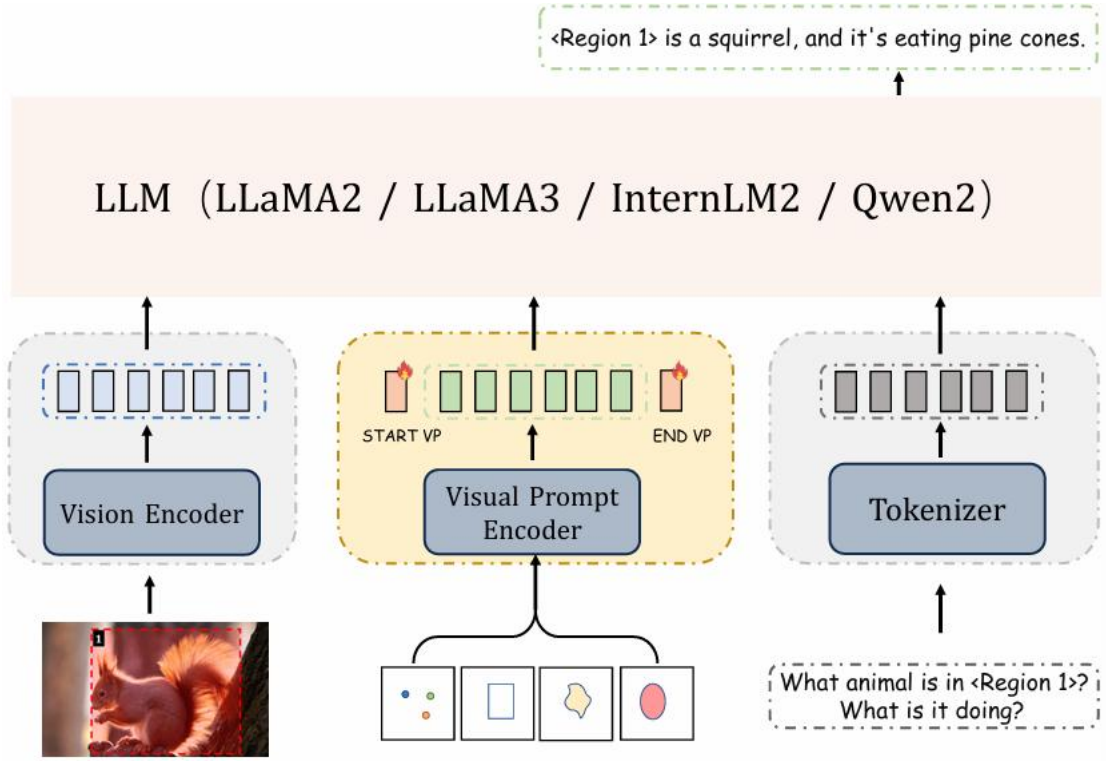
Background: textual method - Elysium

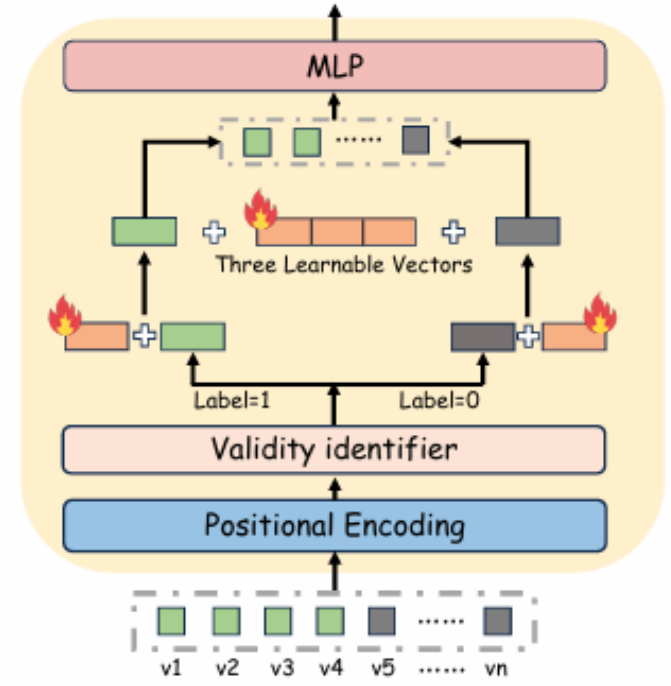


- Encode 2-D bounding-box coordinates as natural-language strings.
- Simple but lack visual semantics.

Elysium: Exploring Object-level Perception in Videos via MLLM, Han Wang, et al., ECCV 2024

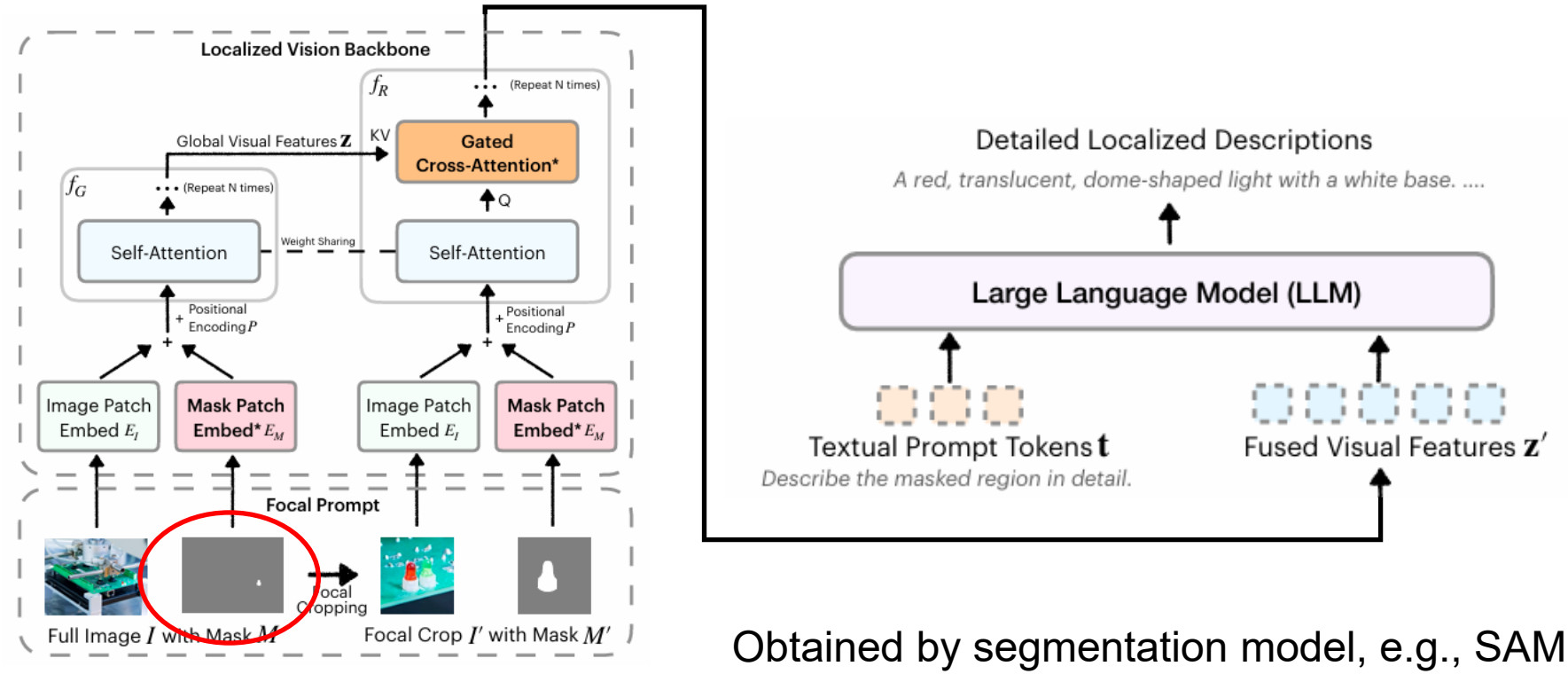
Background: VPE method - VP-MLLM





- Embed regional image features and positional features.
- Lack simultaneous object masks.

Background: RoI-based method - DAM

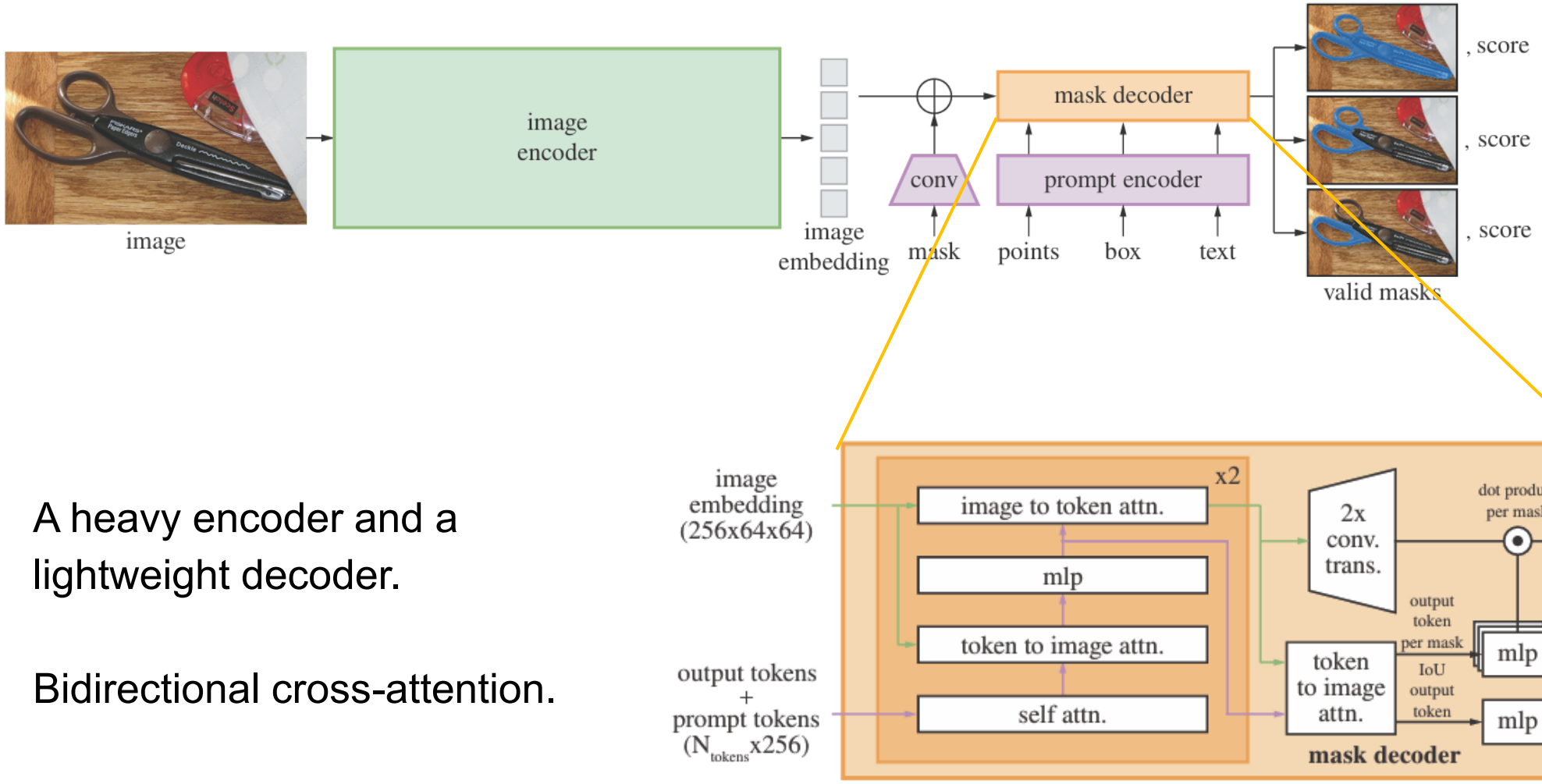


Obtained by segmentation model, e.g., SAM

- Mask is obtained by a segmentation model, e.g., SAM or SAM 2.
- Segmentation model and vision encoder and separated.

Describe Anything: Detailed Localized Image and Video Captioning, Long Lian, et al., arXiv 2025

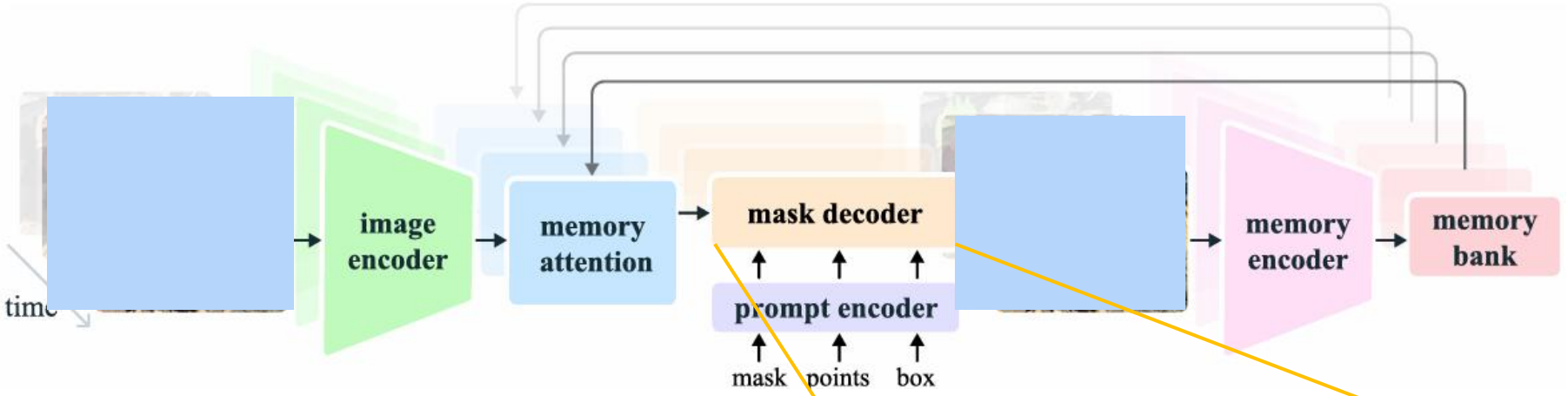
Background: SAM



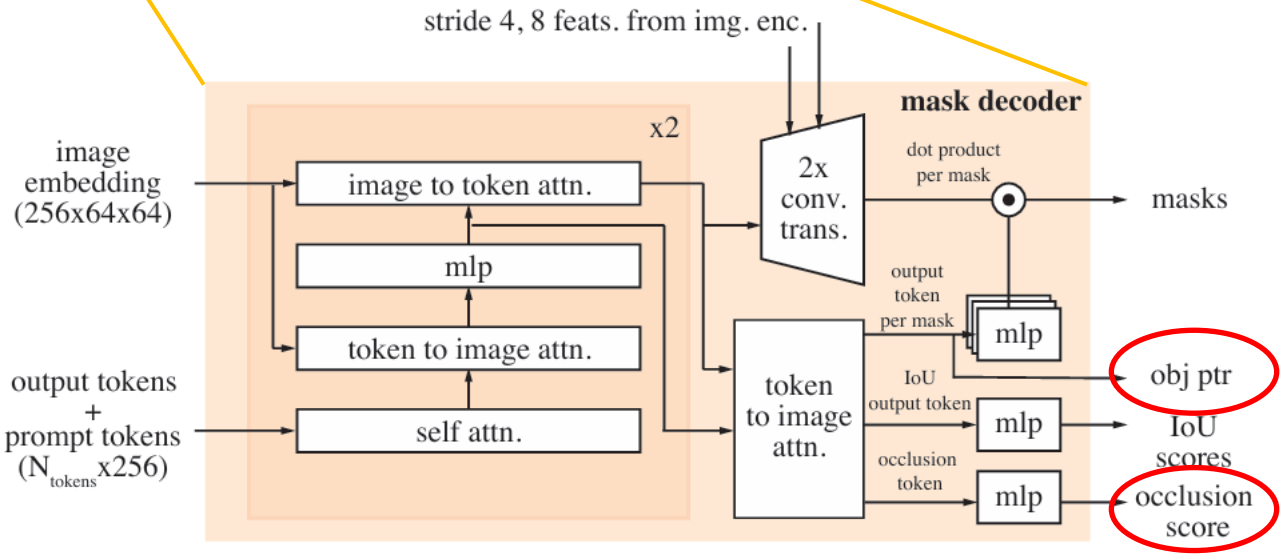
- A heavy encoder and a lightweight decoder.
- Bidirectional cross-attention.

Segment Anything, Alexander Kirillov, et al., ICCV 2023

Background: SAM 2

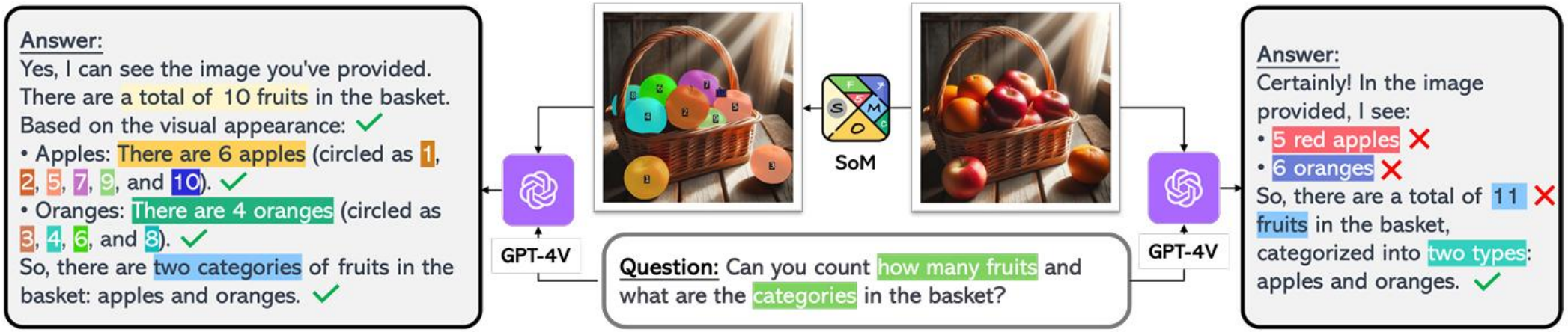


- memory bank to store previous frame and prompt information.
- Object pointer to store high-level semantics.
- Occlusion prediction.



SAM 2: Segment Anything in Images and Videos, Nikhila Ravi, et al., arXiv 2024

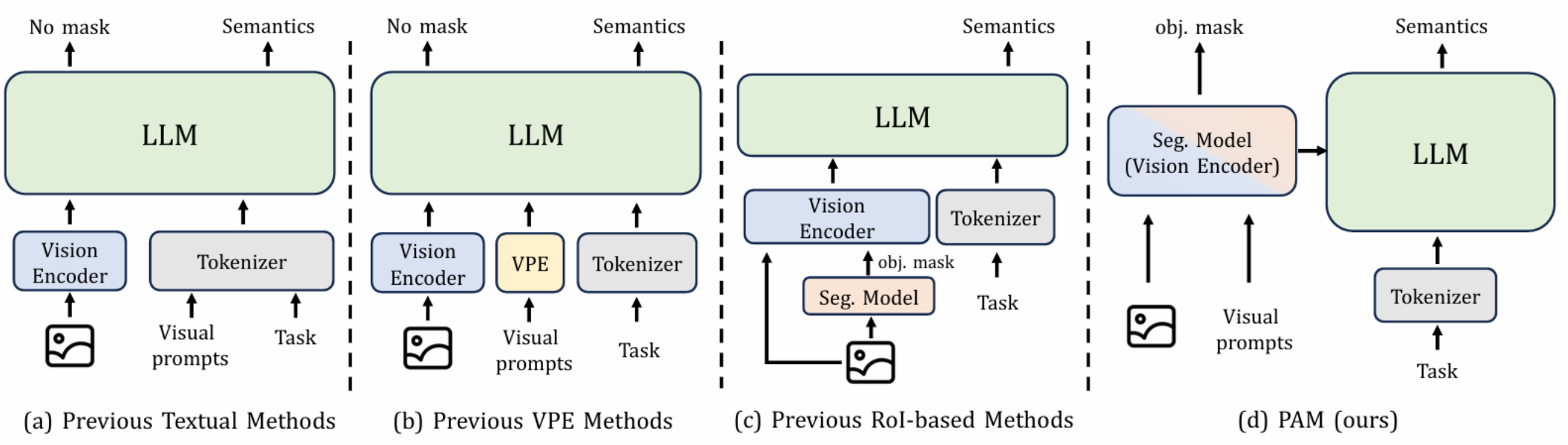
Background: SoM



- Segment and mark the regions in images.
- VLMs like GPT-4V can better ground the regions given these marks.

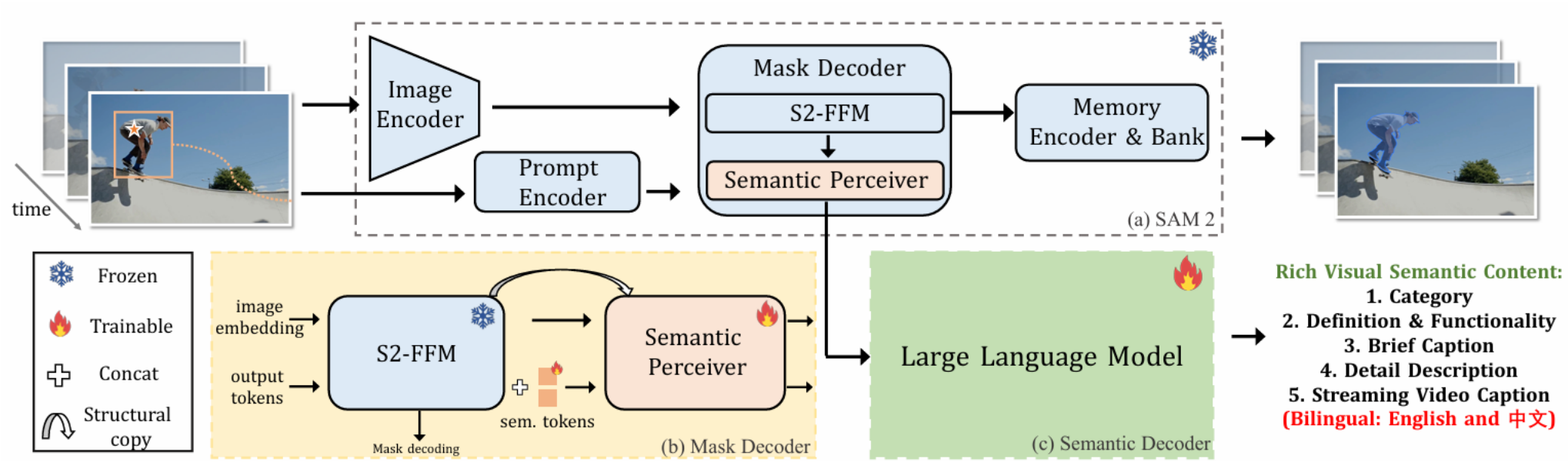
Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V, Nikhila Ravi, et al., arXiv 2023

Method: Overview



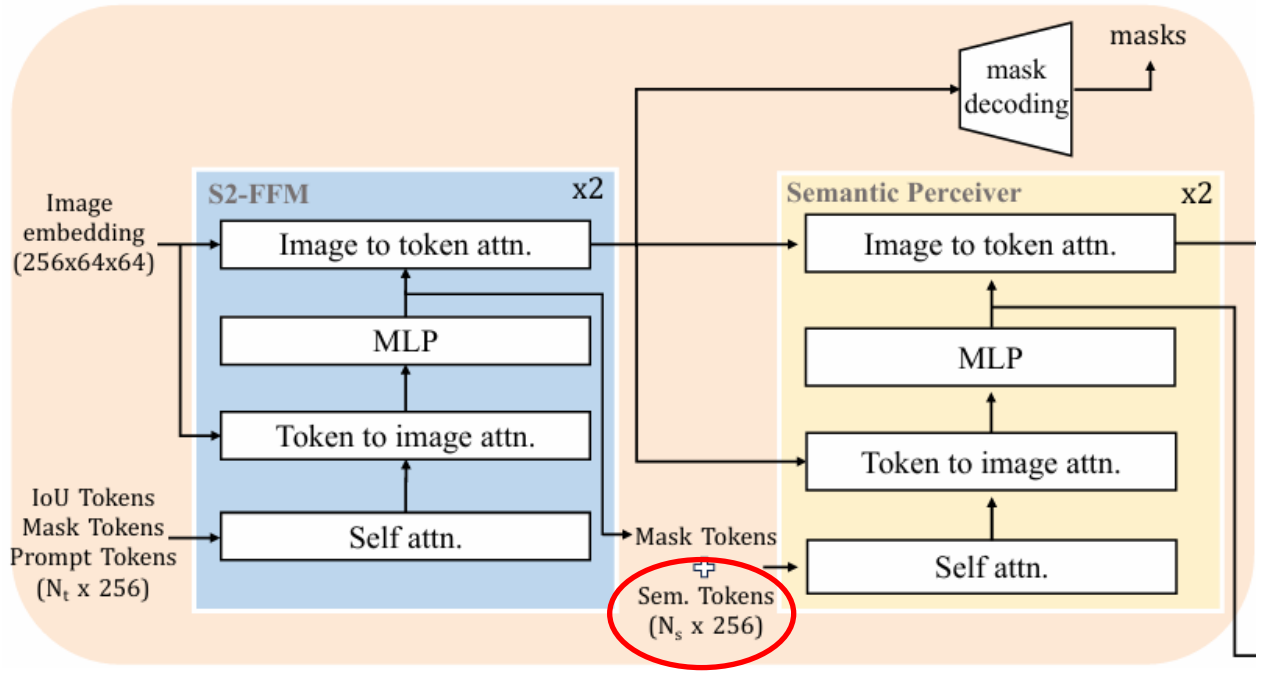
- Obtain object masks and region-level semantics simultaneously.
- Integrate the segmentation model and vision encoder.

Method: Overview



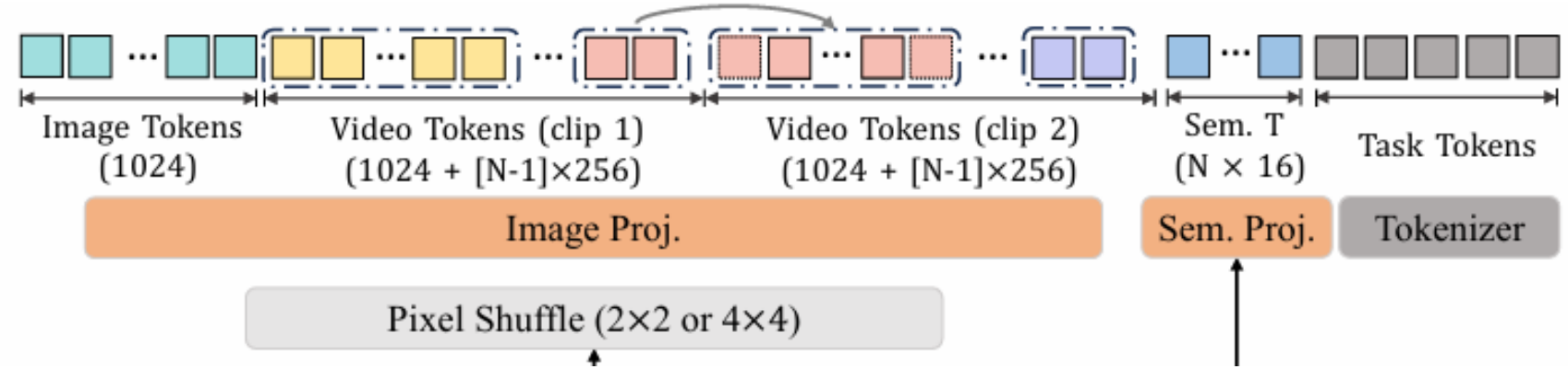
- Semantic Perceiver as a bridge between SAM 2 features and LLM.
- LLM as a semantic decoder to provide visual semantic content.
- Parallel design for mask and semantic decoders.

Method: Semantic Perceiver



- Copy the architecture of SAM 2 mask decoder.
- Add semantic tokens to perceive general visual information better.

Method: Projector



- Pixel shuffle for efficiency
  - 2x2 for images and prompt frames in videos.
  - 4x4 for remaining frames in videos.

- For streaming video processing
  - Additional 2x2 pixel shuffle is applied to the last frame of each clip, and processed together with next clip.
  - Incorporate the previous textual description to augment contextual history.

Method: Training

北京大学 PEKING UNIVERSITY

- **Stage 1**: Image Pretraining and Alignment.
  - Align visual/semantic tokens with LLM's embedding space.
  - Semantic perceiver and projector are trainable.

- **Stage 1.5**: Video-Enhanced Pretraining and Alignment.
  - Extend to videos.
  - Semantic perceiver and projector are trainable.

- **Stage 2**: Multimodal Fine-Tuning.
  - SFT on diverse tasks with proposed dataset.
  - Semantic perceiver, projector and LLM are trainable.

Method: Data Construction - Image

- SoM to identify RoI
- GPT-4o for conceptual explanations and contextual functionalities
- Qwen2.5-VL-72B for detailed descriptions
- Rule-based and manual cleaning

Method: Data Construction - Video

- Storyboard-driven caption expansion
- Trace-Uni to segment events

Experiments: Image Regional Recognition

Model	Classification				OCR	
	LVIS		PACO		COCO Text	Total-Text
	Semantic Sim.	Semantic IoU	Semantic Sim.	Semantic IoU	Acc.(%)	Acc.(%)
Shikra-7B [12]	49.7	19.8	43.6	11.4	–	–
GPT4RoI-7B [84]	51.3	12.0	48.0	12.1	–	–
Osprey-7B [80]	65.2	38.2	73.1	52.7	–	–
Ferret-13B [77]	65.0	37.8	–	–	–	–
VP-LLAVA-8B [41]	86.7	61.5	75.7	50.0	44.8	46.9
VP-SPHINX-13B [41]	87.1	62.9	76.8	51.3	**45.4**	48.8
DAM-8B [38]	**89.0**	77.7	84.2	73.2	–	–
PAM-1.5B (Ours)	87.4	76.5	85.1	73.5	39.4	48.6
PAM-3B (Ours)	88.6	**78.3**	**87.4**	**74.9**	42.2	**52.3**

- Semantic Similarity measures the similarity of predicted/GT labels in a semantics space.
- Semantic IoU reflects the overlap of words.

Experiments: Image Regional Caption

Model	VG		Refcocog		Ref-L4			Ferret Bench	MDVP Bench
	METEOR	CIDEr	METEOR	CIDEr	ROUGE-L	METEOR	CIDEr	Refer. Desc.	Avg.
GLaMM-7B [51]	17.0	127.0	15.7	104.0	23.8	10.1	51.1	-	-
Osprey-7B [80]	-	-	16.6	108.3	-	-	-	72.2	44.3
Ferret-7B [77]	-	-	-	-	22.3	10.7	39.7	68.7	47.6
VP-LLaVA-8B [41]	-	-	22.4	153.6	-	-	-	75.2	70.6
VP-SPHINX-13B [41]	20.6	141.8	23.9	**162.5**	22.6	10.7	32.4	77.4	**74.3**
Omni-RGPT-7B [25]	17.0	139.3	17.0	109.7	-	-	-	-	-
RegionGPT-7B [21]	17.0	145.6	16.9	109.9	25.3	12.2	42.0	-	-
DAM-3B [38]	-	-	-	-	30.8	17.2	56.4	-	-
DAM-8B [38]	-	-	-	-	**37.1**	**19.4**	**70.0**	-	-
PAM-1.5B (Ours)	19.2	132.9	24.7	135.0	26.6	14.9	49.8	72.4	68.4
PAM-3B (Ours)	**20.8**	**142.3**	**26.9**	143.1	31.3	17.2	59.7	**77.5**	72.2

- METEOR, ROUGE-L, CIDEr measure caption quality.
- Ferret/MDVP Bench uses GPT-4o for fine-grained quality assessment (Referring Description here).

Experiments: Video Regional Caption

Model	Elysium	BensMOT	HC-STVG		VideoRefer-Bench-D				
	METEOR		METEOR	CIDEr	SC	AD	TD	HD	Avg.
Elysium-7B [63]	19.1	1.1	–	–	2.35	0.30	0.02	3.59	1.57
Merlin-7B [78]	–	–	11.3	10.5	–	–	–	–	–
Omni-RGPT-7B [25]	9.3	14.6	–	–	–	–	–	–	–
Artemis-7B [49]	–	–	18.0	53.2	3.42	1.34	1.39	2.90	2.26
VideoRefer-7B [81]	–	–	18.7	68.6	4.44	3.27	3.10	3.04	3.46
DAM-3B [38]	–	–	18.2	72.7	3.62	2.86	2.81	2.67	2.99
DAM-8B [38]	–	–	21.0	**91.0**	**4.69**	**3.61**	**3.34**	**3.09**	**3.68**
PAM-1.5B (ours)	20.7	19.1	18.8	58.9	3.63	2.71	2.67	2.89	2.97
PAM-3B (ours)	**24.3**	**21.6**	**23.3**	70.3	3.92	2.84	2.88	2.94	3.14

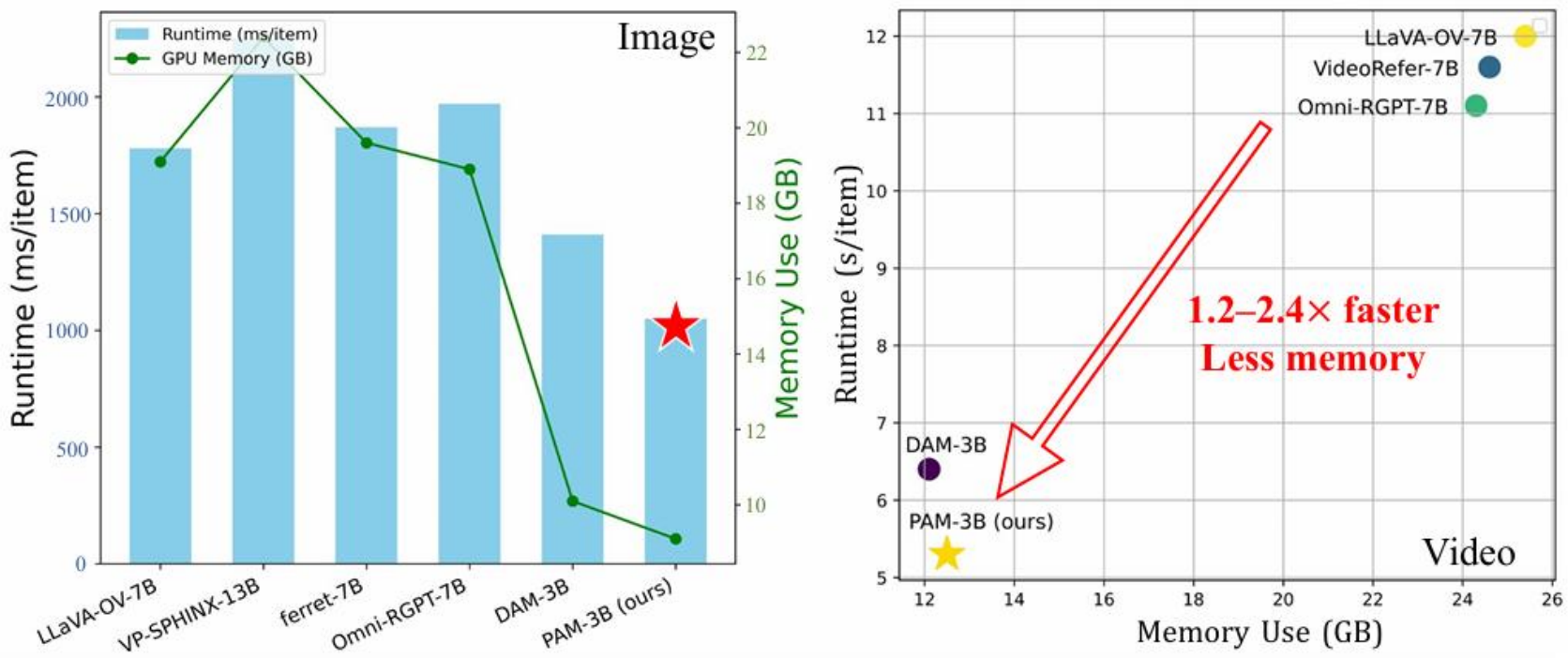
- Subject Consistency (SC) ensures descriptions strictly match the target object.
- Appearance Description (AD) scores the accuracy of visual attributes in descriptions.
- Temporal Description (TD) assesses the correctness of dynamic object states over time.
- Hallucination Detection (HD) identifies fabricated or wrong details in descriptions.

Experiments: Streaming Video Regional Caption

Model	ActivityNet		
	CIDEr	METEOR	G-STDC
VideoRefer-7B [81]	22.1	14.7	1.73
DAM-3B [38]	11.3	14.8	0.94
GIT* [65]	29.8	7.8	–
Vid2Seq* [73]	30.2	8.5	–
Streaming Vid2Seq* [73]	**37.8**	10.0	–
PAM-1.5B (ours)	27.3	22.6	2.15
PAM-3B (ours)	30.1	**27.3**	**2.67**

- GPT-4o-evaluated Spatio Temporal Description Continuity Score (G-STDC) assesses the continuity and entity consistency of descriptions for sequential events.

Experiments: Efficiency



Image

Video

Experiments: Qualitative Results

Label

Face Mask

口罩

Definition & Functionality

A face mask is a **loose-fitting disposable** cover that protects the wearer's **mouth and nose** from the surrounding environment. In this image, the person is likely wearing it for reducing the **spread of droplets**.

口罩是一种用于遮盖口鼻的一次性防护用品。在这张图片里，这位正在练习书法的人佩戴口罩，很可能是出于卫生目的，以减少飞沫的传播。

Caption

The **blue** mask is worn by the person as he concentrate on the art work. It offers a layer of **protection**, ensuring that the person can immerse themselves in the detailed **brushwork**.

这是一个蓝色的口罩，练习书法的人正戴着这个口罩，专注于他的艺术创作。这层口罩提供了保护，让他能够安心地沉浸在对笔墨细节的精心处理之中。

Experiments: Qualitative Results

A man in a white t-shirt and black shorts **stands** as the central figure on the court. Initially, he **holds a basketball**, he then **transitions** into dynamic movement.

一位身穿白色T恤和黑色短裤的男子站在球场上。一开始，他手里拿着篮球，随后开始进行运球动作。

He is **now seen** actively practicing **dribbling**. He **started from** the baseline of the field to the center of the field and then **back into** the penalty area.

现在，他正投入地练习运球。他从球场的右底线开始，移动到中场区域，接着回到罚球区里面。

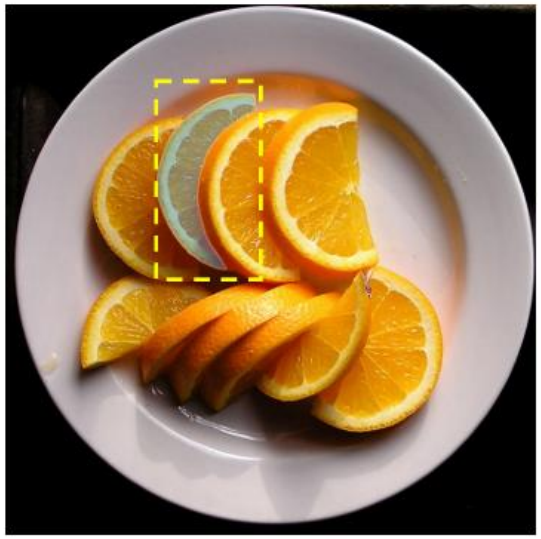
The man is **now positioned** on the court. He **starts** by holding the basketball, **then** he brings the ball up, and shoots it towards the hoop.

他现在在空旷的球场上，面前有一个篮筐。他拿着篮球，接做出投篮动作，球被投入筐中。

Experiments: Qualitative Results

北京大学 PEKING UNIVERSITY

Experiments: Failure Cases



The orange slice, positioned third from the left, has translucent, juicy flesh. Its rind is a soft, light yellow. It rests closely against the other slices on the white plate, offering a gentle contrast in color.

This is a small, spherical container holding coffee, and it glows with a metallic rose-red color. Its glossy surface brightly reflects the light. The letters 'E' and 'L' are delicately inscribed on its polished roundness, possibly indicating it is handmade.

The circular sign is red and white. Its large, round surface is primarily red, making the white lettering stand out clearly. The central letters are 'Tcct', with other white text curving around the sign's edge. The sign juts out from the building on a sturdy bracket, acting as a clear marker above the sidewalk.

- Expected to be mitigated by broader data coverage and reinforcement training.

Conclusion - Contributions

北京大学 PEKING UNIVERSITY

- Efficiently extending SAM 2 with a Semantic Perceiver allows direct bridging of its rich visual features to an LLM for semantic understanding.
- A parallel decoding design for mask and semantic generation enhances computational efficiency.
- The development of a high-quality, multi-granularity dataset, including pioneering region-level streaming video captions.

Conclusion - Limitations

- Limited capability for general understanding tasks, e.g., VQA.
- Its real-time streaming video region captioning capability is still hindered by the excessive number of visual tokens.
- Trade-off between efficiency and granularity when processing long videos.

Thanks for listening!

Experiments: Ablation

Method	LVIS (S.IoU)	RefCOCOg (METEOR)	HC-STVG (METEOR)	time (ms/it)
+ 4	78.9	26.1	22.5	972
+ 16	79.6	26.9	23.3	980
+ 64	**80.0**	**27.0**	**23.5**	1143
+ w/o(0)	77.6	24.6	21.3	**967**

- Number of semantic tokens.

Method	LVIS (S.IoU)	RefCOCOg (METEOR)	HC-STVG (METEOR)
All in one	78.7	25.8	21.6
S1→2	**79.7**	26.7	22.4
S1→1.5 →2	79.6	**26.9**	**23.3**

- Training stages.

Method	LVIS (S.IoU)	RefCOCOg (METEOR)	HC-STVG (METEOR)
I.E. pre S2-FFM	78.4	25.0	21.9
I.E. after S2-FFM	**79.6**	**26.9**	**23.3**
all T. + sem.T	**79.9**	26.8	23.3
mask T. + sem.T	79.6	**26.9**	23.3

- Features from SAM 2.