Osprey

Pixel Understanding with Visual Instruction Tuning

Yuqian Yuan*, Wentong Li*, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin,

Lei Zhang, Jianke Zhu

*Equal Contribution

https://github.com/CircleRadon/Osprey

Content

Motivation

Object Category: person

Part Taxonomy: body

Attribute: color, position …

Caption: region short / detailed description

SAM ”Segment Everything” Predictions

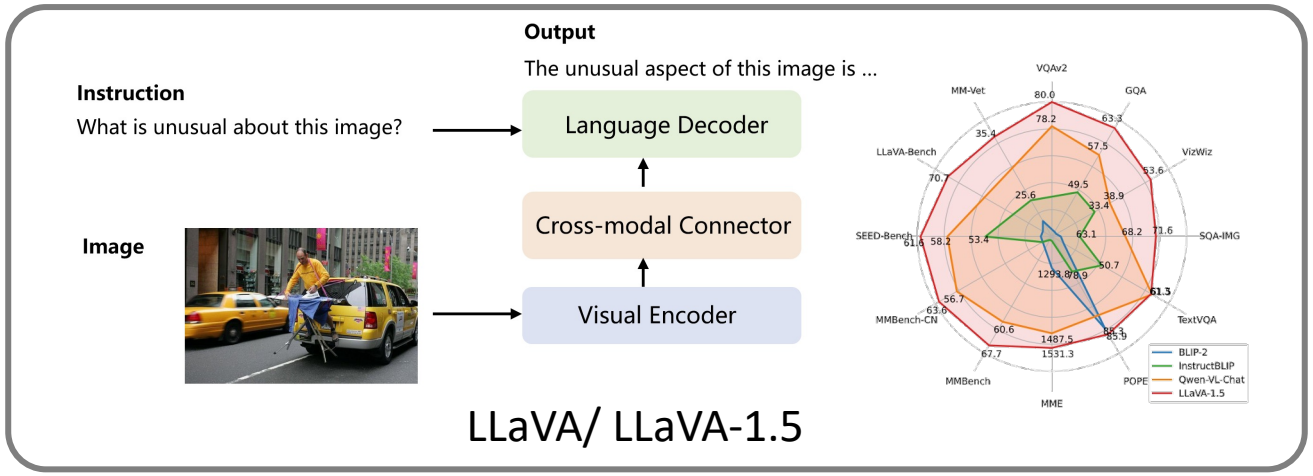
☹ **No semantic information**

Fine-grained Region/Pixel Understanding

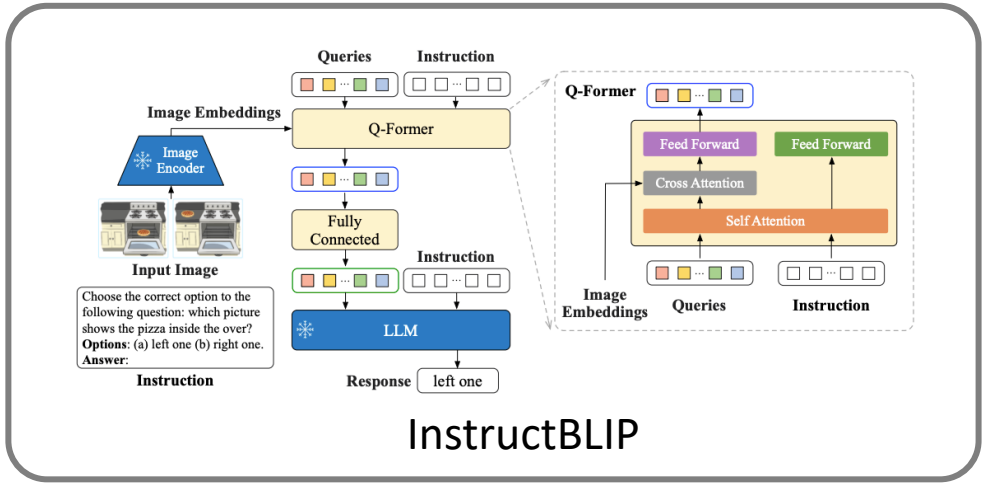
☺ **Rich semantic information containing different granularities**

Background — Multimodal Large Language Models

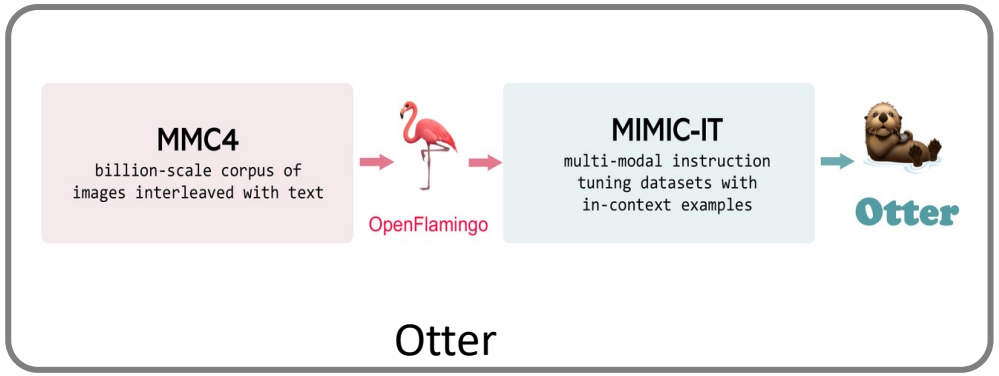
Image-level understanding



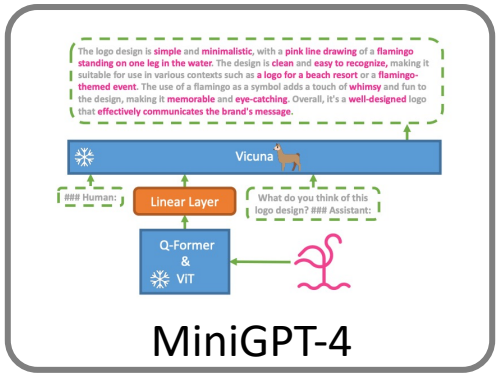
LLaVA/ LLaVA-1.5



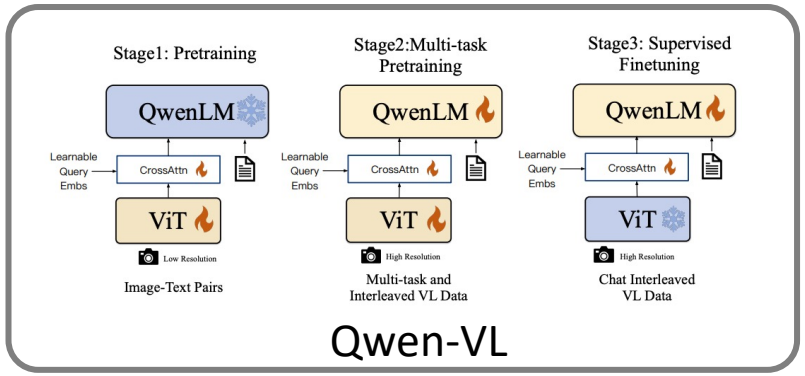
InstructBLIP



Otter



MiniGPT-4

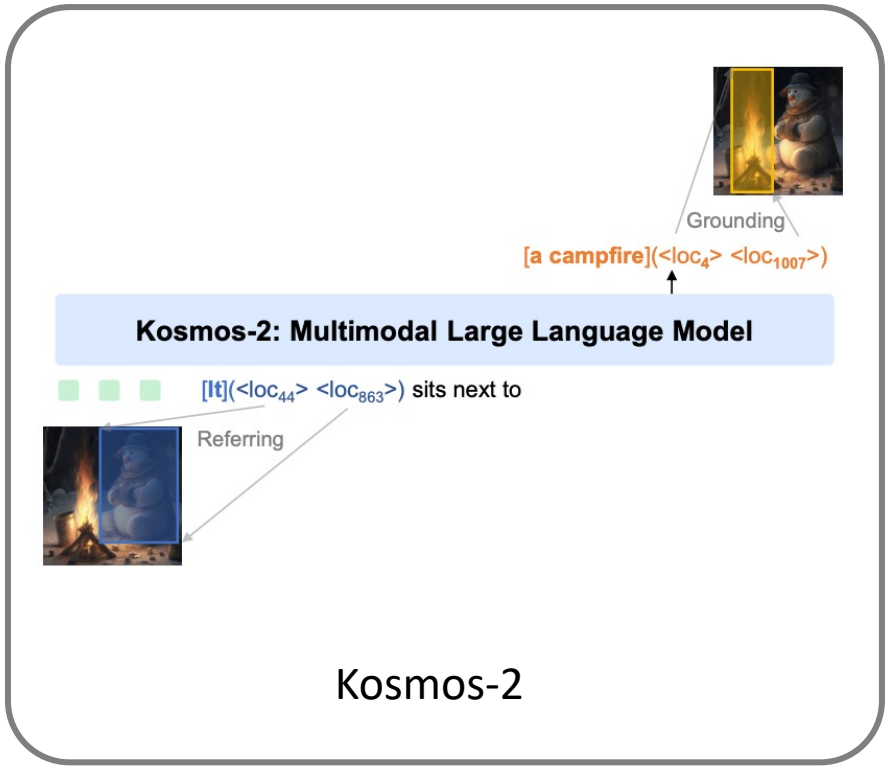


Qwen-VL

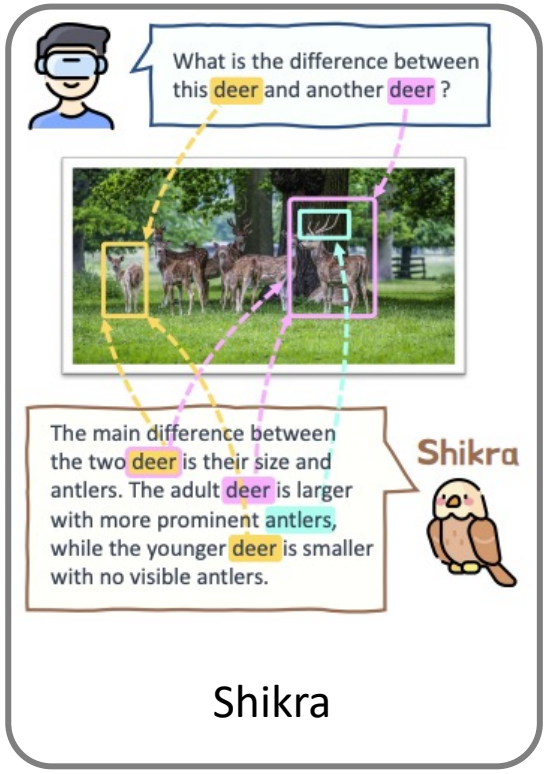
[Liu et al. 2023] Visual instruction tuning.
[Liu et al. 2023] Improved baselines with visual instruction tuning.
[Dai et al. 2023] Instructblip: Towards general-purpose visionlanguage models with instruction tuning.
[Li et al. 2023] Otter: A multi-modal model with in-context instruction tuning.
[Zhu et al. 2023] Minigpt-4: Enhancing vision-language understanding with advanced large language models.
[Bai et al. 2023] Qwen-vl: A frontier large vision-language model with versatile abilities.

Background — Multimodal Large Language Models

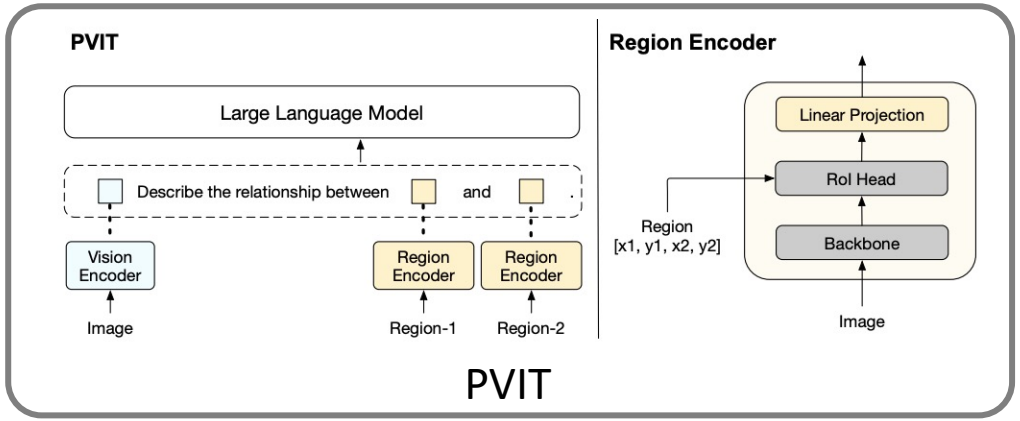
Region-level understanding



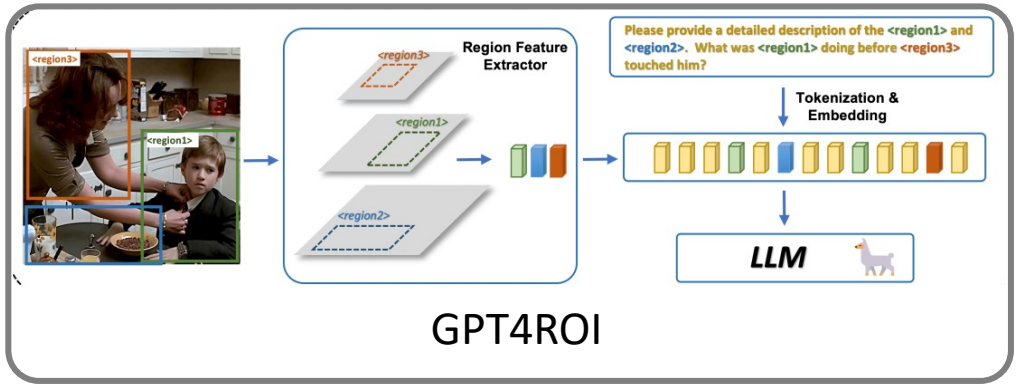
Kosmos-2



Shikra



PVIT



GPT4ROI

[Peng et al. 2023] Kosmos-2: Grounding multimodal large language models to the world.
[Chen et al. 2023] Shikra: Unleashing multimodal llm's referential dialogue magic.
[Chen et al. 2023] Position-enhanced visual instruction tuning for multimodal large language models.
[Zhang et al. 2023] Gpt4roi: Instruction tuning large language model on region-of-interest.

Comparisons with box-level referring

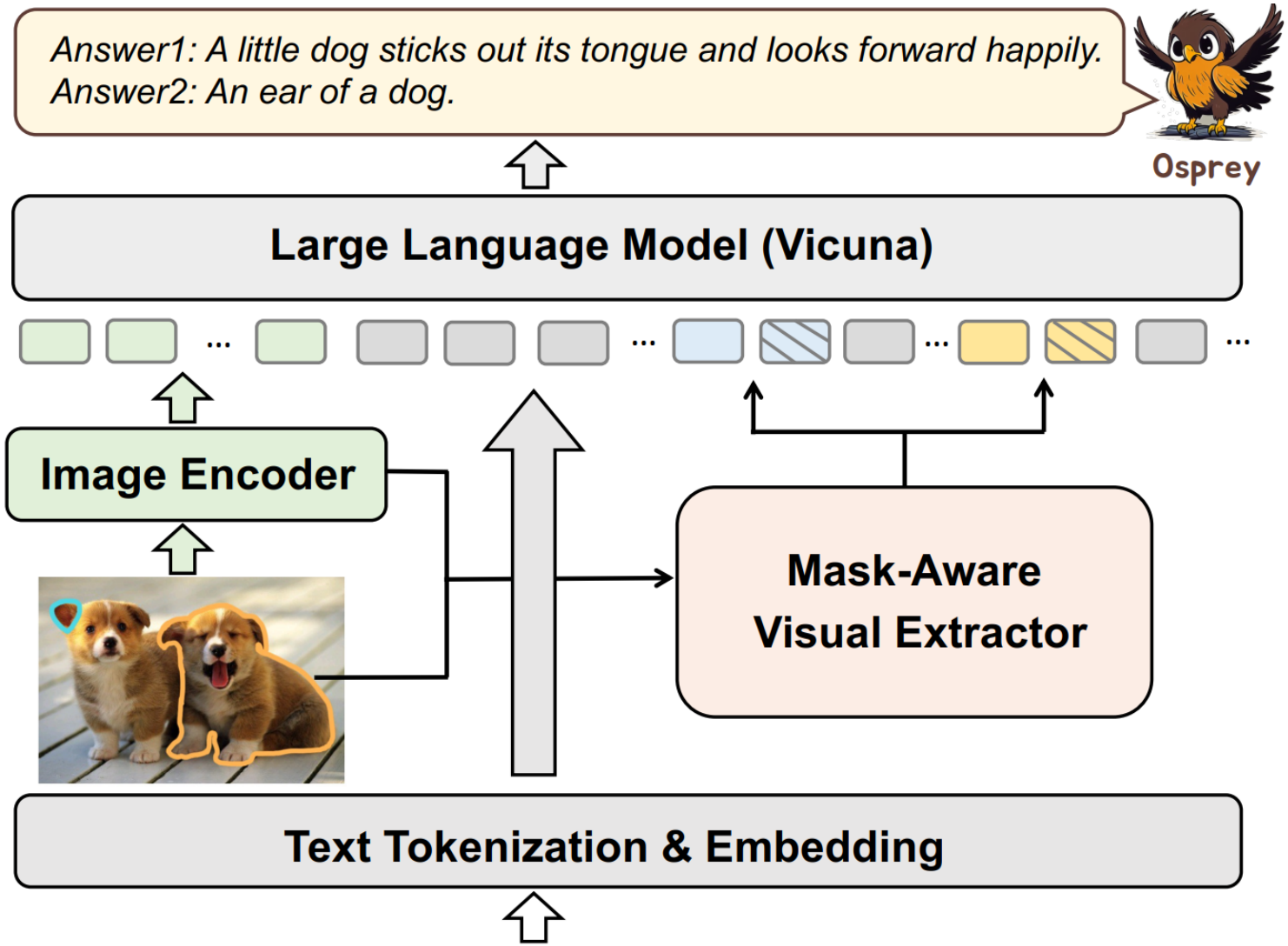
Box-level referring

- involve **irrelevant background features** and lead to **inexact region-text pair alignment** for visual instruction tuning on LLM.
- not be able to precisely indicate the object, resulting in **semantic deviation**.

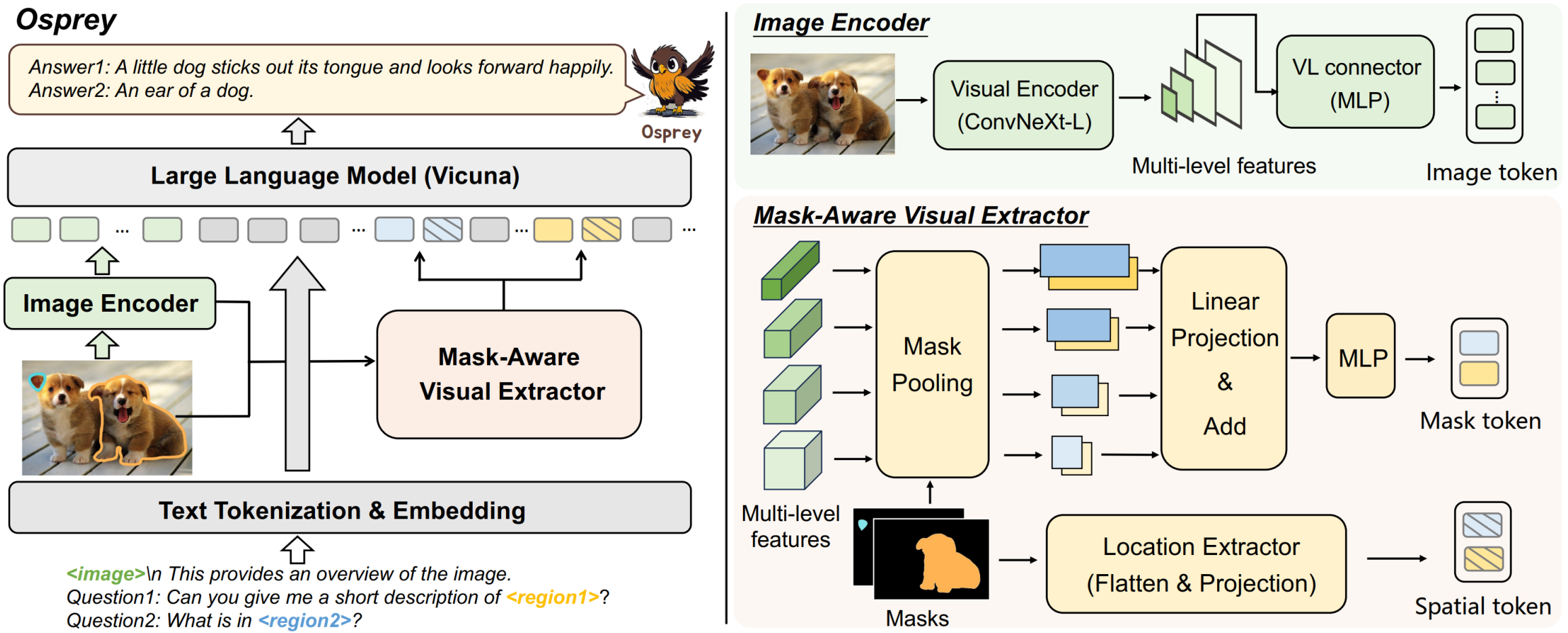
Ours (Mask-level referring)

- Precise regional representation.
- Higher resolution for input images.
- Understand regions with various granularity.

Architecture



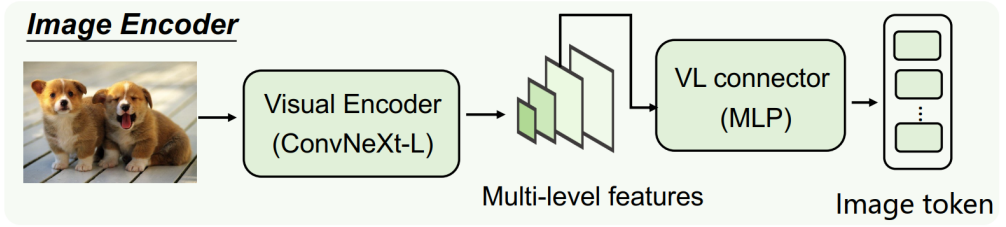
Architecture



Method

Convolutional CLIP Vision Encoder

- **ConvNeXt-Large** CLIP model
- **512x512** input image size
- Multi-level image features
- Adopt the **“res4”** stage as the image-level features

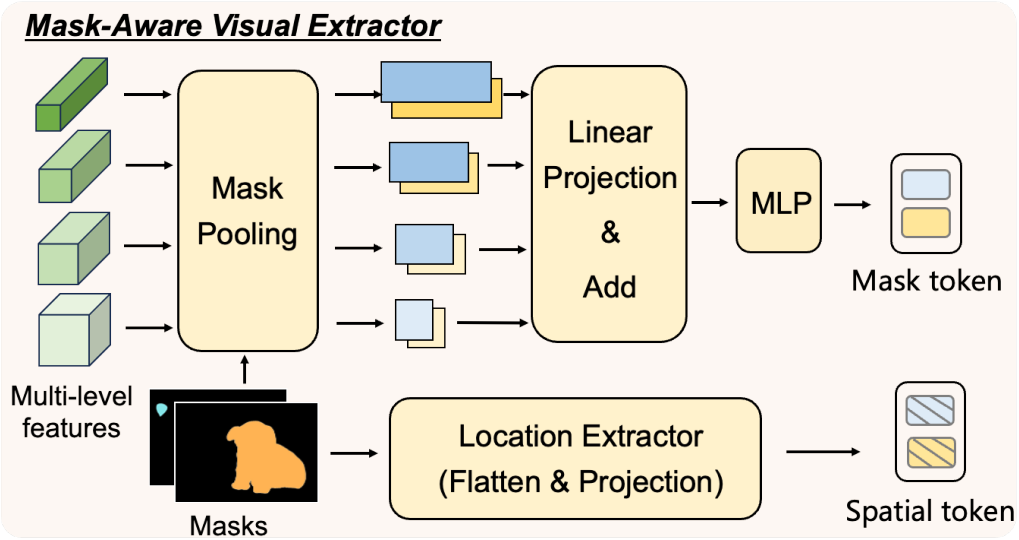


CLIP Vision Encoder	224	448	672	896	1120
ViT-Surgery-L [27]	**26.52**	28.15	27.26	25.18	24.61
ConvNeXt-L [31]	23.35	**34.36**	**40.57**	**43.04**	**43.33**

Table 6. Panoptic segmentation comparisons (PQ) using different vision encoders with different input sizes on ADE20K-150 [54]. The ground truth mask is used for recognition evaluation.

Mask-Aware Visual Exactor

- **Mask Token**
  - Mask Pooling: $V_{ij} = \mathcal{MP}(\mathbf{R}_i, \mathbf{Z}(x)_j)$
  - $t_i = \sigma(\sum_{j=1}^{4} \mathbf{P}_j(V_{ij}))$
- **Spatial Token**
  - Resize 224, flatten and project to s_i



Tokenization for LLM Model

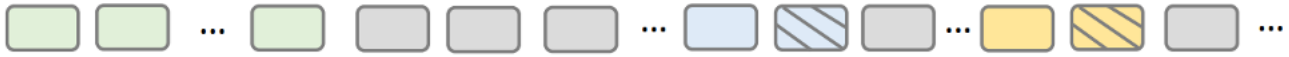
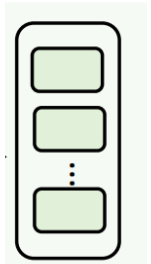
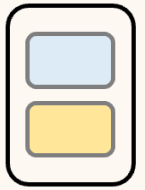


Image: Special token **<image>** → Image-level embeddings from the output of **ConvNeXt-Large CLIP** model.

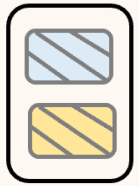


Text: Using the pre-trained LLM's tokenizer.

Mask-based region: Special token <region>: <mask><position>.



Mask token



Spatial token

Prefix prompt: "**<image>** \n This provides an overview of the picture."

How to create mask-text instruction data?

Current Data

COCO、RefCOCO、RefCOCOg、RefCOCO+ …

- Segmentation & Detection
- Categories
  - person
- Mask region captions
  - the lady with the blue shirt.
  - the back of an older woman with her hair in a barrette with a blue jacket on.
  - navy blue shirt.
  - woman back in blue.
  - a woman is wearing blue sweater.

Expected data

Question: Can you give me a detailed description of **<region>**?
Answer: The woman in **<region>** is wearing a gray shirt, jeans, and glasses. She has short hair and a watch on her wrist. She is shopping in the lively market, browsing through the fresh produce, such as fruits and bags.

Question: What color is the woman's sweater in **<region>**?
Answer: Blue.

Question: Please describe the atmosphere of the market.
Answer: The market has a bustling and lively atmosphere. There's a sense of excitement and energy as people gather around the fresh fruits and bags for sale. The variety of items available, as well as the interactions among customers, contribute to the vibrant ambience of the market scene.

Mask-text instruction data creation

Use strong LLM-based teacher GPT-4

Image-level description (LLaVA-115K)

The image presents a lively market scene with a group of people buying fruits and bags. There are multiple individuals in the market, all browsing through the fresh produce available…

Categories & bboxes (COCO)

person: [0.507,0.409,0.698,0.740], person: [0.243,0.496,0.558,0.746], handbag: [0.473,0.110,0.607,0.201], orange: [0.761,0.537,0.820,0.569], …

Mask region captions (RefCOCO、 RefCOCO+、 RefCOCOg)

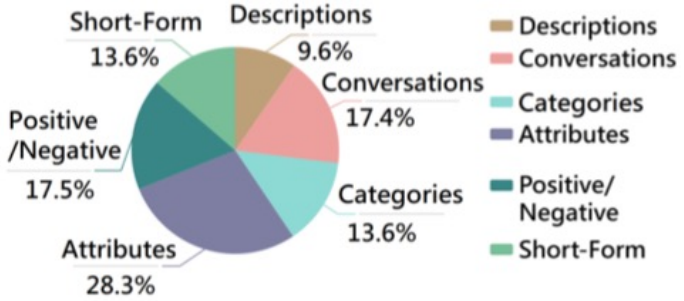
<region1>: gray shirt wearing glasses.
<region2>: the lady with the blue shirt.
<region3>: a woman in glasses shops in an open air fruit market.

different prompts ↓

GPT-4 → different types of responses

Overview of Osprey-724K dataset

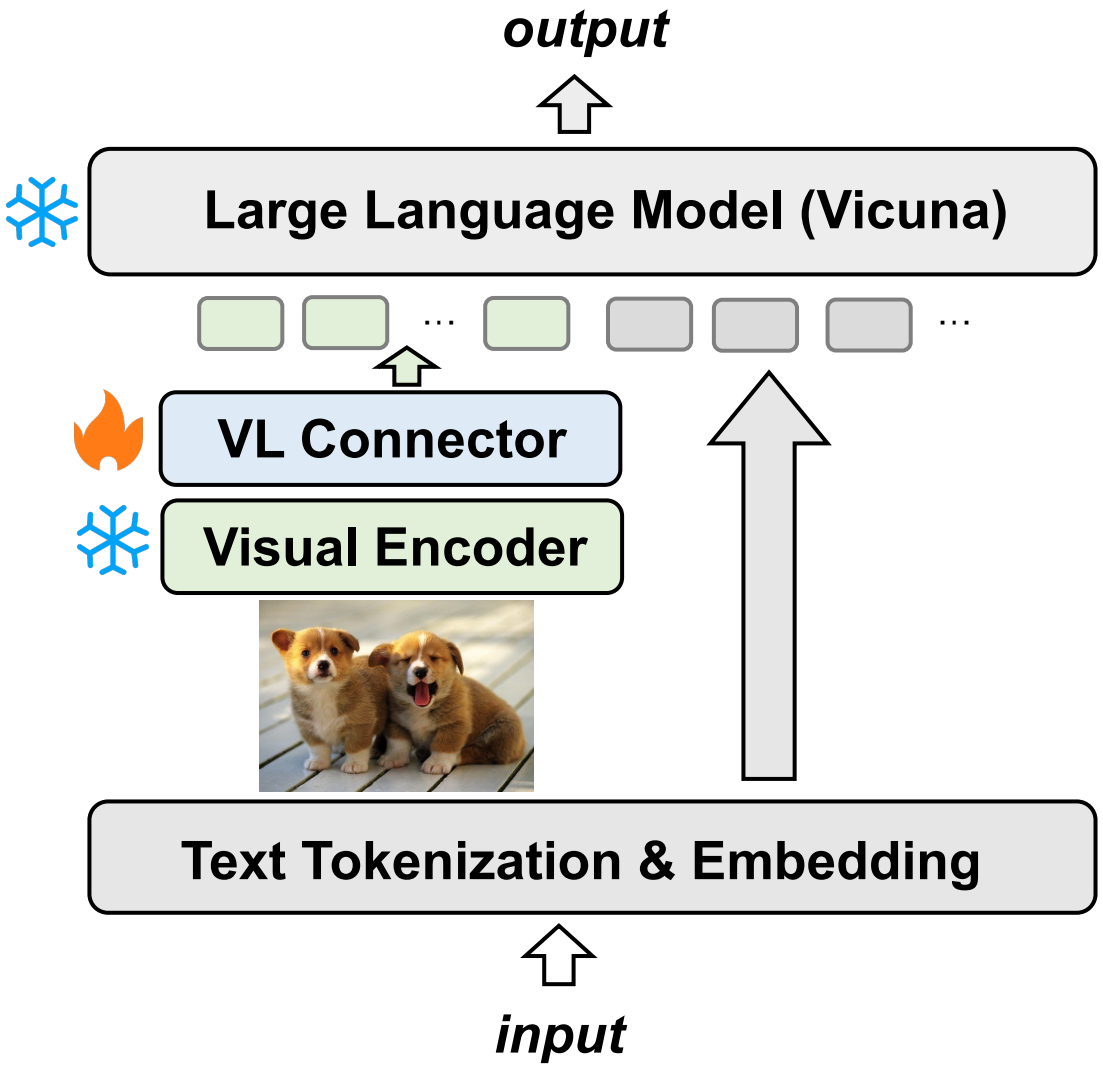
Type	Form	Raw Data	GPT-4	#Samples
Object-level	Descriptions	COCO/RefCOCO/RefCOCO+/ RefCOCOg/LLaVA-115K	✓	70K
	Conversations		✓	127K
Part-level	Categories	PACO-LVIS	✓	99K
	Attributes		✓	207K
Robustness &Flexibility	Positive/Negative	COCO/RefCOCO/RefCOCO+/ RefCOCOg/LLaVA-115K/LVIS	✗	64K/64K
	Short-Form		✓	99k



Training Stage 1: Image-Text Alignment Pre-training

Target: Train the image-level feature and language **connector** for **image-text feature alignment**.

Data: a filtered **CC3M subset (595K)** introduced in LLaVA[1].



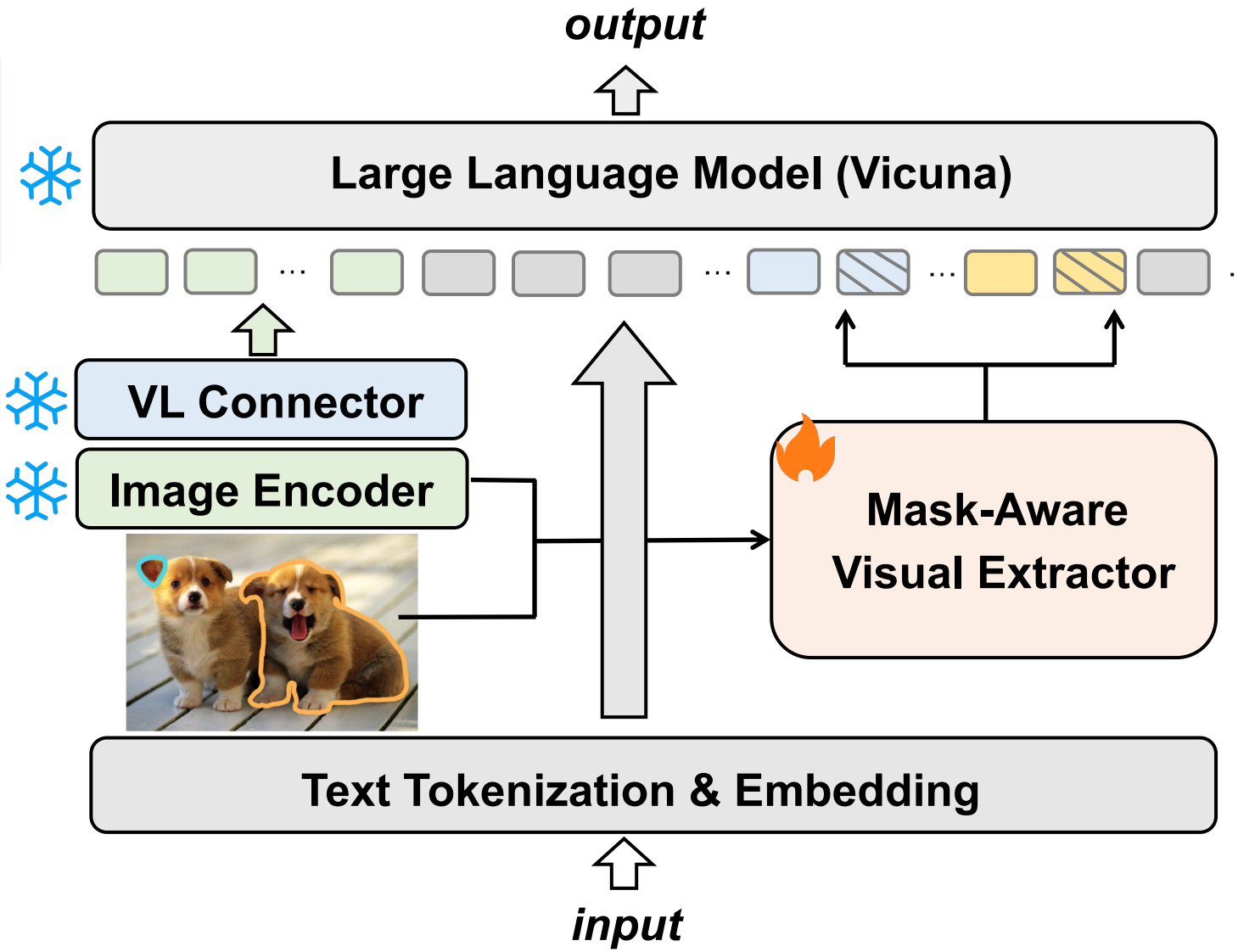
[1] [Liu et al. 2023] Visual instruction tuning.

Training Stage 2: Mask-Text Alignment Pre-training

Target: Align mask-based region features with language embeddings.

Data:

- **Object-level datasets** (COCO, RefCOCO, RefCOCO+)
- **Part-level datasets** (Pascal Part, Part Imagenet)

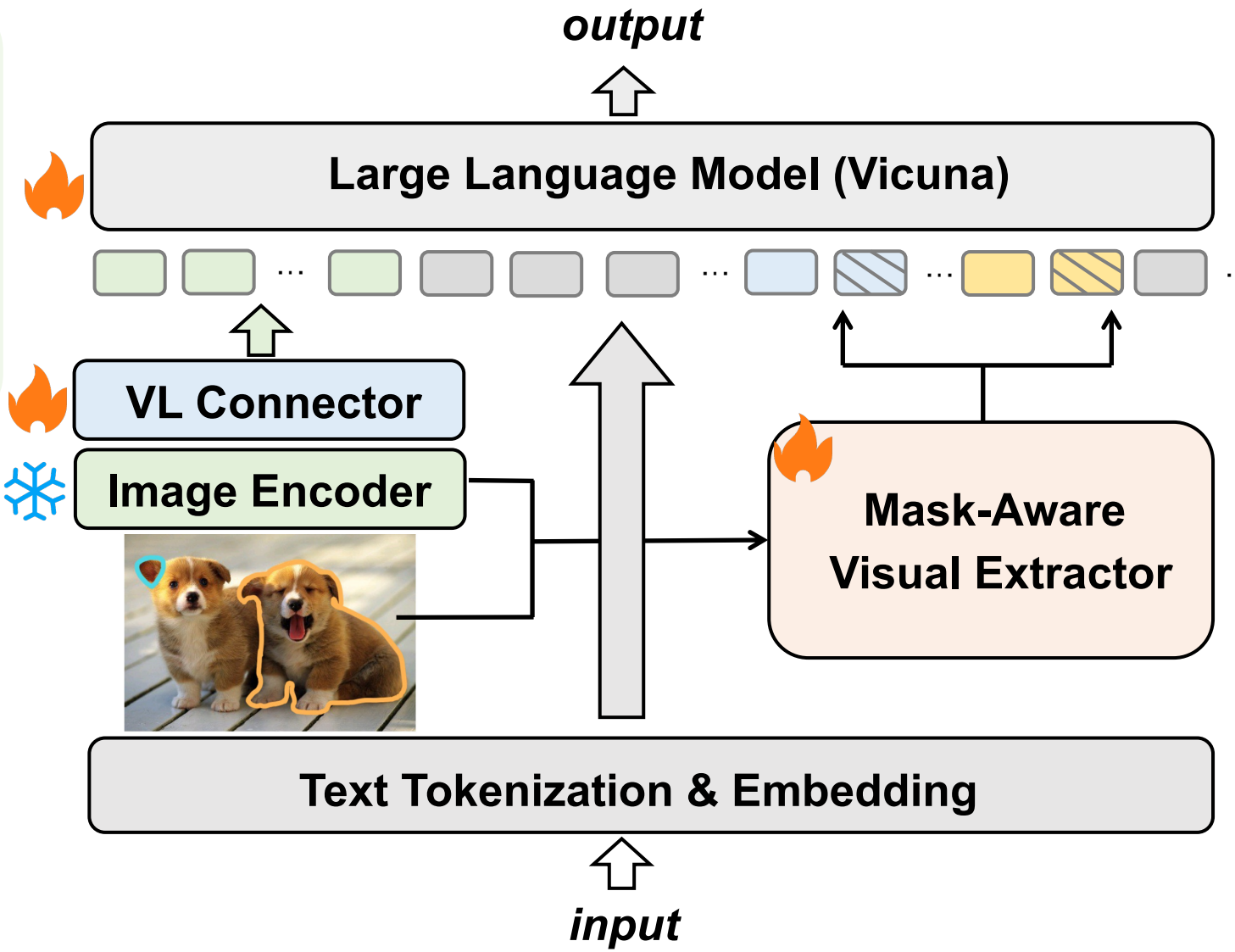


Training Stage 3: End-to-End Fine-tuning

Target: Extending the capability of Osprey to **follow user instructions** and tackle complex pixel-level region understanding tasks.

Data:

- Osprey-724K
- Visual Genome (VG)
- Visual Commonsense Reasoning (VCR)



Experiments

Strong Region Understanding ability

- Open-vocabulary Segmentation

Method	Type	Cityscapes			ADE20K-150		
		PQ	AP	mIoU	PQ	AP	mIoU
CLIP-ConvNeXt-L [40]	Mask	22.53	12.07	23.06	36.86	**39.38**	28.74
CLIP-Surgery-ViT-L [27]	Mask	27.24	28.35	21.92	26.55	29.70	21.42
Kosmos-2 [37]	Box	12.09	9.81	13.71	6.53	4.33	5.40
Shikra-7B [5]	Box	17.80	11.53	17.77	27.52	20.35	18.24
GPT4RoI-7B [53]	Box	34.70	21.93	36.73	36.32	26.08	25.82
Ferret-7B [49]	Mask	35.57	26.94	38.40	39.46	29.93	**31.77**
Osprey-7B (Ours)	Mask	**50.64**	**29.17**	**49.78**	**42.50**	31.72	29.94

- Referring object classification

Method	LVIS		PACO	
	Semantic Similarity	Semantic IoU	Semantic Similarity	Semantic IoU
LLaVA-1.5 [29]	48.95	19.81	42.20	14.56
Kosmos-2 [37]	38.95	8.67	32.09	4.79
Shikra-7B [5]	49.65	19.82	43.64	11.42
GPT4RoI-7B [53]	51.32	11.99	48.04	12.08
Ferret-7B [49]	63.78	36.57	58.68	25.96
Osprey-7B (Ours)	**65.24**	**38.19**	**73.06**	**52.72**

- Detailed region description

Method	Detailed Description
LLaVA-1.5 [29]	71.11
Kosmos-2[37]	40.89
Shikra-7B [5]	40.97
GPT4RoI-7B [53]	49.97
Osprey-7B (Ours)	**77.54**

Experiments

Explore the effects of each components

- Region Caption

Method	Type	METEOR	CIDEr
GRIT [46]	Box	15.2	71.6
Kosmos-2[37]	Box	14.1	62.3
GLaMM [42]	Box	16.2	105.0
Osprey-7B (Ours)	Mask	**16.6**	**108.3**

- Different input **image sizes**

Input	#Image Tokens	Speed	Semantic Similarity	Semantic IoU
224	196	**6.0**	53.20	26.12
336	441	5.8	56.70	28.90
512	1024	3.5	65.24	38.19
800	2500	1.9	**68.29**	**42.66**

- Performance with **different vision encoders**

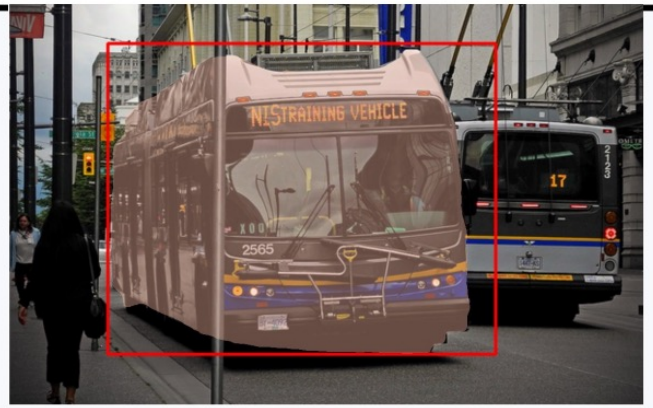
Vision Encoder	PQ	AP	mIoU
ViT-L	38.86	29.02	29.51
ConvNeXt-L	**42.50**	**31.72**	**29.94**

- With and without **short-from prompt** and **positive/negative samples**

Method	**LVIS**		**PACO**	
	Semantic Similarity	Semantic IoU	Semantic Similarity	Semantic IoU
w/o Short-form	56.41	25.65	50.26	23.29
w/o Pos./Neg.	63.55	36.70	71.59	50.39
Osprey-724K	**65.24**	**38.19**	**73.06**	**52.72**

Comparison with other approaches

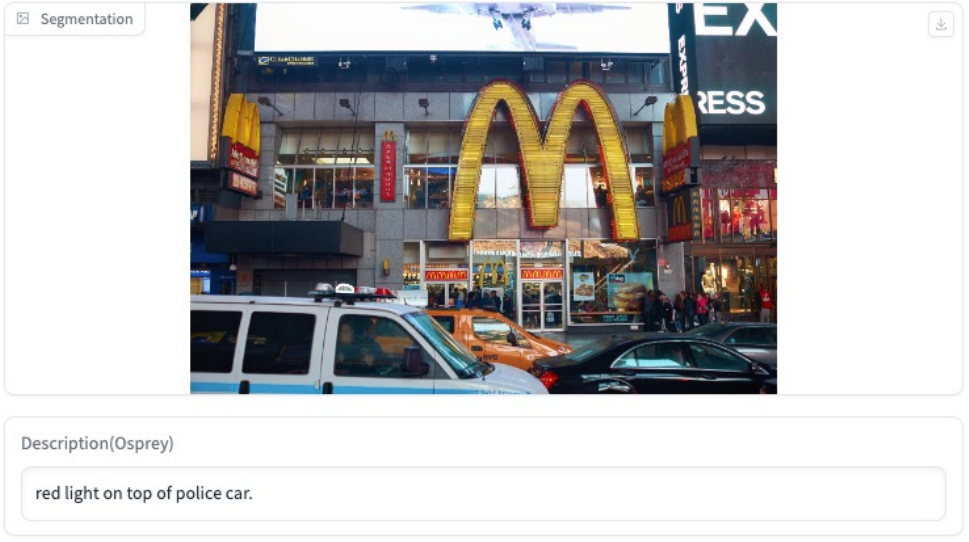
Detailed and accurate region description

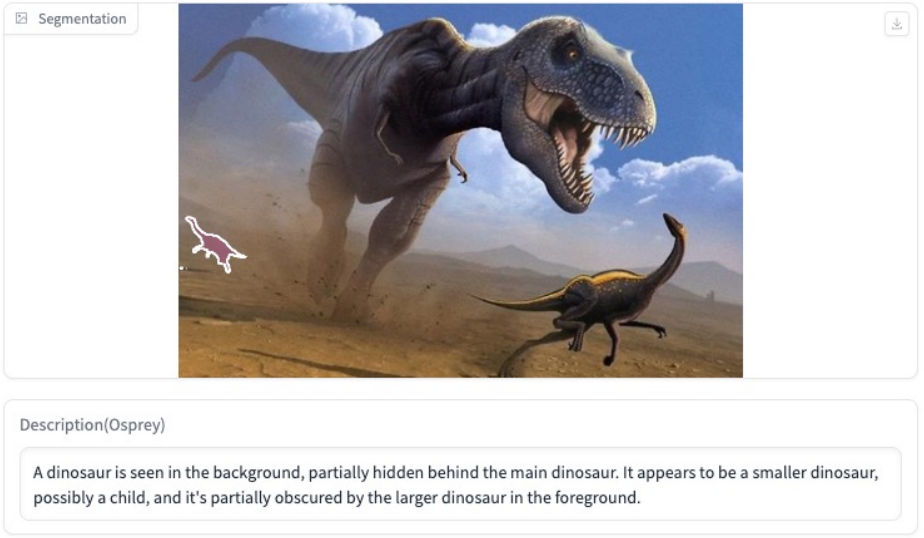


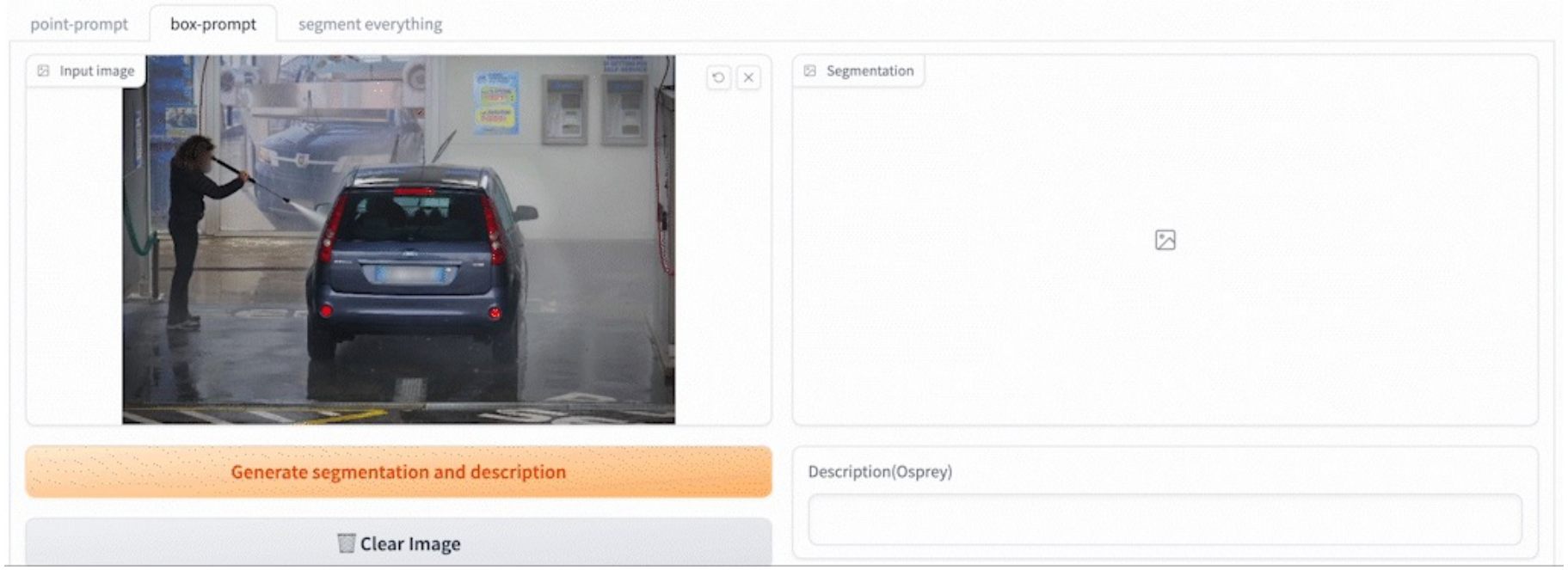
Osprey	A bus with the **number 2565** on it is parked on the left side of the street, next to a bus with the **number 17** on it. The bus with the number 2565 is the closest to the viewer, and it has **a yellow sign** on the front.
Kosmos	A bus and a truck driving down a street.
Shikra	Bus with "training vehicle" written on it.
GPT4RoI	A bus heading to vancouver.
LLaVA-1.5	The image depicts a city street with a large black and white bus driving down the road. The bus is in motion, and it is passing by a bus stop where a person is waiting. The bus stop is equipped with a bus stop sign, which is visible in the scene. The street is lined with buildings, creating an urban atmosphere. There is also a traffic light present in the scene, indicating that the bus is traveling on a busy street with controlled intersections.

Visualizations from online Demo

Fine-grained Understanding ability



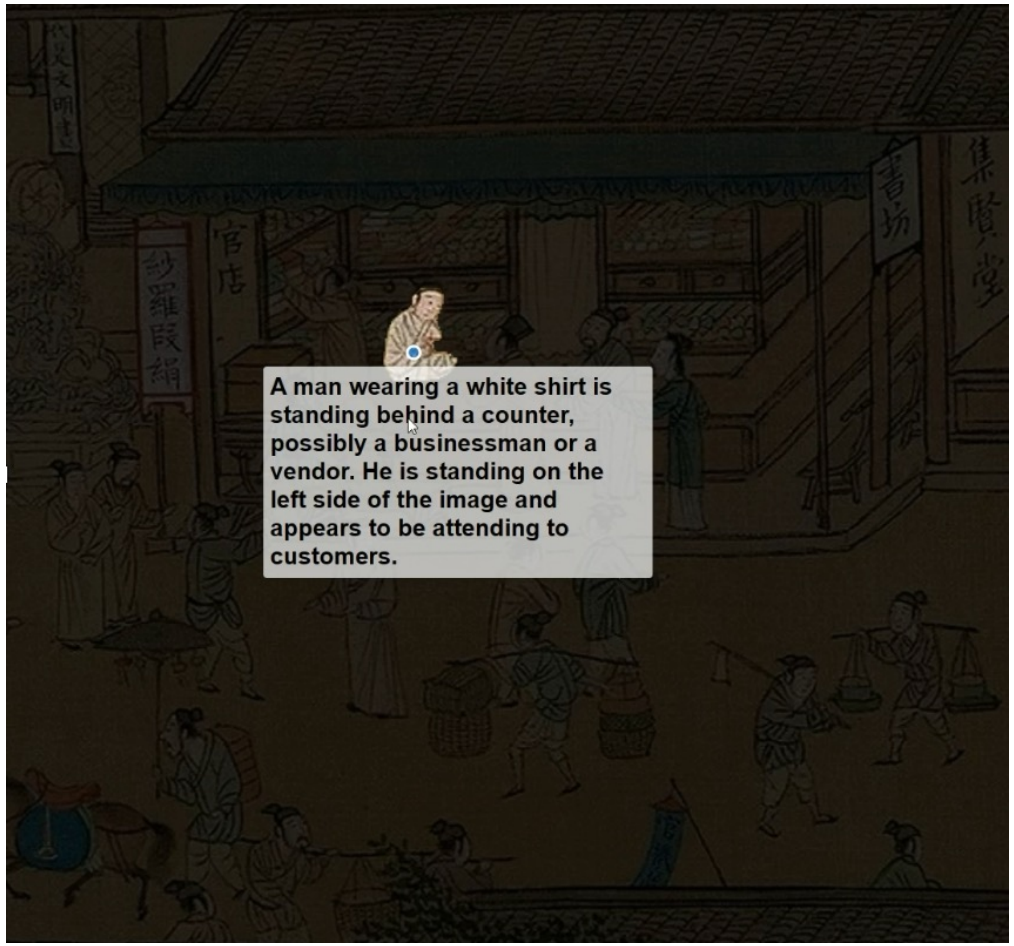




Combining with SAM

Powerful generalization ability

Video Demo



A part of *Along the River During the Qingming Festival*

清明上河图

Conclusion

- Osprey
  - a novel approach to incorporate pixel-level mask region references into language instructions.
  - seamlessly integrate with SAM to generate the semantics.
- Osprey-724K
  - Contains data of different granularities and levels of detail.
  - Robustness & Flexibility.
- And open-source!
- **Future work**
  - Chatting + Fine-grained grounding.

Thanks for your attention!

Demo, code, data, model

can be found at:

https://github.com/CircleRadon/Osprey