VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks

Wenhai Wang*[1], Zhe Chen*[2,1], Xiaokang Chen*[3,1], Jiannan Wu*[4,1], Xizhou Zhu[5,1],
Gang Zeng[3], Ping Luo[4,1], Tong Lu[2], Jie Zhou[6], Yu Qiao[1], Jifeng Dai†[6,1]
[1]OpenGVLab, Shanghai AI Laboratory [2]Nanjing University [3]Peking University [4]The University of Hong Kong
[5]SenseTime Research [6]Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Code

arXiv

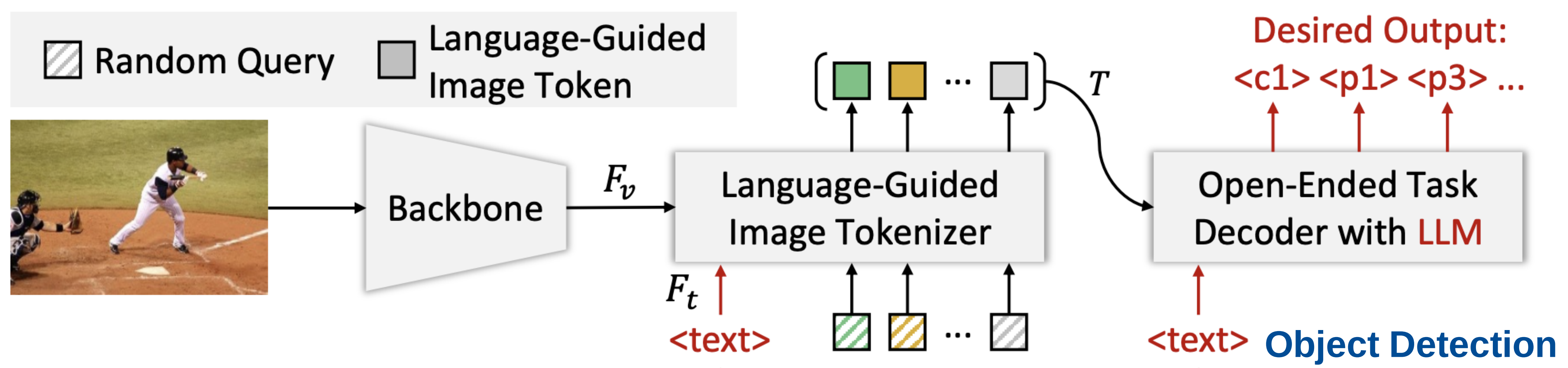
1. Motivation

- **Lack of open-ended task abilities in the field of computer vision**. Large language models (LLMs) like ChatGPT have notably accelerated progress towards artificial general intelligence (AGI), with their impressive zero-shot capacity for user-tailored tasks. However, despite the availability of numerous powerful vision foundation models (VFMs), they are still restricted to tasks in a pre-defined form, struggling to match the open-ended task capabilities of LLMs.

2. Contributions

- We propose VisionLLM, **the first LLM-based framework** that leverages the power of LLMs to address vision-centric tasks in an open-ended and customizable manner.
- We overcome many difficulties when porting LLMs to vision-centric tasks, by designing **unified language instruction** that matches the format of language models and covers various vision-centric tasks. Correspondingly, we develop a **language-guided image tokenizer** and an **LLM-based task decoder** that can handle open-ended tasks according to the given language instructions based on the LLMs' reasoning and parsing capabilities.
- We demonstrate the remarkable generality of our models, showcasing their ability to handle diverse scenarios, including **random object categories, random output formats, and random task descriptions**. Our model also yields an impressive mAP score of 60+% on the COCO detection dataset.

3. Methodology

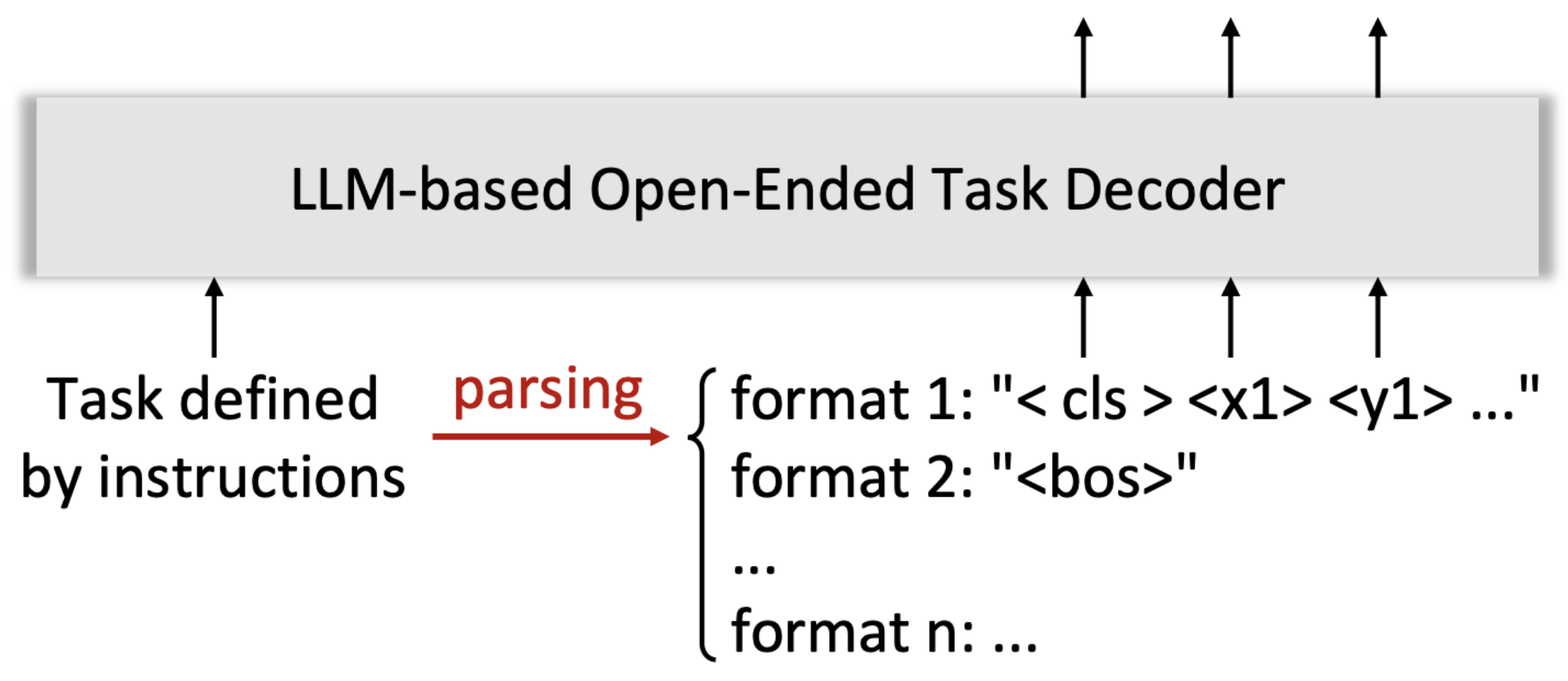


Language Instructions <text>

Vision-language example: "*Describe the image* <image> *in details.*"

Vision-only example: "*For each object in image* <image> *that is a member of class set* <class>, *output a tuple with the class label and the coordinates of a polygon with 16 points that encloses the object. The coordinates should be within range* <range>. *The output format should be (c, x1, y1, ...).*"

(a)


LLM-based Open-Ended Task Decoder
Task defined by instructions
parsing
format 1: "< cls > <x1> <y1> ..."
format 2: "<bos>"
...
format n: ...

(b)

(a) Overview of VisionLLM. It consists of three parts: a unified language instruction designed to accommodate both vision and vision-language tasks, an image tokenizer that encodes visual information guided by language instructions, and an LLM-based open-ended task decoder that executes diverse tasks defined by language instructions.

(b) Illustration of the "output-format-asquery" decoding process. "<cls><x1>" denote the queries of the object's class index and boundary points, and "<bos>" denotes the beginning of string.

4. Experimental Results

- **Ablation studies**

w/ BERT	Freeze	COCO	RefCOCO
-	-	44.7	48.1
✓	-	44.8	84.1
✓	✓	1.3	34.3

(a) Effect of text encoder in the language-guided image tokenizer.

Tokenization	AP
Average Pooling	23.1
Ours	44.8

(b) Effect of image tokenization method.

LoRA	Randomness	AP
✗	✗	45.2
✗	✓	1.2
✓	✓	44.8

(c) Effect of LoRA.

#Bins	AP
257	34.9
513	40.8
1025	44.8
2049	44.8

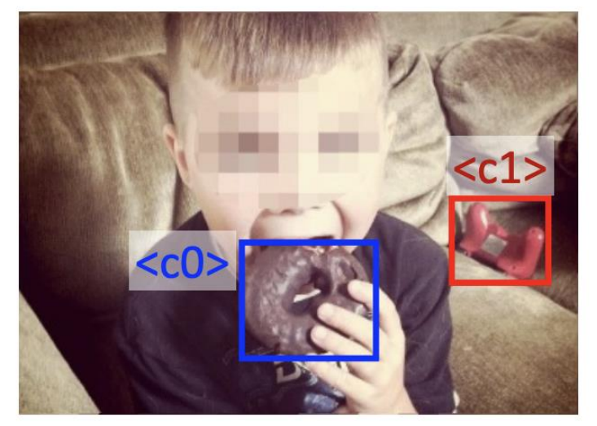
(d) Effect of #Bins.

Seq2Seq	AP
✓	-
✗	44.8

(e) Effect of Seq2Seq.

- **Visualizations**

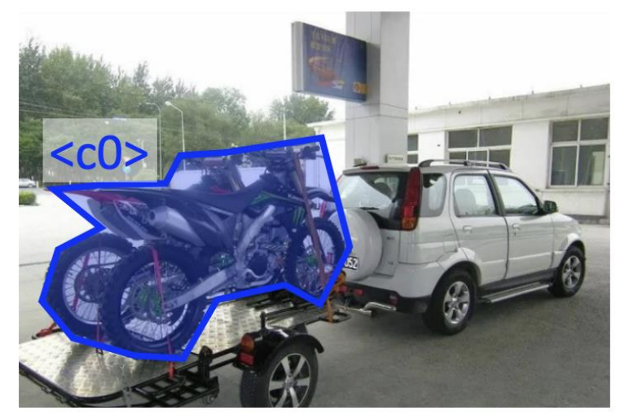
Human: "*Identify the objects in* <image> *that belong to {'What is the child eating?': <c0>, 'red gamepad': <c1>} and draw a bounding box around each one. The output should be a list of tuples in the format (c, x1, y1, x2, y2), where c is the class label and x1, y1, x2, y2 are the coordinates of the top-left and bottom-right corners of the bounding box. The coordinates should be within the range 1024.*"



VisionLLM: "*The bounding boxes are [(<c0>, 226.4, 229.8, 363.1, 347.4), (<c1>, 441.1, 183.5, 538.6, 269.9)].*"

(a) Object Detection with Customized Class Set.

Human: "*For each object in* <image> *that is a member of {'motorbike': <c0>}, output a tuple with the class label and the coordinates of a polygon with 16 points that encloses the object. The coordinates should be within the range 1024. The output format should be (c, x1, y1, x2, y2, ..., x16, y16).*"



VisionLLM: "*The detected polygons are [(<c0>, 135.3, 95.7, 123.4, 53.4, 84.9, 57.6, 66.8, 60.5, 60.1, 72.3, 34.2, 71.4, ..., 124.9, 119.3)].*"

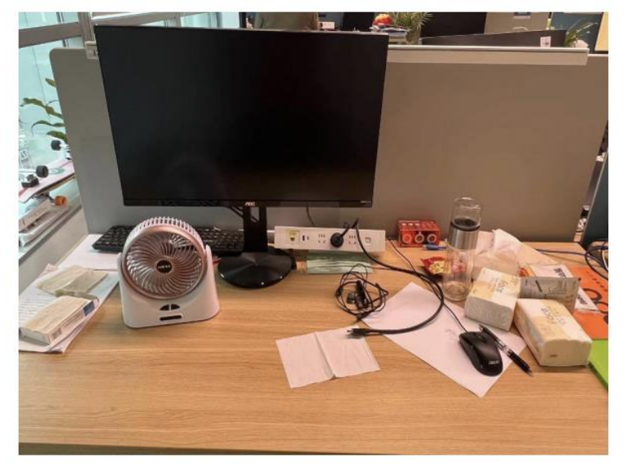
(b) Object Location with Editable Output Format.

Human: "*Provide a long description of the* <image>."

VisionLLM: "*The image features a white dog laying down on a foldable outdoor cooling pad, which is placed on top of a wooden table. The dog has its head in the middle of the cooling pad, and a fan is positioned behind it, blowing air onto the pad. The table is outdoors, in a grassy area, and there are two chairs placed around it, one on each side. The scene appears to be set in a sunny weather, with a bright and sunny sky, which contributes to the overall outdoor atmosphere.*"

(c) Image Description with Controllable Text Length.

Human: "*With the objects in the* <image>, *if you feel hot, what will you do?*"



VisionLLM: "*First, You can also try turning off the computer and other electrical appliances in the room to reduce the heat generated by their operation. As an alternative, you can use the desk fan provided in the image to create a cooling effect in the room. You can plug in the fan and adjust its settings to ensure it blows air in a direction that reaches the room's occupants effectively.*"

(d) Visual Question Answer with Complex Reasoning.

- **Results on standard vision-centric tasks**

Method	Backbone	Open-Ended	Detection			Instance Seg.			Grounding	Captioning	
			AP	AP$_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$	P@0.5	BLEU-4	CIDEr
Specialist Models											
Faster R-CNN-FPN [48]	ResNet-50	-	40.3	61.0	44.0	-	-	-	-	-	-
DETR-DC5 [7]	ResNet-50	-	43.3	63.1	45.9	-	-	-	-	-	-
Deformable-DETR [82]	ResNet-50	-	45.7	65.0	49.1	-	-	-	-	-	-
Mask R-CNN [22]	ResNet-50	-	41.0	61.7	44.9	37.1	58.4	40.1	-	-	-
Polar Mask [69]	ResNet-50	-	-	-	-	30.5	52.0	31.1	-	-	-
Pix2Seq [8]	ResNet-50	-	43.2	61.0	46.1	-	-	-	-	-	-
UNITER [11]	ResNet-101	-	-	-	-	-	-	-	81.4	-	-
VILLA [19]	ResNet-101	-	-	-	-	-	-	-	82.4	-	-
MDETR [27]	ResNet-101	-	-	-	-	-	-	-	86.8	-	-
VL-T5 [13]	T5-B	-	-	-	-	-	-	-	-	-	116.5
Generalist Models											
UniTab [72]	ResNet-101	-	-	-	-	-	-	-	88.6	-	115.8
Uni-Perceiver [83]	ViT-B	-	-	-	-	-	-	-	-	32.0	-
Uni-Perceiver-MoE [81]	ViT-B	-	-	-	-	-	-	-	-	33.2	-
Uni-Perceiver-V2 [28]	ViT-B	-	58.6	-	-	50.6	-	-	-	35.4	116.9
Pix2Seq v2 [9]	ViT-B	-	46.5	-	-	38.2	-	-	-	34.9	-
VisionLLM-R50$_{sep}$	ResNet-50	-	44.8	64.1	48.5	25.2	50.6	22.4	84.4	30.8	112.4
VisionLLM-R50	ResNet-50	✓	44.6	64.0	48.1	25.1	50.0	22.4	80.6	31.0	112.5
VisionLLM-H	Intern-H	✓	60.2	79.3	65.8	30.6	61.2	27.6	86.7	32.1	114.2