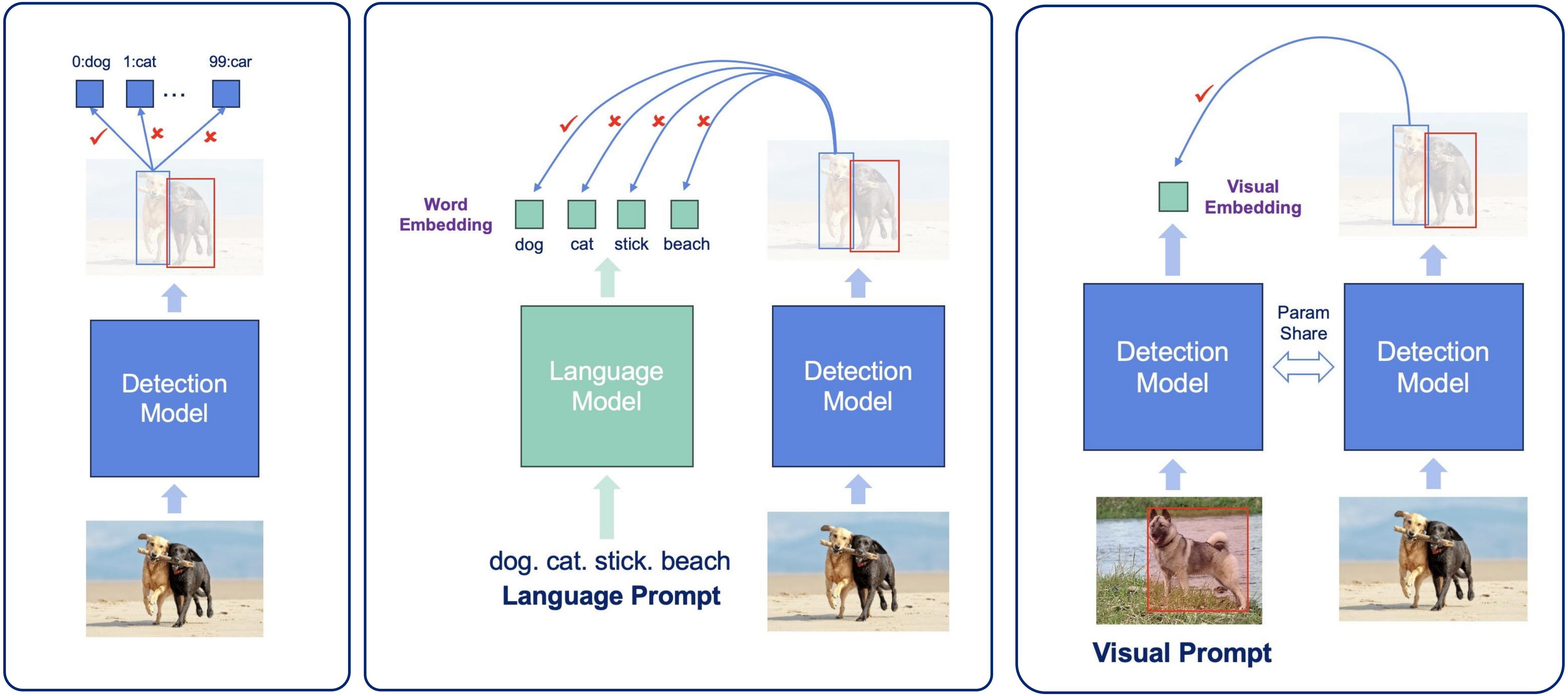
idea

Taming Multimodal LLM for Object Perception

蒋擎

4-15

What is the Next Step for Object Detection?

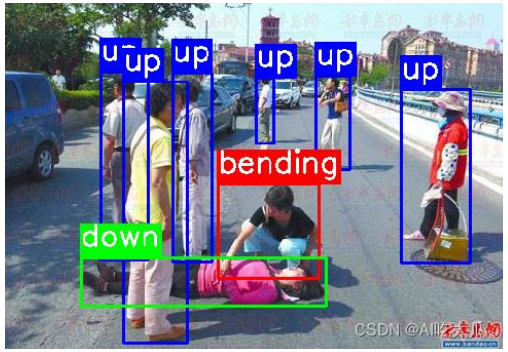


Closed-Set → Open-vocabulary → Visual Prompt

DETR | Grounding DINO | T-Rex

What is the Next Step for Object Detection?

Most Detection Entities Can be Described in Language



摔倒检测

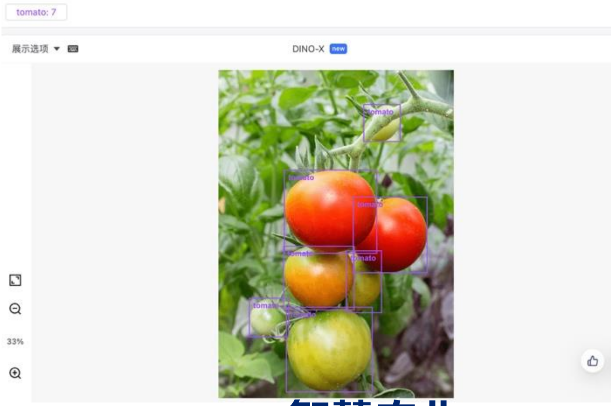
“person fallen”

佩戴安全帽检测

“person that are not wearing helmet”

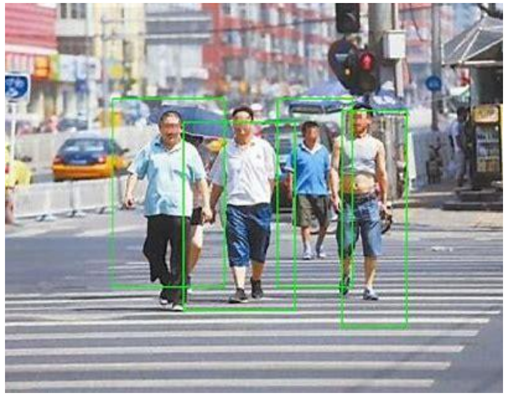
工位睡觉检测

“person that is sleeping”



智慧农业

“tomato that are not ripe”



行人安全检测

“person on the crossroad”

抽烟检测

“person that are smoking”

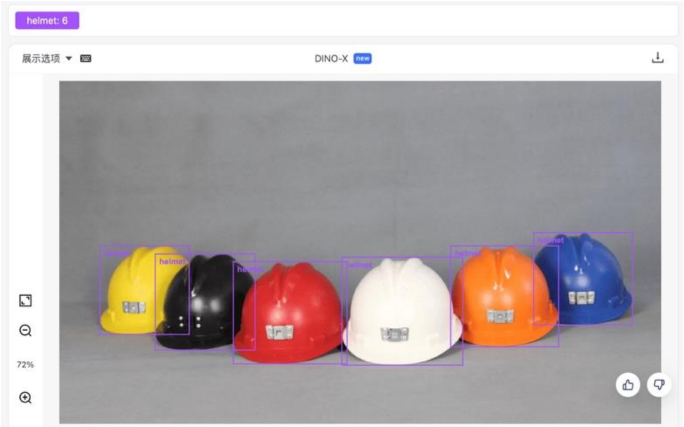
交通管理

“cars that are crushed”

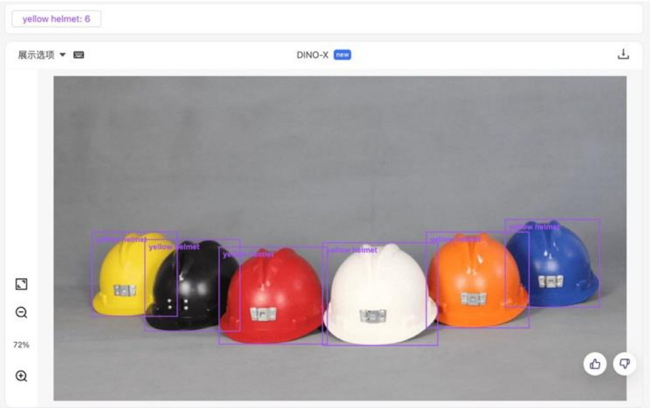
What is the Next Step for Object Detection?

Finding 1: State-of-the-art Open-set detection models lack language comprehension capabilities

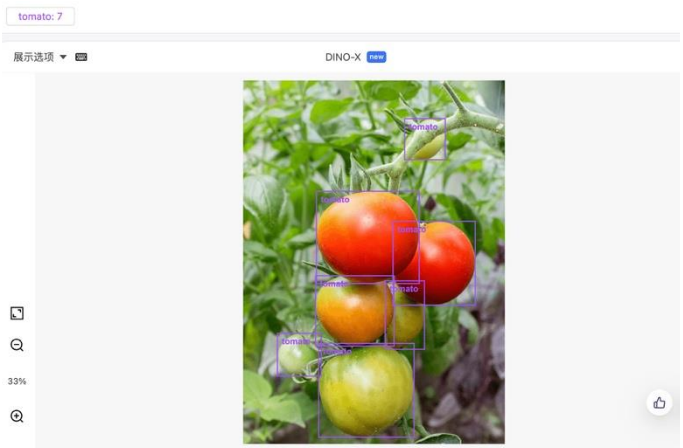
"helmet"



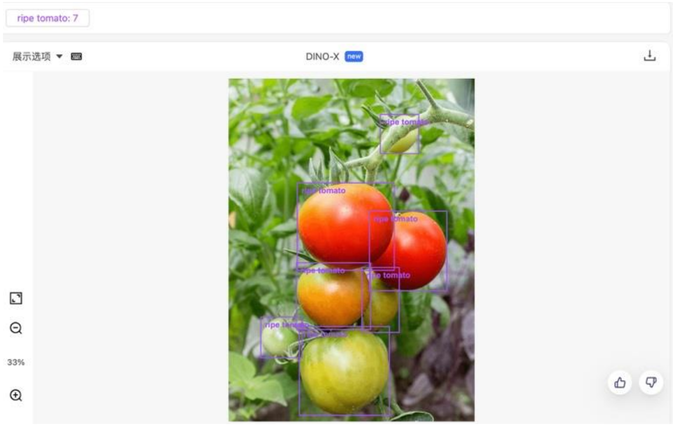
"yellow helmet"



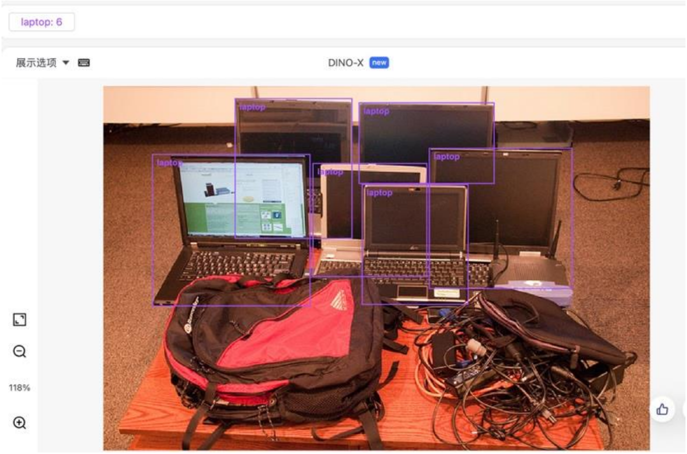
"tomato"



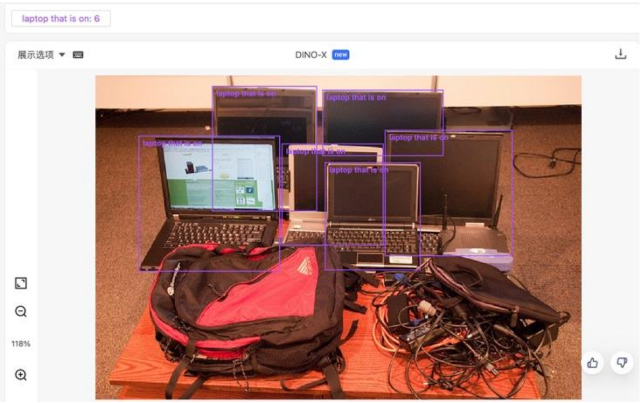
"ripe tomato"



"laptop"



"laptop that is on"



In Open-Vocabulary object detection, the term "open" only refers to openness in categories,

What is the Next Step for Object Detection?

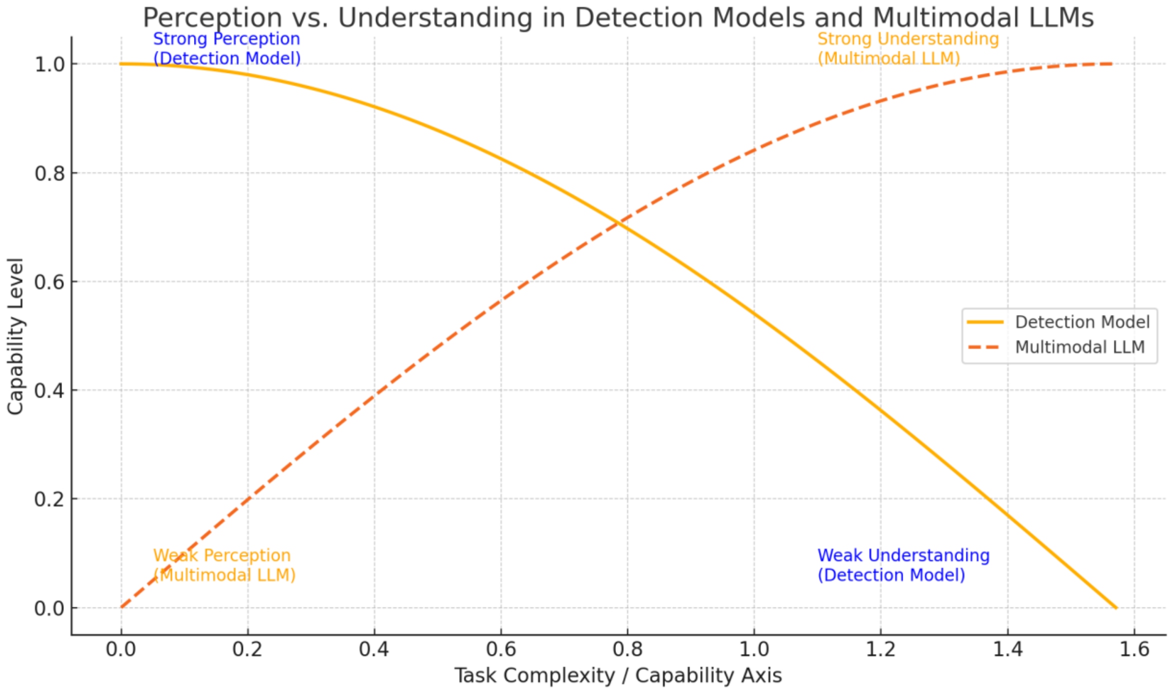
Finding 2: State-of-the-art Multimodal LLMs lack fine-grained perception capabilities

User: Please help me detect person in this image

MLLMs:

“Sure, here is person [[90, 70, 120, 340], [110, 70, 125, 400]]”

What is the Next Step for Object Detection?



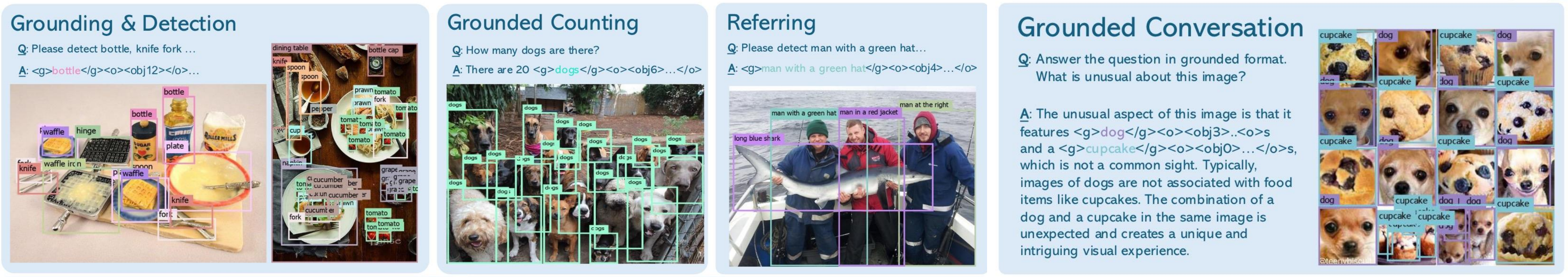
Detection Model：Strong Perception, Weak Understanding

Multimodal LLMs：Weak Perception, Strong Understanding

Next Step: A mode with both **strong perception** and **understanding** capabilities

ChatRex: Taming Multimodal LLM for Joint Perception and Understanding

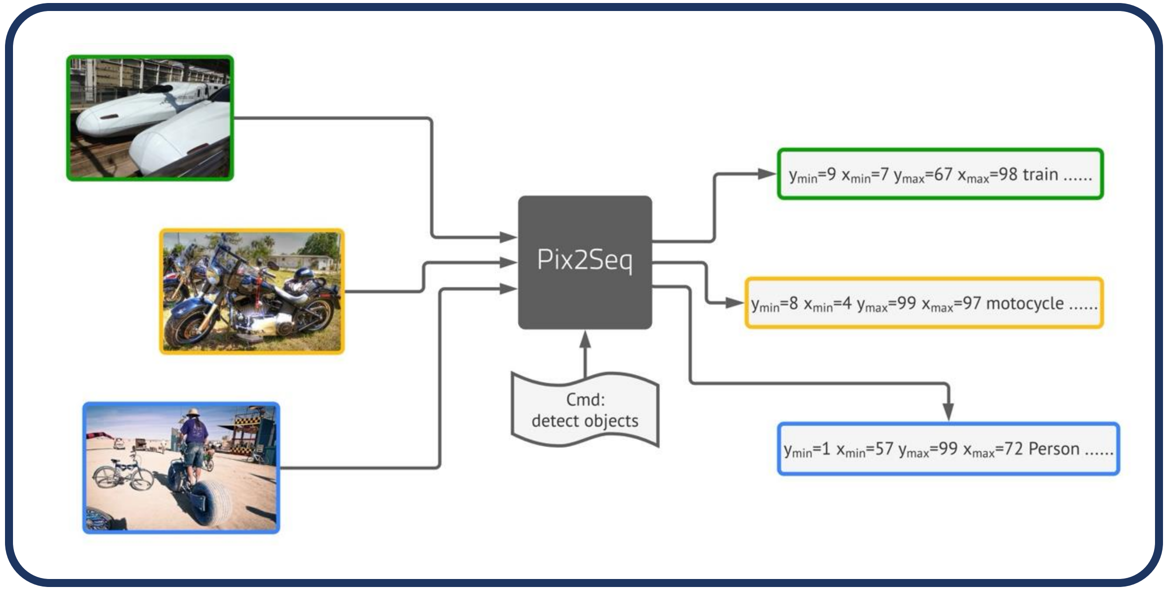
Qing Jiang[1,2] , Gen Luo[1] , Yuqin Yang[1,2] , Yuda Xiong[1] , Zhaoyang Zeng[1]
Yihao Chen[1] , Tianhe Ren[1] , Lei Zhang[1,2†]
[1]International Digital Economy Academy (IDEA)
[2]South China University of Technology
mountchicken@outlook.com , leizhang@idea.edu.cn



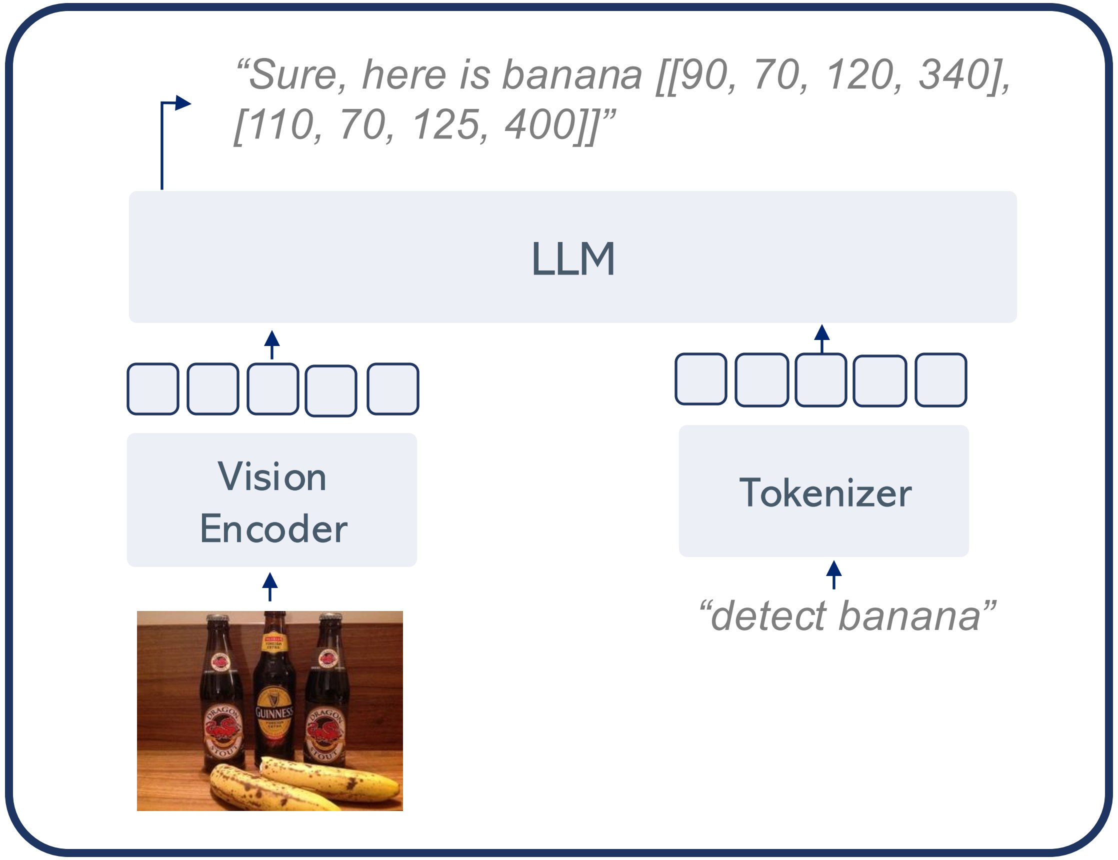
Jiang Q, Luo G, Yang Y, et al. Chatrex: Taming multimodal llm for joint perception and understanding[J]. arXiv preprint arXiv:2411.18363, 2024.

Motivation: How MLLMs do Detection?

How MLLMs do object detection: directly predict quantized coordinates as text tokens[1].



Pix2Seq[1]



Modern MLLMs

[1] Chen T, Saxena S, Li L, et al. Pix2seq: A language modeling framework for object detection[J]. arXiv preprint arXiv:2109.10852, 2021.

Motivation: How MLLMs do Detection?

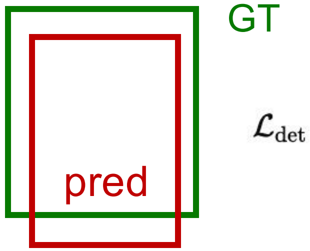
But the performance is not good enough

Method	Type	COCO-Val			LVIS-Mini Val					
		P@0.5	R@0.5	mAP	P@0.5	R@0.5	mAP	AP-R	AP-C	AP-F
Faster-RCNN [70]	Closed-set Detection Model	-	-	42.0	-	-	-	-	-	-
DETR [8]		-	-	43.3	-	-	-	-	-	-
Pix2Seq [12]		-	-	43.2	-	-	-	-	-	-
DINO [102]		-	-	49.4	-	-	-	-	-	-
Florence2 [88]	Open-set Detection Model	-	-	43.4	-	-	-	-	-	-
GLIP [39]		-	-	**49.8**	-	-	37.3	28.2	34.3	41.5
T-Rex2 [29]		-	-	46.5	-	-	**47.6**	**45.4**	46.0	**49.5**
Grounding DINO [52]		-	-	48.4	-	-	33.0	22.2	30.7	38.8
Shikra-7B [10]	MLLM	40.3	21.5	-	52.8	14.5	-	-	-	-
Ferret-7B [94]		66.3	33.5	-	72.9	25.2	-	-	-	-
Groma-7B [61]		69.9	28.9	-	76.3	10.9	-	-	-	-
InternVL2-7B [14]		45.3	24.5	-	51.6	13.1	-	-	-	-
Qwen2-VL-7B [85]		59.3	43.9	-	77.0	34.7	-	-	-	-
ChatRex-7B		**73.5**	**72.8**	48.2	**80.3**	**58.9**	42.6	44.6	**48.4**	37.2

Low Recall Rate

Motivation: What's the Challenge?

1. **Directly predict the coordinates is a hard task:** Regression V.S. Classification

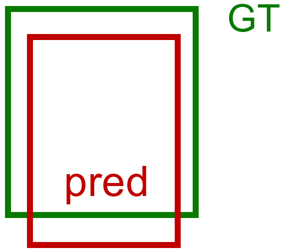


Loss

$$\mathcal{L}_{\text{det}} = \lambda_1 \cdot \text{L1}(\hat{b}, b) + \lambda_2 \cdot \text{Giou}(\hat{b}, b)$$

Low Loss

detection model training



Loss

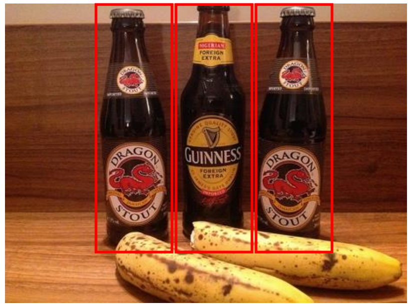
$$\mathcal{L}_{\text{MLLM}} = \text{CE}(\hat{y}, y)$$

High Loss

MLLM training

2. **Error Propagation:** Each box requires at least 9 tokens and can cause cascading errors.

3. **Ambiguity in Prediction Order:** Auto-regressive prediction needs a predefined sequence order.

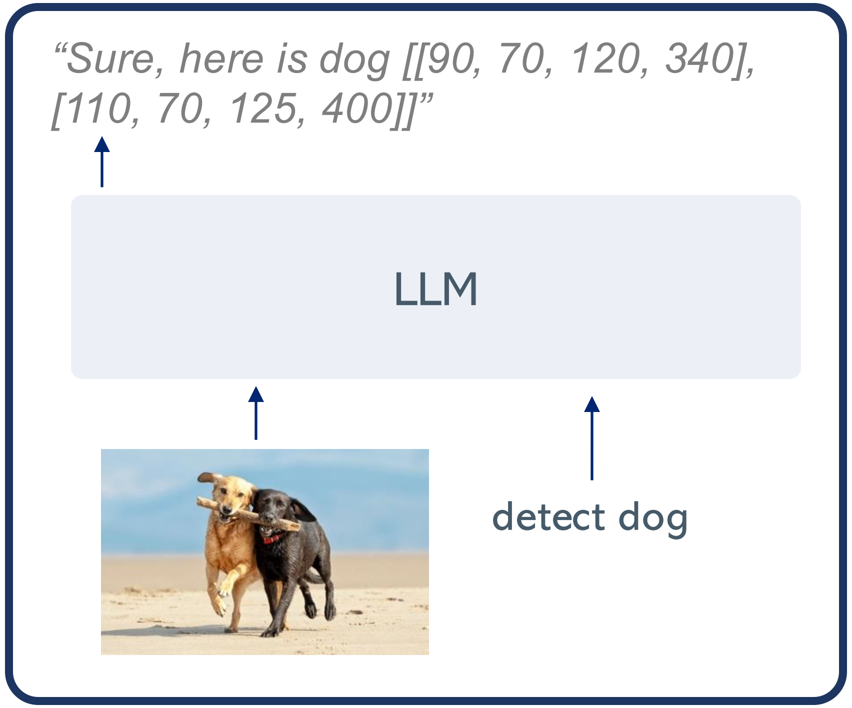


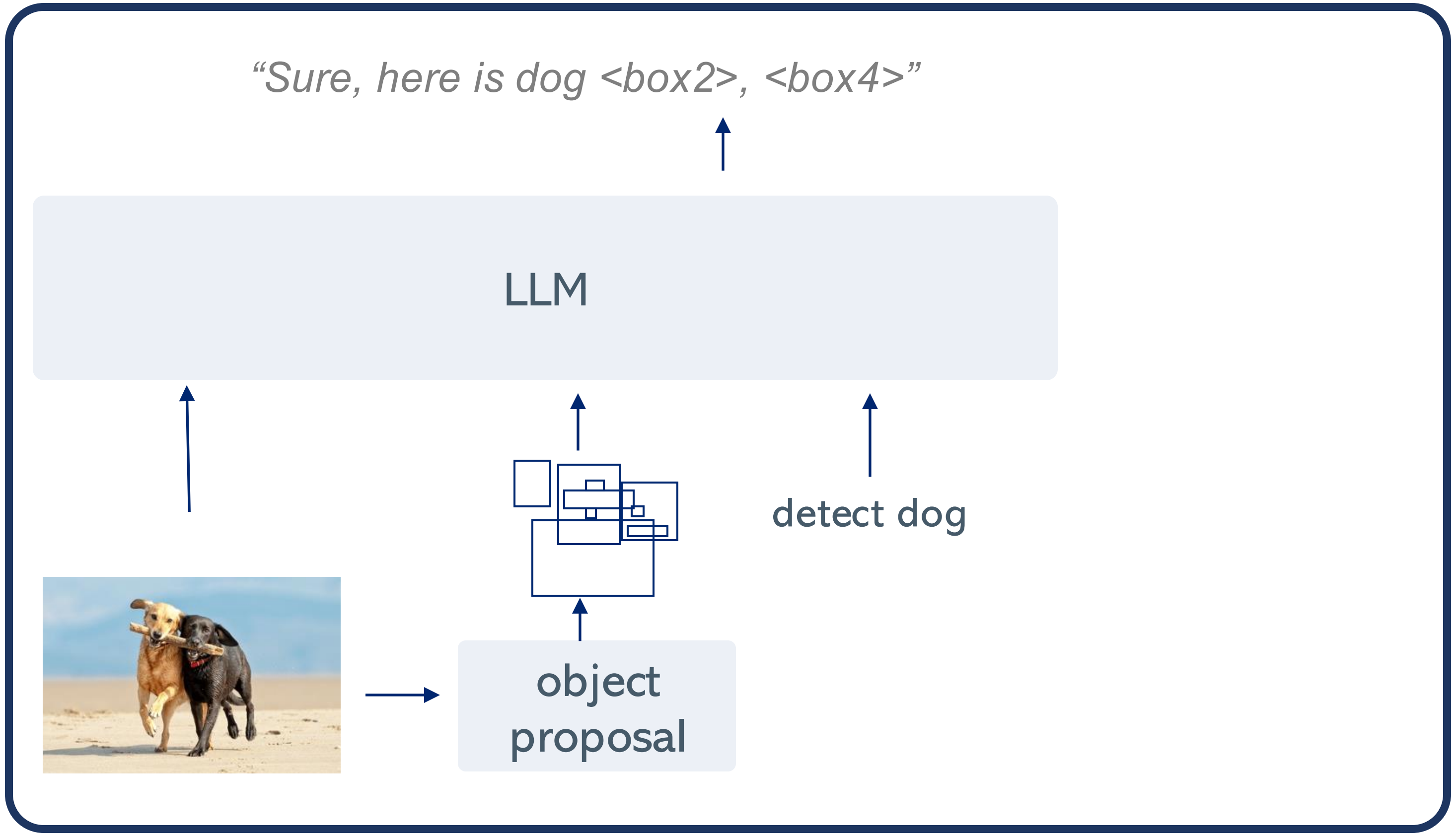
"bottle1, bottle2, bottle3" *"bottle3, bottle2, bottle1"* *"bottle2, bottle1, bottle3"*

4. **Quantization Range Limitation:** Large image (>1000 px) input can lead to quantization error.

Solution: Retrieval-based Perception MLLM

Core idea: LLM has strong understanding capability, while detection model has strong perception capability

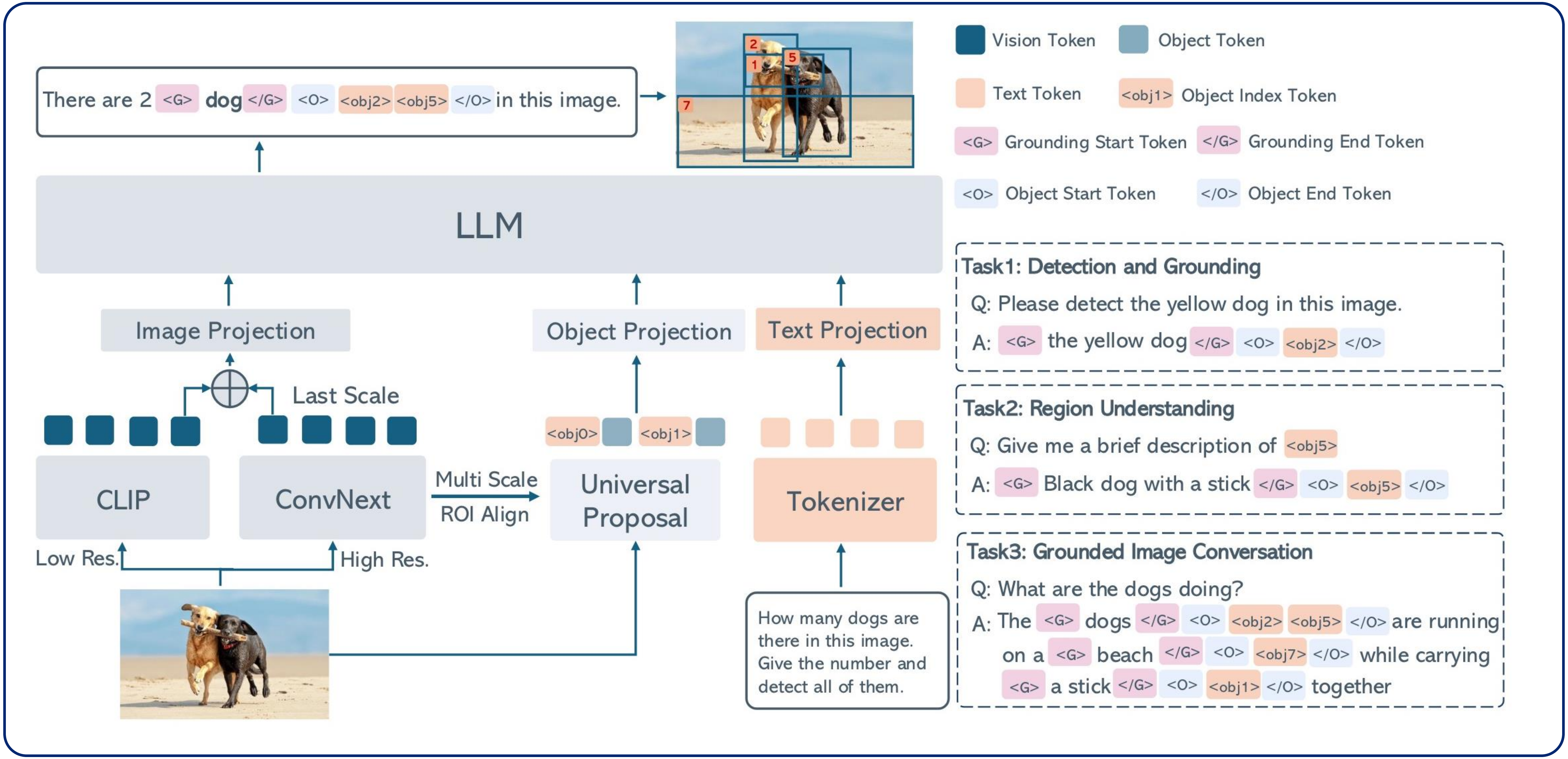




Retrieval based method

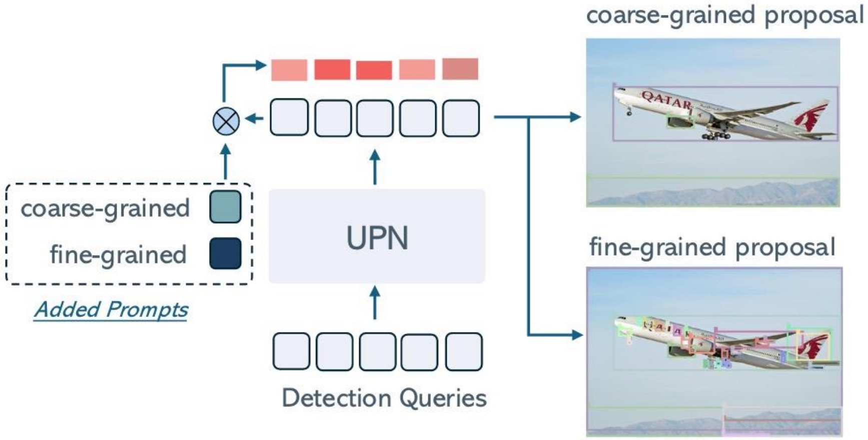
ChatRex: Detection-Oriented MLLM

idea

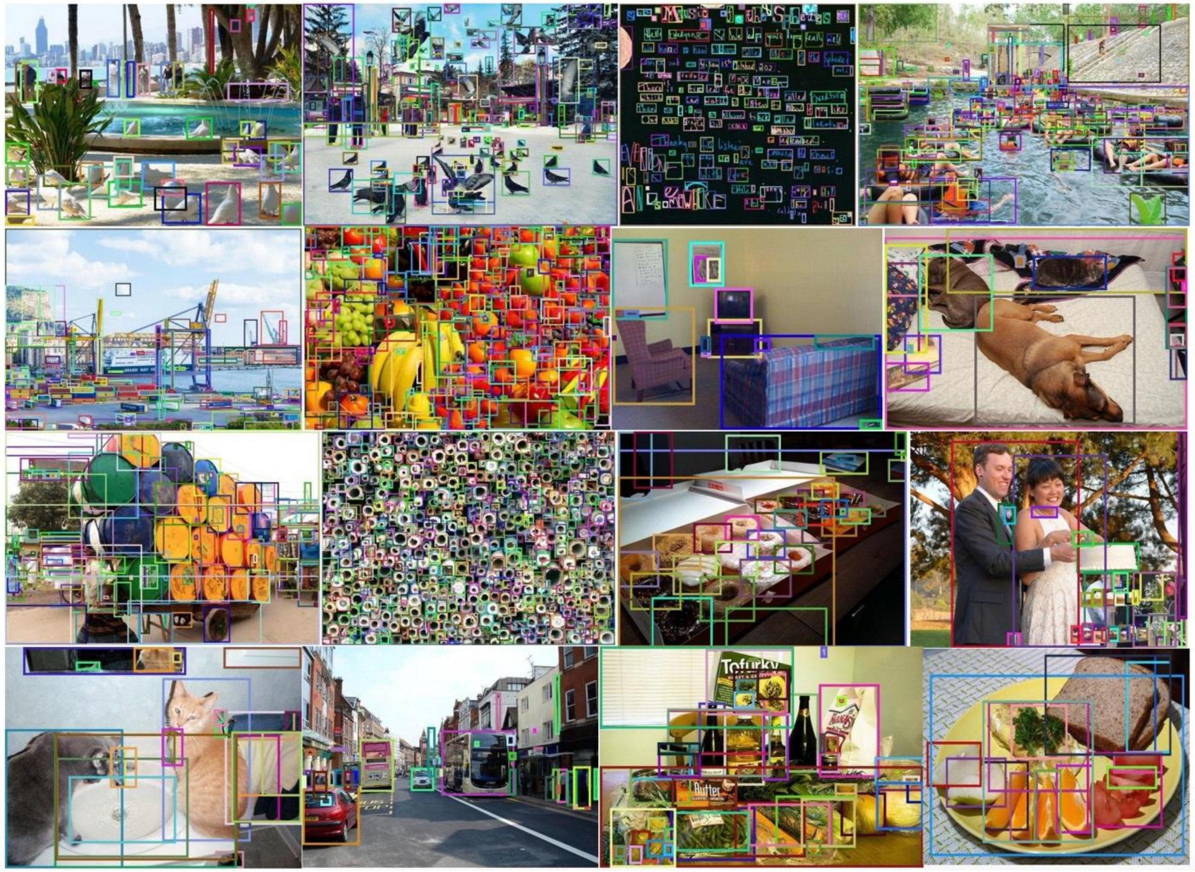


ChatRex: Detection-Oriented MLLM

ChatRex Architecture



Universal Proposal Architecture



ChatRex: Detection-Oriented MLLM

idea

Use the input box indices as the output of detection

Input and Output Formats for Detection Related Tasks.

Single-Object Detection

User: Can you help me detect dog in this image? Answer the question with object indices.
Assistant: <g>dog</g><o><objm>...<objn></o>.

Multi-Object Detection

User: Can you help me detect a yellow flower; a man wearing a blue shirt; a dog at the left in this image? Answer the question with object indices.
Assistant: <g>a yellow flower</g><o><objm>...<objn></o>\n
<g>a man wearing a blue shirt</g><o><objm>...<objn></o>\n
<g>a dog at the left</g><o><objm>...<objn></o>

ChatRex: Detection-Oriented MLLM

Use the input box indices as the output of detection

Input and Output Formats for Region Understanding Related Tasks.

Single-Object Description

User: What is the category name of <objm>.
Assistant: <g>cat</g><o><objm></o>.

Multi-Object Description

User: Can you provide a brief description for <objm>; <objn>?
Assistant: <g>a cat lying an the bed</g><o><objm></o>\n
<g>an opened tomato can</g><o><objn></o>\n

Region QA

User: What might be the species of <objm>?
Assistant: The species of <objm> appears to be a dog, likely a bread with black and white fur.

Region Relationship Analysis

User: Can you tell me about the relationship between <objm> and <objn>?
Assistant: <objm> and <objn> are standing next to each other, which suggests they might be companions or partners

ChatRex: Detection-Oriented MLLM

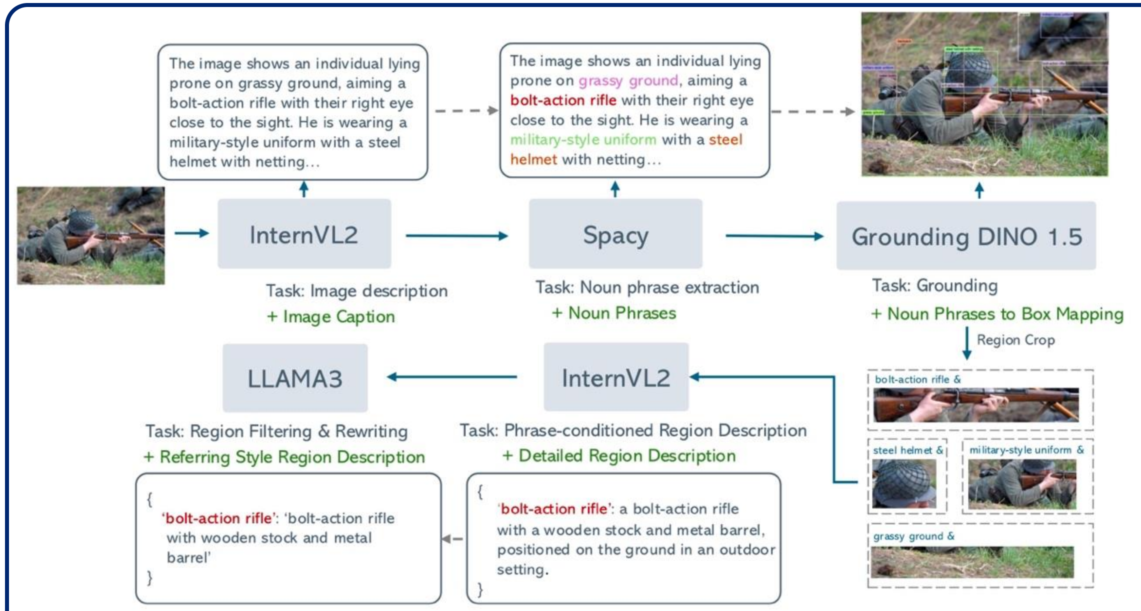
Use the input box indices as the output of detection

Input and Output Formats for Grounded Conversation Task.

User: Please briefly describe this image and detect all the mentioned objects. Answer with grounded object indexes.

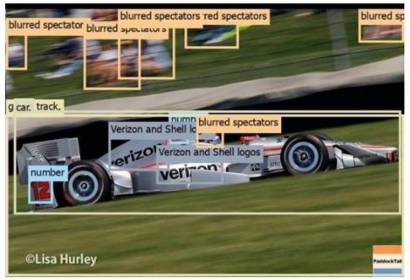
ChatRex: A <g>man</g><o><objm></o> in a <g>white tuxedo</g><o><objn></o> with a <g>red bow tie</g><o><objm></o> is holding an <g>Oscar statuette</g><o><objn></o> and standing on a stage with a microphone, while a large, ornate Oscar statue is visible in the background.

Experiments: Training Data and Recipe



RexVerse-2M Dataset Engine

The image features a unique combination of objects: a green, semi-precious stone ring with a polished and faceted surface, resting on a large, irregularly shaped, clear crystal rock, surrounded by various other colorful gemstones and moss in the background, creating a visually appealing and intricate display



A high-speed Formula 1 race car, number 12, adorned with Verizon and Shell logos, navigates a turn on a green track with blurred spectators in the background, captured in motion by the sharp focus on the car.

RexVerse-2M Dataset Example

Stage	Task	# Samples	Datasets
Stage1	Image Caption	976K	ALLAVA-4V-Caption [9]
Stage2	Grounding & Region Understanding	2.07M	COCO [46], O365 [75], LVIS [25], RefCOCO/+/g [30, 62, 95], Rexverse-2M
Stage3	Grounding & Counting & Region Understanding & Grounded Conversation Conversation & QA	3.8M	Rexverse-2M, COCO, O365, LVIS, RefCOCO/+/g [30, 62, 95], PACO [68] MVDP [47], Osprey [97], CrowdHuman [74], VCR [99] , ALLAVA-4V-Instruct [9] , LLAVA-1.5 [49], LLaVA-Onevision [37]

Stage1: Alignment Training

Stage2: Perception Training

Stage3: Joint Perception and Understanding Training

Experiments: Perception and Understanding

Method	Type	COCO-Val			LVIS-Mini Val						RefCOCO			RefCOCO+			RefCOCOg	
		P@0.5	R@0.5	mAP	P@0.5	R@0.5	mAP	AP-R	AP-C	AP-F	val	testA	testB	val	testA	testB	val	test
Faster-RCNN [70]	Closed-set Detection Model	-	-	42.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DETR [8]		-	-	43.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pix2Seq [12]		-	-	43.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DINO [102]		-	-	49.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Florence2 [88]	Open-set Detection Model	-	-	43.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIP [39]		-	-	**49.8**	-	-	37.3	28.2	34.3	41.5	-	-	-	-	-	-	-	-
T-Rex2 [29]		-	-	46.5	-	-	**47.6**	**45.4**	46.0	**49.5**	-	-	-	-	-	-	-	-
Grounding DINO [52]		-	-	48.4	-	-	33.0	22.2	30.7	38.8	89.2	91.9	86.0	81.1	87.4	74.7	84.2	84.9
Shikra-7B [10]	MLLM	40.3	21.5	-	52.8	14.5	-	-	-	-	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
Ferret-7B [94]		66.3	33.5	-	72.9	25.2	-	-	-	-	-	-	-	-	-	-	-	-
Groma-7B [61]		69.9	28.9	-	76.3	10.9	-	-	-	-	89.5	92.1	86.3	83.9	88.9	78.1	86.4	87.0
InternVL2-7B [14]		45.3	24.5	-	51.6	13.1	-	-	-	-	87.1	91.1	80.7	79.8	87.9	71.4	82.7	82.7
Qwen2-VL-7B [85]		59.3	43.9	-	77.0	34.7	-	-	-	-	**91.7**	93.6	**87.3**	85.8	90.5	**79.5**	87.3	87.8
ChatRex-7B		**73.5**	**72.8**	48.2	**80.3**	**58.9**	42.6	44.6	**48.4**	37.2	91.0	**94.1**	87.0	**89.8**	**91.9**	79.3	**89.8**	**90.0**

Perception Capability

ChatRex achieves strong performance in object detection tasks, while also demonstrating competitive performance on multimodal benchmarks.

Model	MME	MMB	SEED^I	MMStar	MMVet	MMMU	AI2D	OCRBench	TextVQA	POPE	Hallusion
BLIP-2 [38]	1293.8	-	49.7	-	22.4	-	-	-	-	85.3	-
InstructBLIP [16]	1212.8	-	-	-	-	-	-	-	-	78.9	-
Mini-Gemini-HD-8B [42]	1606.0	72.7	73.2	-	-	37.3	73.5	47.7	70.2	-	-
LLaVA-HR [58]	1554.0	-	64.2	-	31.2	-	-	-	67.1	87.6	-
LLaVA-NeXT-7B [51]	1498.0	68.7	72.2	38.4	42.2	35.3	69.0	531	64.6	86.7	29.1
Eagle-X5-7B [76]	1579.0	68.8	73.5	41.7	42.6	36.3	77.2	574	71.2	88.8	37.8
MM1.5-7B [103]	1514.9	-	73.4	-	42.2	41.8	72.2	635	76.5	88.6	-
Cambrian-8B [81]	1547.1	75.9	74.7	47.1	48.9	41.6	73.6	610	71.7	86.8	39.4
LLaVA-OV-7B [37]	1577.8	83.2	76.7	61.9	51.9	47.9	82.4	622	78.5	88.4	31.6
InternVL2-8B [14]	1639.7	81.7	75.4	61.5	54.2	49.8	83.0	794	77.4	84.2	45.0
Qwen2-VL-7B [85]	1639.2	83.0	76.0	60.7	62.0	54.1	83.0	845	84.3	88.4	50.6
ChatRex-7B	1544.0	81.1	74.4	57.5	41.5	46.7	79.1	626	69.1	87.6	39.1

Table 3. Comparison of different models on multimodal benchmarks.

Understanding Capability

Applications: Common/Long-tailed Object Detection

QA Example:

User: Please detect person; cup in this image. Answer the question with object indexes.

ChatRex: <g>person</g><o><obj1><obj5><obj16><obj21></o>\n
<g>cup</g><o><obj12><obj14><obj33></o>

Figure 7: Visualization on Common Object Detection Task.

QA Example:

User: Please detect saluki; folding chair in this image. Answer the question with object indexes.

ChatRex: <g>saluki</g><o><obj12></o>\n
<g>folding chair</g><o><obj19><obj23></o>\n

Visualization:

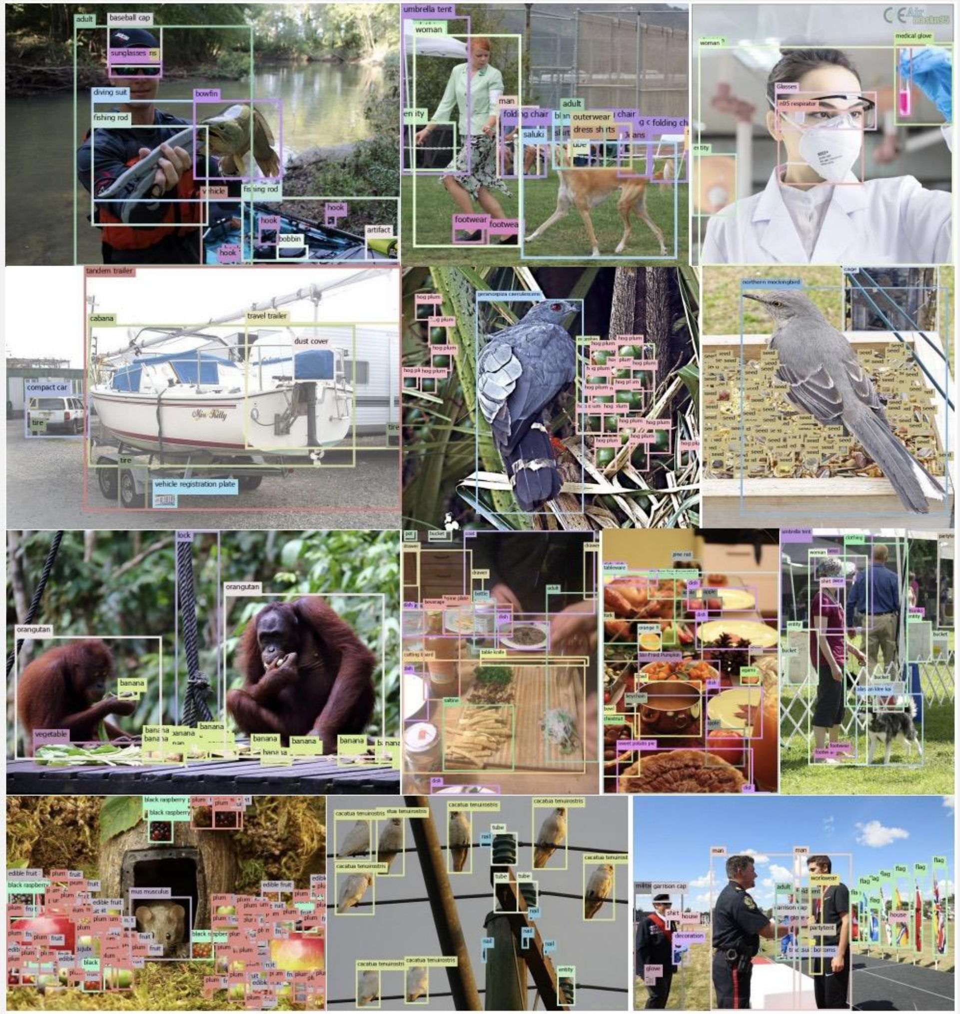


Figure 8: Visualization on Long-tailed Object Detection Task.

Applications: Short Phrase/Referring Detection

QA Example:

User: Please detect bamboo cutting board; fresh green onions in this image. Answer the question with object indexes.

ChatRex: <g>bamboo cutting board</g><o><obj1></o>\n
<g>fresh green onions</g><o><obj52><obj66></o>\n

Visualization:

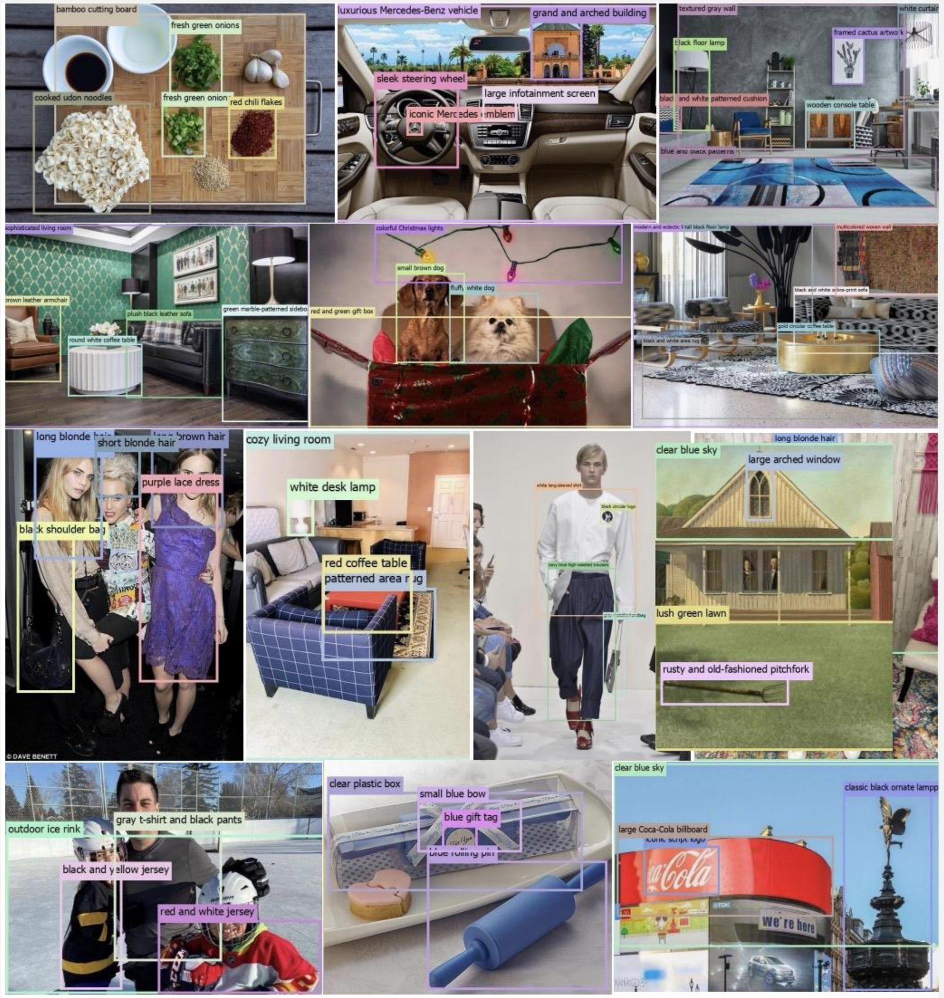


Figure 9: Visualization on Short-Phrase Object Detection Task.

User: Please detect man step on a skateboard; man wearing a hat and a phone in his hand in this image. Answer the question with object indexes.

ChatRex: <g>man step on a skateboard</g><o><obj26></o>\n
<g>man wearing a hat and a phone in his hand</g><o><obj11></o>\n

Visualization:

Figure 10: Visualization on Referring Object Detection Task.

Applications: Region Understanding

QA Example:

User: What is <obj0> wearing and what is he doing now?

ChatRex: <obj0> is wearing a black jacket and a black beanie. He is standing in a store, looking at something on the shelf.

Visualization:

Figure 12: Visualization on Region QA Task.

QA Example:

User: Give a short phrase that describes <obj0>;<obj1>.

ChatRex: <g>Starbucks logo</g><o><obj0></o>\n
<g>a coffee cup</g><o><obj1></o>\n

Visualization:

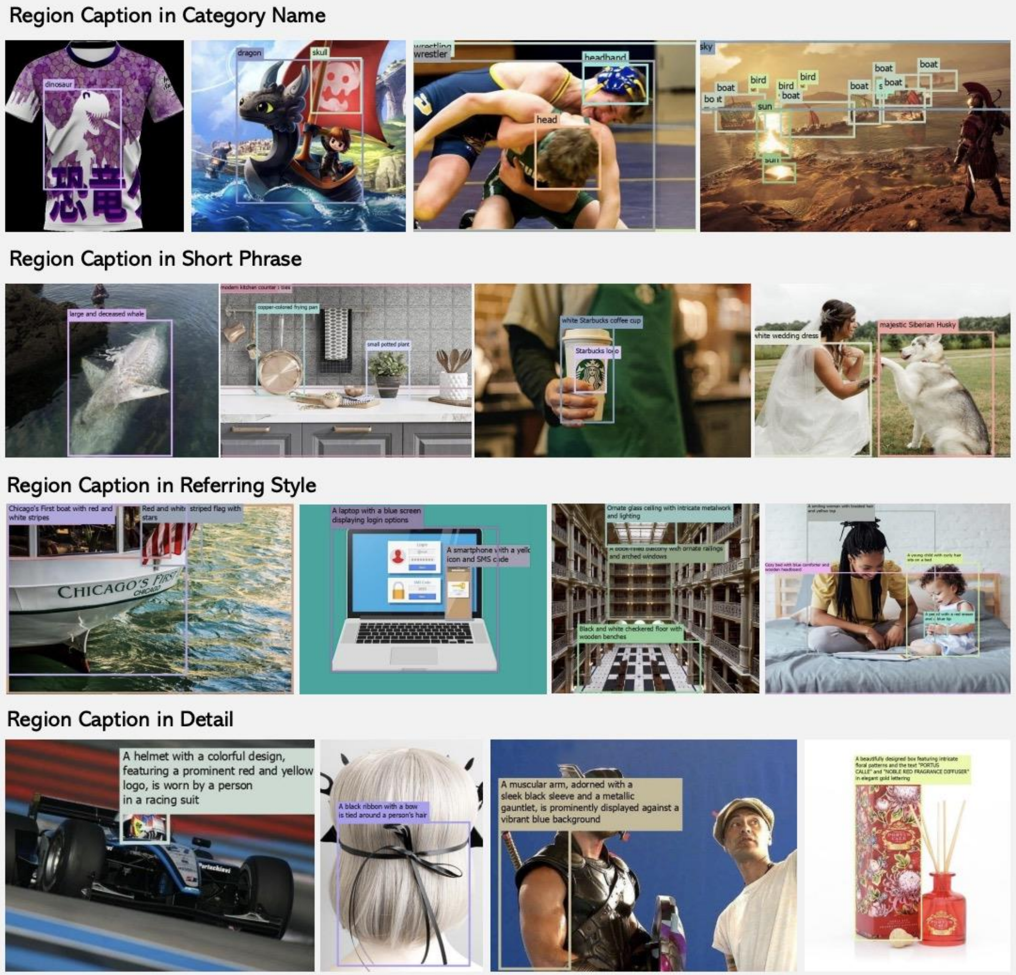


Figure 11: Visualization on Detailed Grounded Image Caption Task.

Applications: Grounded Conversation

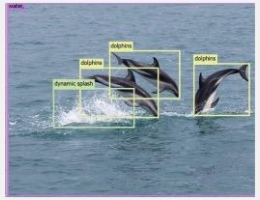
QA Example:

User: Please briefly describe this image and detect all the mentioned objects. Answer with grounded object indexes.

ChatRex: A <g>man</g><o><obj13></o> in a <g>white tuxedo</g><o><obj17></o> with a <g>red bow tie</g><o><obj16></o> is holding an <g>Oscar statuette</g><o><obj14></o> and standing on a stage with a microphone, while a large, ornate Oscar statue is visible in the background.

Visualization:

A man in a white tuxedo with a red bow tie is holding an Oscar statuette and standing on a stage with a microphone, while a large, ornate Oscar statue is visible in the background



Three dolphins are leaping out of the water, creating a dynamic splash as they dive and play in the ocean.

A person is slicing fresh cilantro a wooden cutting board, surrounded by various ingredients including a bowl of chopped tomatoes, a bowl of grated cheese a bowl of chopped herbs, a whole avocado and a bowl of grated cheese.

The image is a vintage advertisement for Fitzgerald & Co., featuring a hand pouring a bottle of whiskey into a crescent moon with the text "Fitzgerald & Co. Dublin" prominently displayed.

A woman with glasses is smiling and holding a white envelope with a handwritten address , standing in a cozy room with a window and a bookshelf filled with books in the background.

A woman with long brown hair is enjoying a cone of ice cream on a beach, with a backdrop of waves and seagulls in flight.

The image showcases a variety of ingredients laid out on a white surface, including a plate of raw beef, a bowl of sliced onions, a plate of fresh green lettuce, a plate of sliced cheese, a bowl of nuts, and a plate of bread.

The image features a charming black dog-shaped keychain with a textured surface, adorned with a pink tongue and a small brown tag attached to a brown leather strap, all set against a white background.

Figure 13: Visualization on Brief Grounded Image Caption Task.

QA Example:

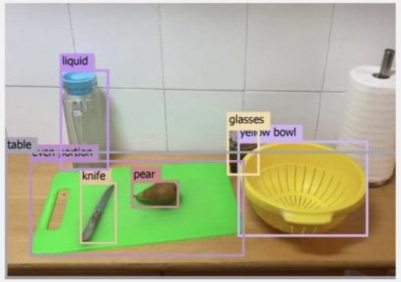
User: Answer the question in grounded format. How would you organize this space?

ChatRex: To organize the space, I would sort the items into categories such as stationery, electronics, personal items, and accessories. For stationery, I would group the <g>pen</g><o><obj23></o>, <g>pencil</g><o><obj12></o>...

Visualization:

Question: Answer the question in grounded format. How would you organize this space? Sort and arrange items on the surface based on their function or category, ensuring the area is tidy and visually pleasing.

ChatRex: To organize the space, I would sort the items into categories such as stationery, electronics, personal items, and accessories. For stationery, I would group the pen, pencil, and ruler together. Electronics would include the laptop, mouse, and keyboard. Personal items would be grouped with the notebooks, wallet, and keys. Accessories would be separated from the other categories and include the scissors, bobby pins, and tape. After sorting, the items would be arranged neatly on the surface, with similar items grouped together and the larger items placed at the edges to create a visually pleasing and organized space.



Question: Answer the question in grounded format. Use the tools available on the table to prepare the pear for consumption. Ensure any necessary cleaning, peeling, or cutting is performed appropriately, and leave the workspace tidy afterward.

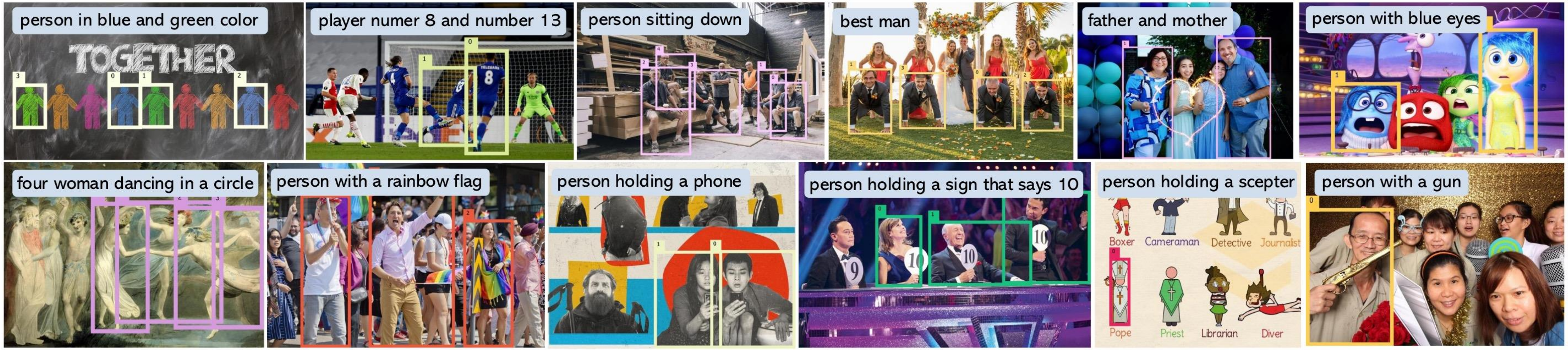
ChatRex: To prepare the pear for consumption, I would first ensure that the pear is clean and free of any dirt or blemishes. If necessary, I would use the knife to peel the pear, removing the skin in a smooth and even manner. After peeling, the pear would be placed in the yellow bowl for further preparation. If the pear is to be cut, the knife would be used to make the necessary cuts, ensuring that the pear is sliced into even portions. The glasses on the table could be used to rinse the pear under running water to remove any residual water or liquid. After the pear is prepared, the knife would be cleaned and stored properly, and the table would be tidied up to maintain a clean and organized workspace.

Figure 16: Visualization on Detailed Grounded Conversation Task.

Referring to Any Person

Referring to Any Person

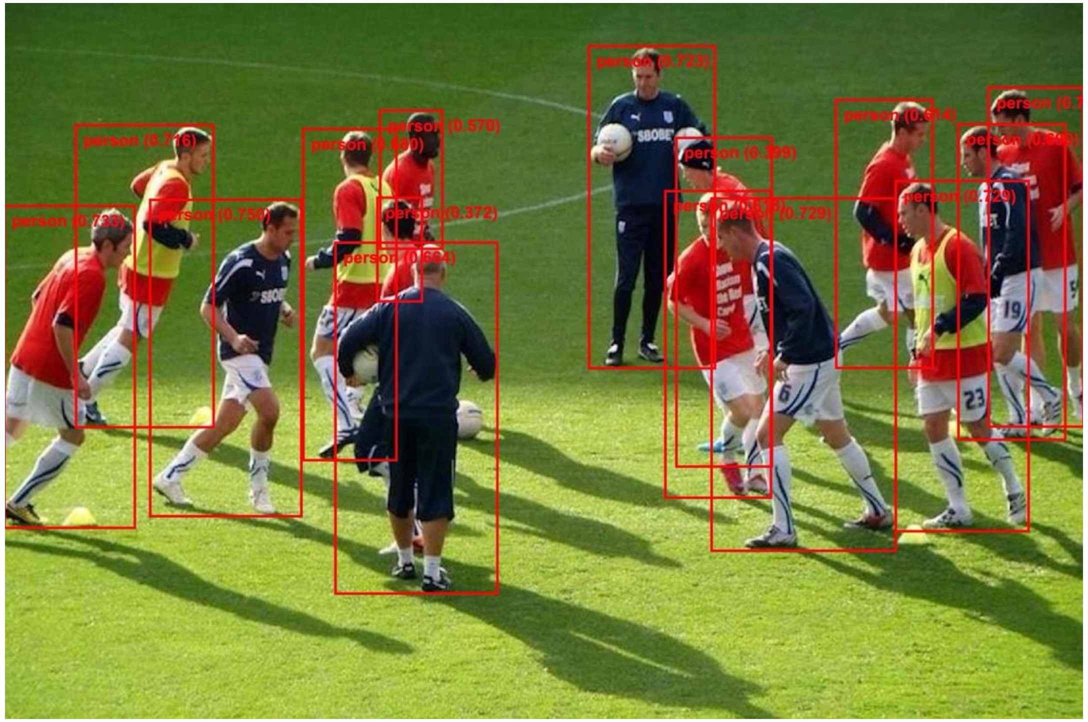
Qing Jiang[1,2], Lin Wu[1,2], Zhaoyang Zeng[1], Tianhe Ren[1], Yuda Xiong[1]
Yihao Chen[1], Liu Qin[1], Lei Zhang[1,2†]
[1]International Digital Economy Academy (IDEA)
[2]South China University of Technology
mountchicken@outlook.com , leizhang@idea.edu.cn



Jiang Q, Wu L, Zeng Z, et al. Referring to Any Person[J]. arXiv preprint arXiv:2503.08507, 2025.

Referring V.S. Detection

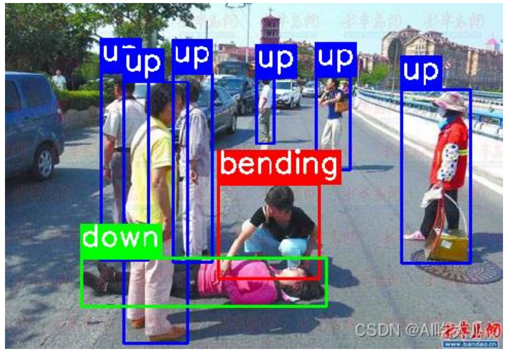
idea

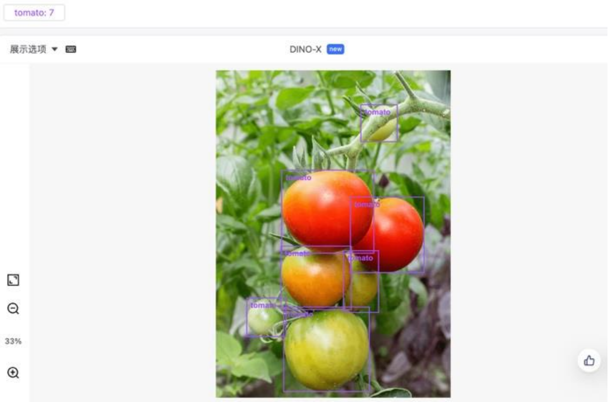


Detection: "person"

Referring: "person who is holding two footballs"

Most Detection Tasks Can be formulated as Referring





摔倒检测

"person fallen"

佩戴安全帽检测

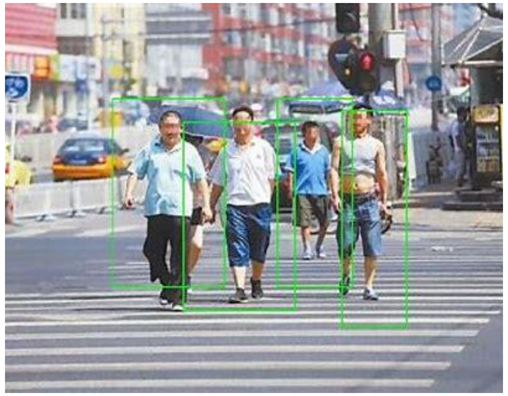
"person that are not wearing helmet"

工位睡觉检测

"person that is sleeping"

智慧农业

"tomato that are not ripe"



行人安全检测

"person on the crossroad"

抽烟检测

"person that are smoking"

交通管理

"cars that are crushed"

Referring V.S. Detection

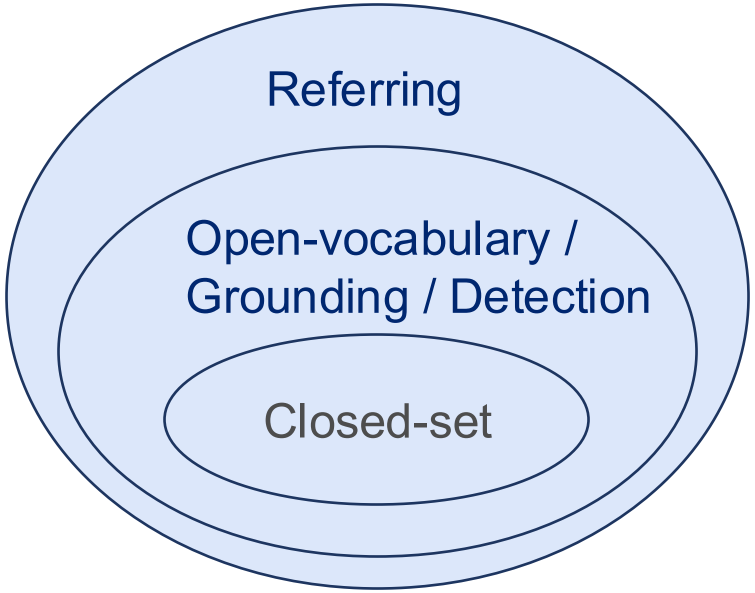
Detection: Category name e.g. man

Referring: Category name +

attributes	color	material	gender	age	wearing glasses
position	left	right	right	on a table	next to someone
affordance	cut	cook	fill water		
action	standing	smiling	running		

E.g.

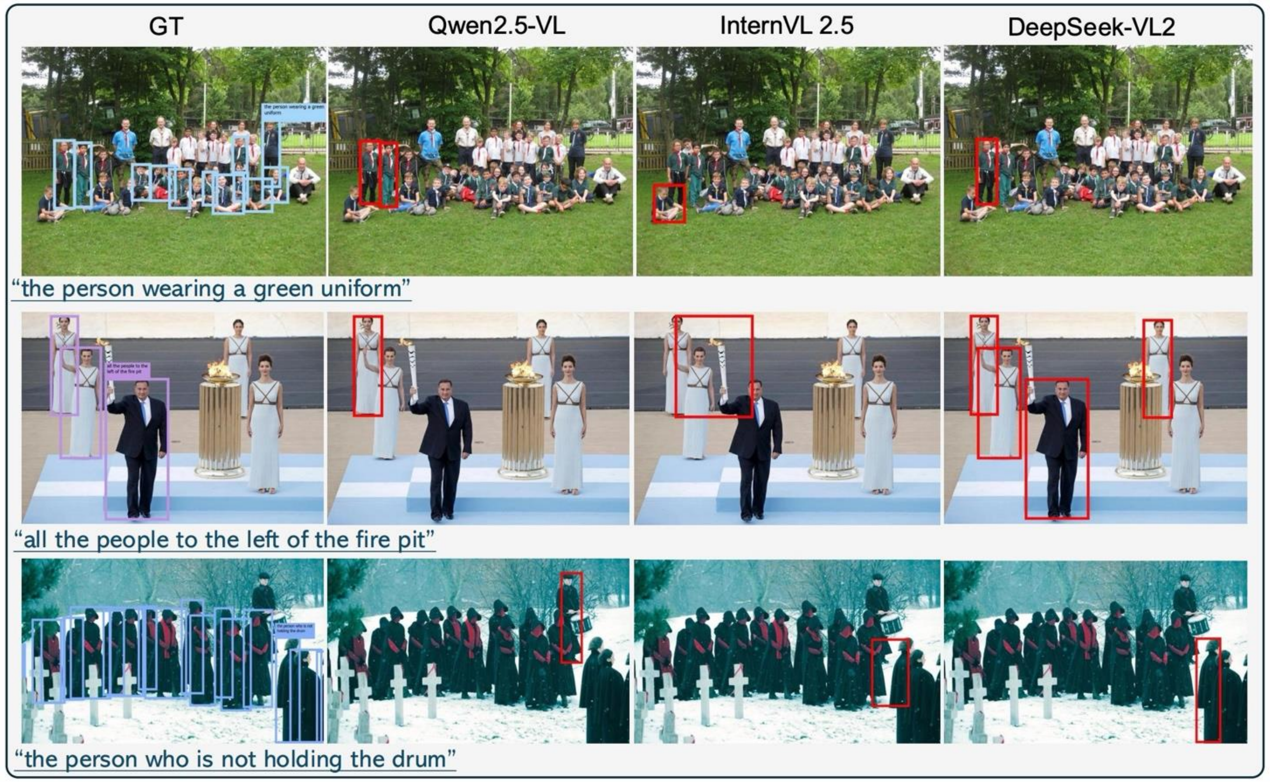
- a white man
- the second white man from the left
- The second white man from the left that is wearing a blue hat
- The second white man from the left that is wearing a blue hat and is smiling



Motivation: Current SOTA models lack usability

Datasets	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B
Refcoco$_{val}$	93.7	92.7	90.0
Refcoco$_{testA}$	95.6	94.6	92.5
Refcoco$_{testB}$	92.5	89.7	85.4
Refcoco+$_{val}$	90.4	88.9	84.2
Refcoco+$_{testA}$	94.7	92.2	89.1
Refcoco+$_{testB}$	86.9	83.7	76.9
Refcocog$_{val}$	92.7	89.9	87.2
Refcocog$_{test}$	92.2	90.3	87.2

High Performance in **existing benchmarks**



Low Performance in real-world scenarios

1. Designing flaws in existing benchmarks
2. Current MLLMs are still less capable

Solutions: Data + Model

HumanRef Dataset

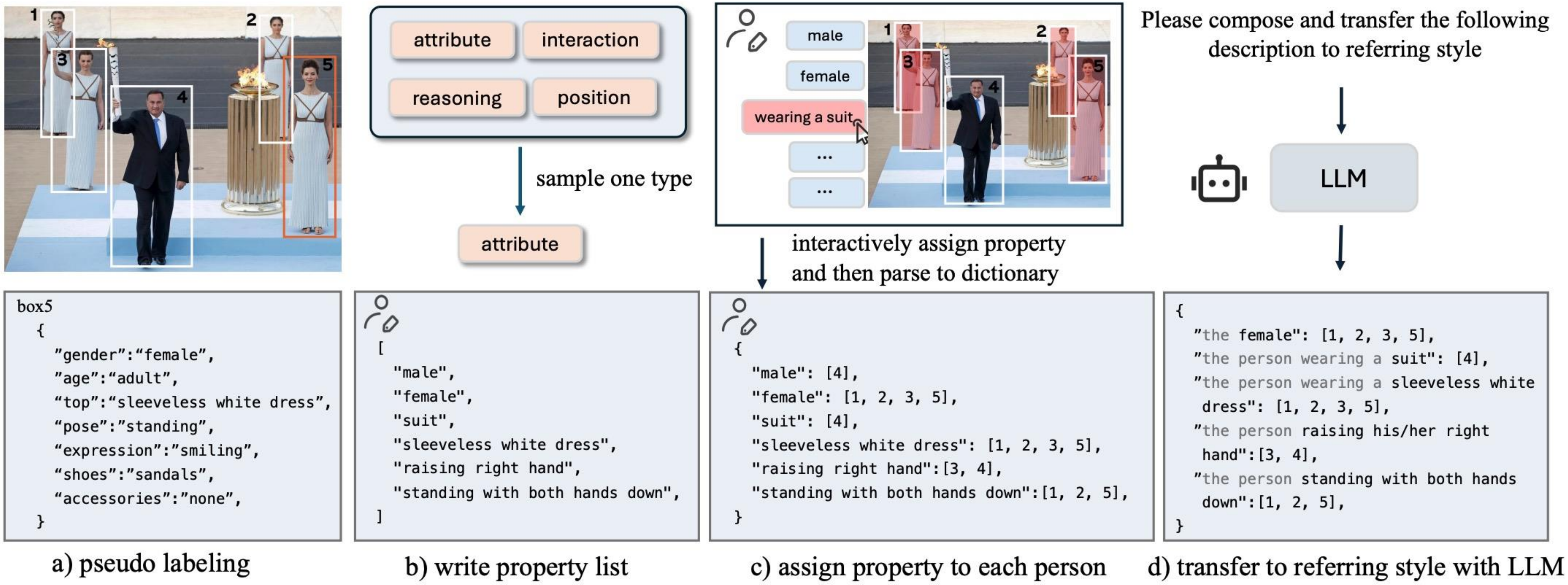


Figure 3. Overview of the mannual annotation pipeline of the HumanRef dataset.

Solutions: Data

HumanRef Dataset V.S. RefCOCO/+/g

Datasets	images	refs	vocabs	avg. size	avg. person/image	avg. words/ref	avg. boxes/ref
RefCOCO [75]	1,519	10,771	1,874	593x484	5.72	3.43	1
RefCOCO+ [75]	1,519	10,908	2,288	592x484	5.72	3.34	1
RefCOCOg [50]	1,521	5,253	2,479	585x480	2.73	9.07	1
HumanRef	5,732	6,000	2,714	1432x1074	8.60	6.69	2.2

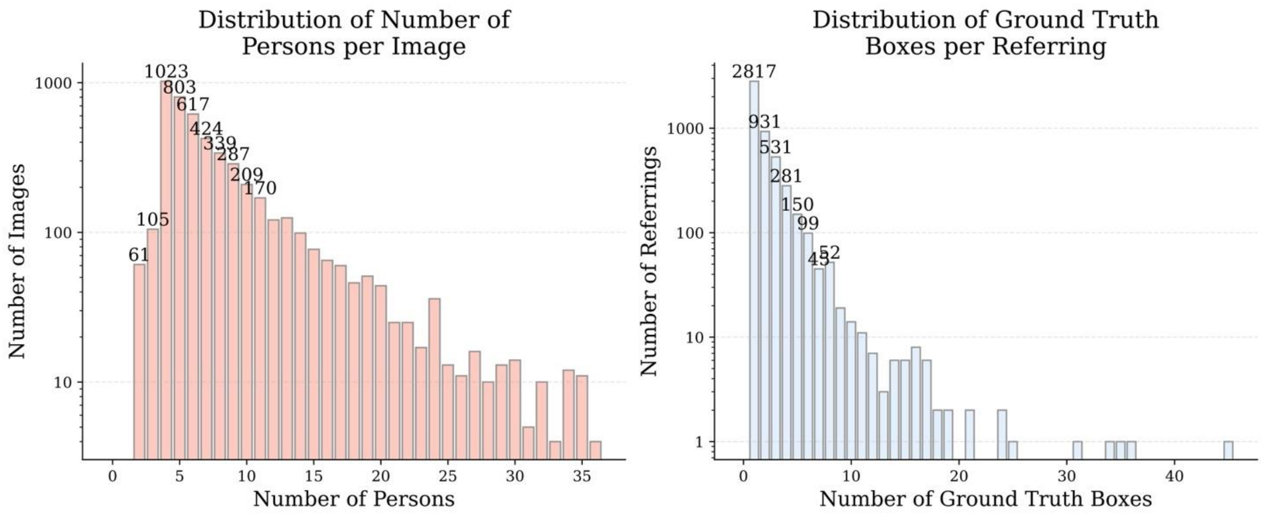


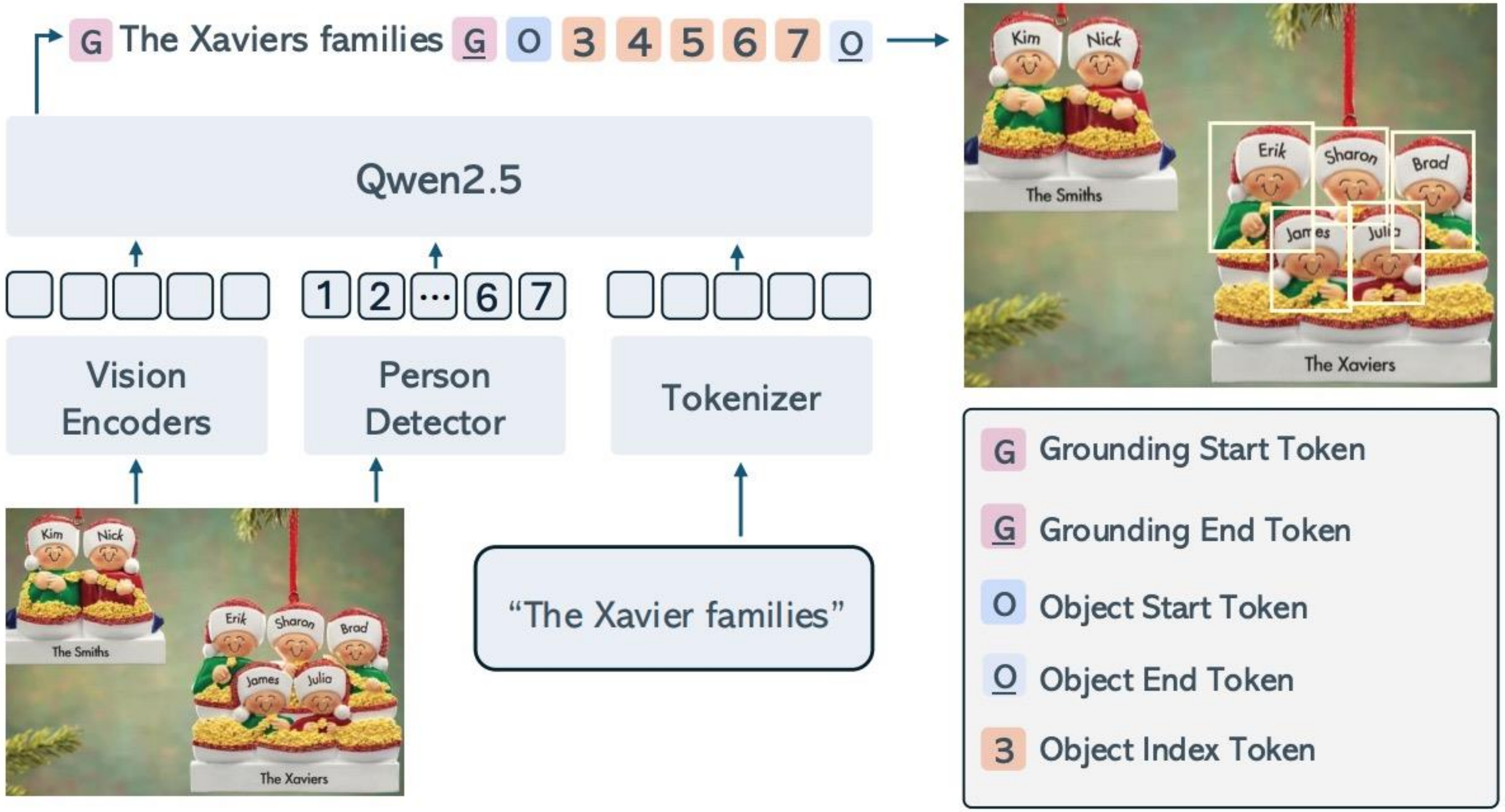
Figure 5. Distribution of the number of individuals per image and the number of individuals referenced by each referring expression.

1. Multi-Instance Referring

2. Multi-Instance Discrimination

Solutions: Model

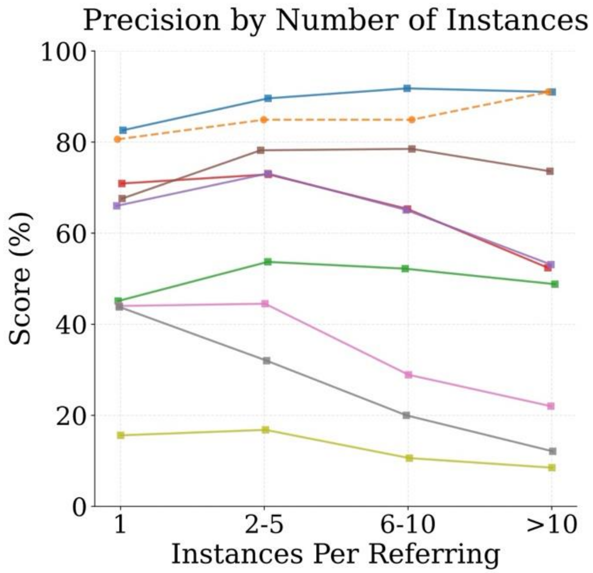
RexSeek

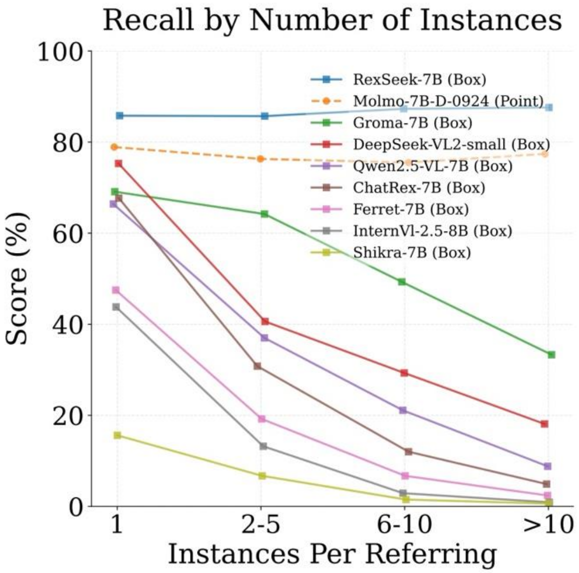


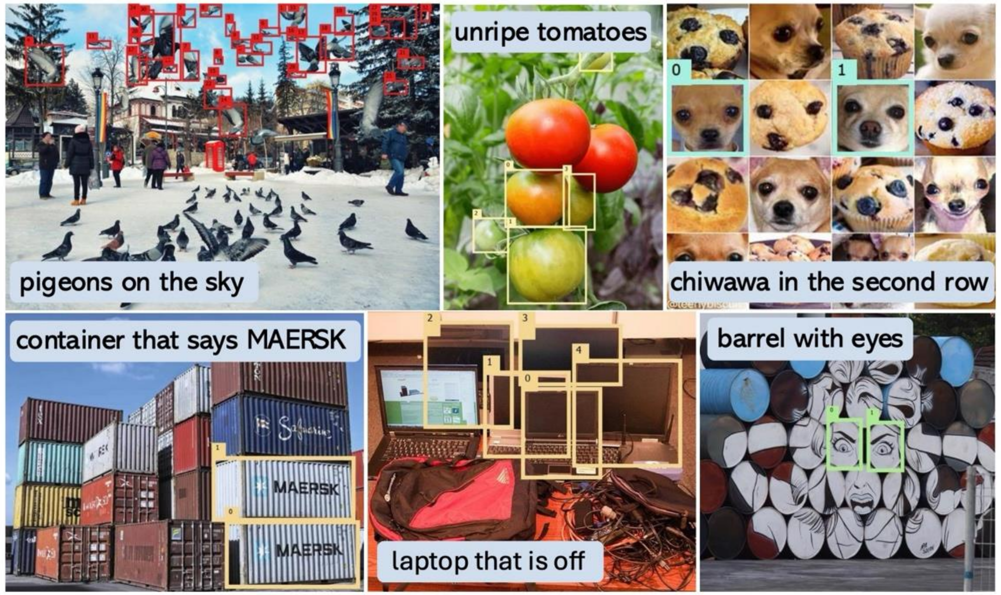
1. Strong perception capability
2. Strong language comprehension

Experiments

Method	Property			Position			Interaction			Reasoning			Celebrity			Average		
	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1
Baseline†	100.0	37.2	24.2	100.0	28.5	15.9	100.0	32.5	19.4	100.0	42.6	30.3	100.0	14.4	4.9	100.0	31.0	18.9
DINOX [60]	59.5	28.8	20.9	78.8	28.1	17.6	67.3	28.5	18.9	76.2	32.1	22.2	94.1	48.0	37.0	75.2	33.1	23.3
InternVL-2.5-8B [14]	23.5	39.0	27.1	23.0	28.0	24.3	27.8	40.1	31.3	17.5	22.8	18.9	57.4	59.3	58.0	29.8	37.8	31.9
Ferret-7B [74]	27.9	44.4	30.4	30.2	36.2	29.8	30.8	41.8	31.2	19.7	33.7	22.8	63.2	60.0	57.5	34.4	43.2	34.3
Groma-7B [49]	67.5	47.8	38.6	63.2	43.1	37.2	66.6	48.1	40.6	59.1	41.4	34.8	73.2	63.3	59.1	65.9	48.7	42.1
ChatRex-7B [25]	44.3	78.0	51.8	48.0	66.7	52.5	49.6	74.8	56.5	36.6	65.1	42.8	73.7	76.5	74.2	50.4	72.2	55.6
Qwen2.5-VL-7B [3]	49.1	71.3	54.4	50.2	61.7	52.8	48.2	66.3	53.2	34.6	61.2	40.3	80.3	81.9	80.1	52.5	68.5	56.2
DeepSeek-VL2-small [70]	52.3	78.0	57.7	56.4	66.1	58.1	55.4	75.7	60.7	46.6	61.7	50.1	85.9	74.3	70.7	59.3	71.2	59.5
Molmo-7B-D* [20]	82.7	86.4	76.3	78.0	80.6	72.4	69.9	77.7	66.1	72.1	80.4	65.5	85.9	87.5	82.9	77.7	82.5	72.6
RexSeek-7B	87.2	86.8	81.5	86.1	86.3	83.8	84.8	84.6	80.7	87.8	84.7	81.5	83.4	86.5	84.2	85.9	85.8	82.3



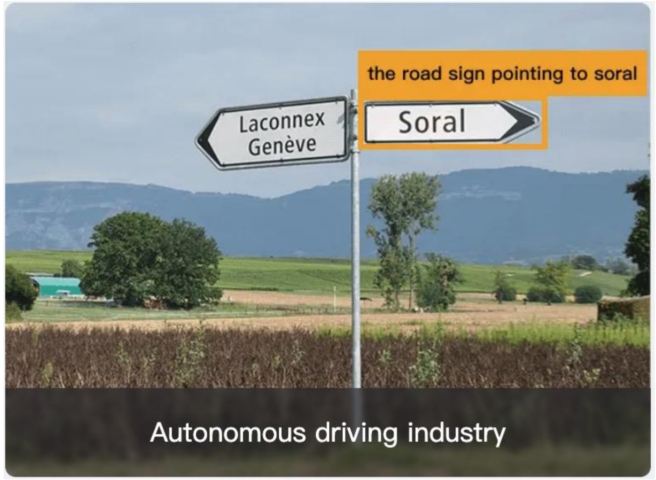


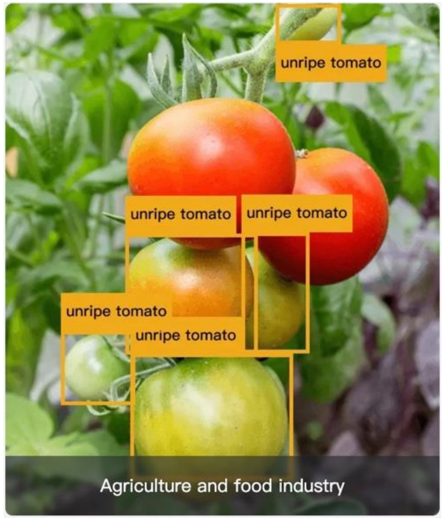


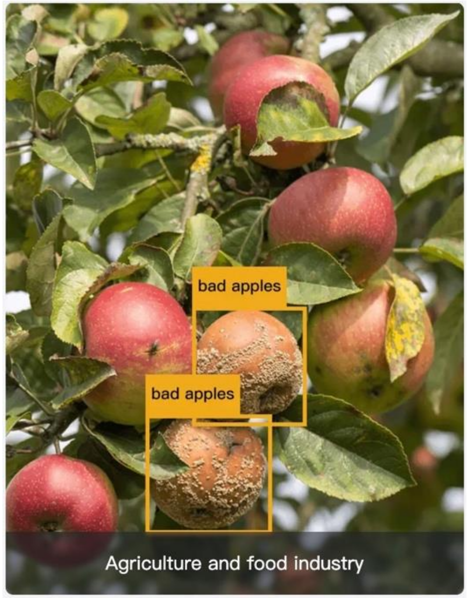
Generalize to Any Object

Applications: Referring Any Person

Applications: Industry Referring







Conclusions

- Current open-set detection models can only handle category-level openness; they still lack more advanced language understanding capabilities.
- Multimodal large language models possess strong understanding capabilities, but they lack fine-grained perception abilities.
- Integrating detection models with multimodal large language models enables a complementary approach

ChatRex:
Paper: https://arxiv.org/abs/2411.18363
Code: https://github.com/IDEA-Research/ChatRex

RexSeek:
Paper: https://arxiv.org/abs/2503.08507
Code: https://github.com/IDEA-Research/RexSeek
Demo: https://cloud.deepdataspace.com/playground/dino-x

Taming Multimodal LLM for Object Perception

Thanks!