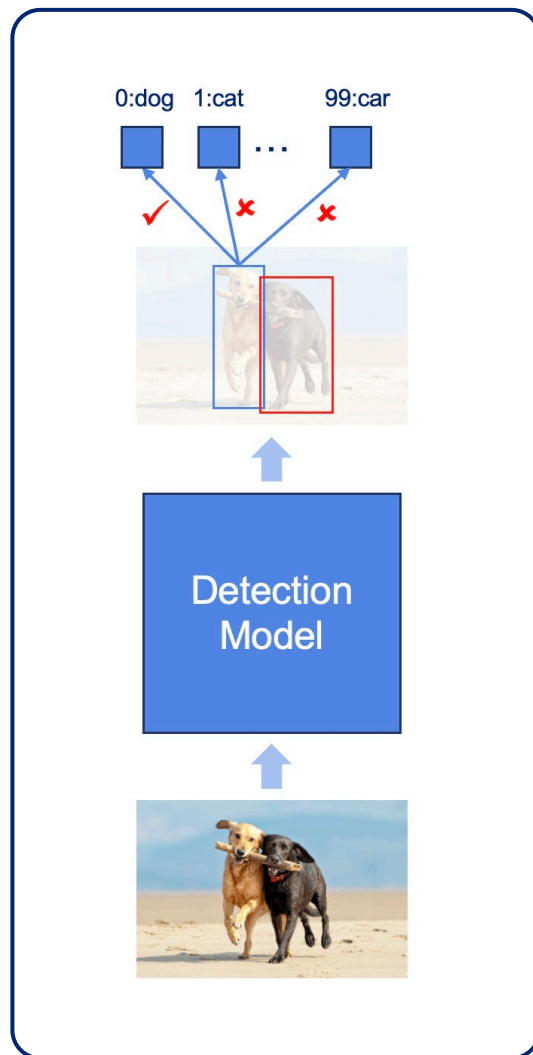


Taming Multimodal LLM for Object Perception

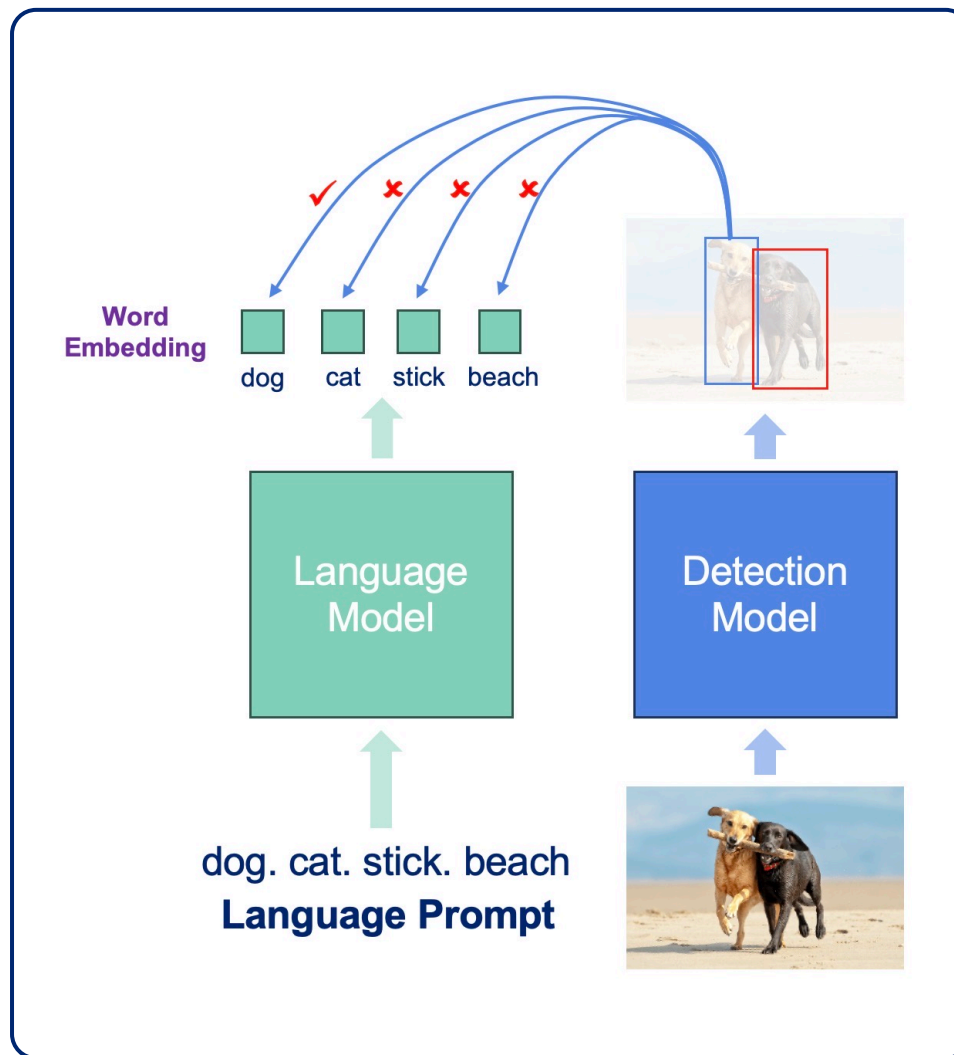
蒋擎

7-15

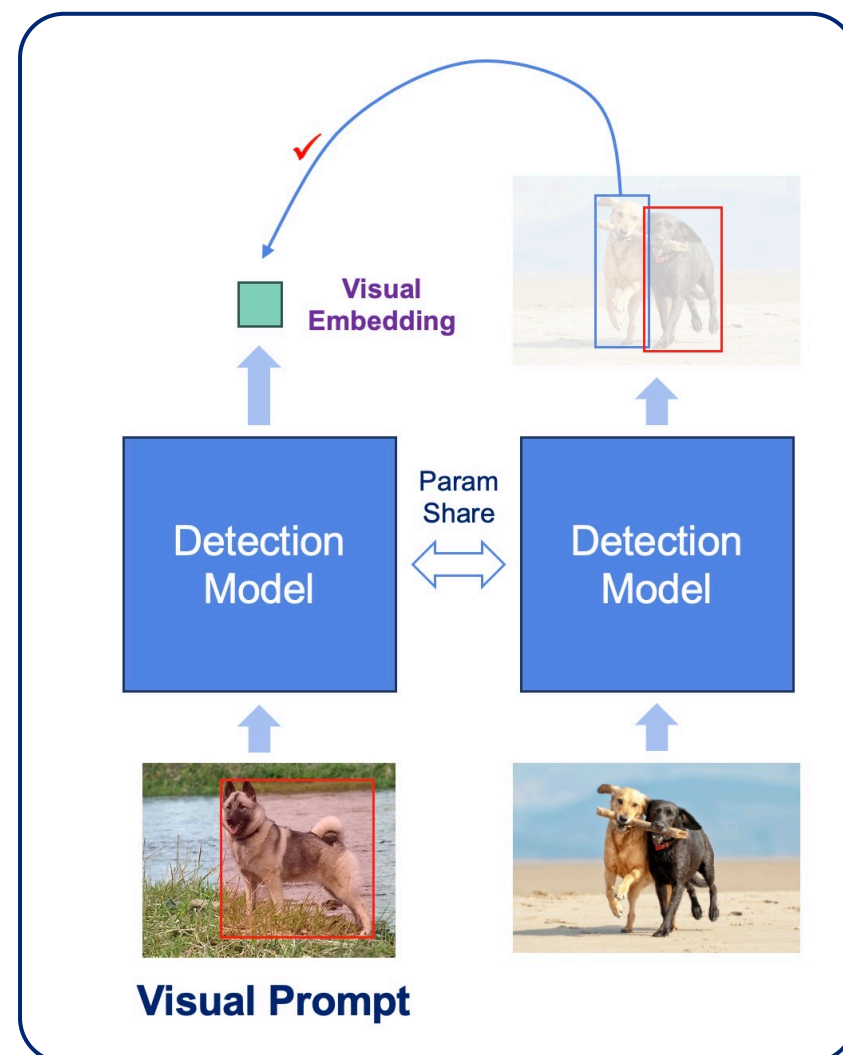
What is the Next Step for Object Detection?



Closed-Set
DETR



Open-vocabulary
Grounding DINO

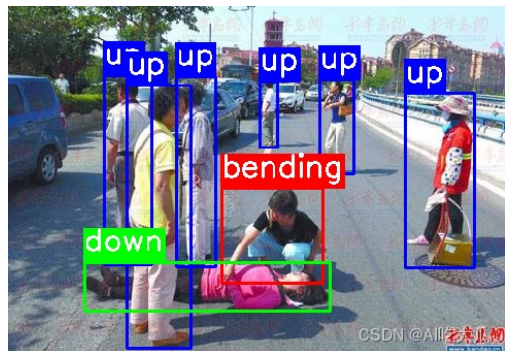


Visual Prompt
T-Rex2

What is the Next Step for Object Detection?

idea

most detection entities can be described in language



摔倒检测

“person fallen”



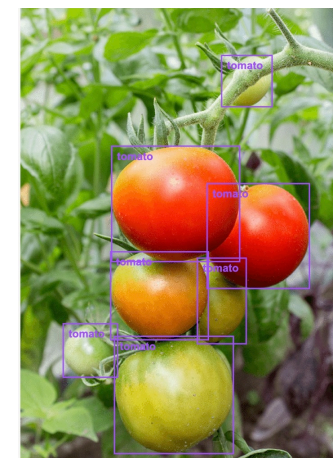
佩戴安全帽检测

“person that are not wearing helmet”



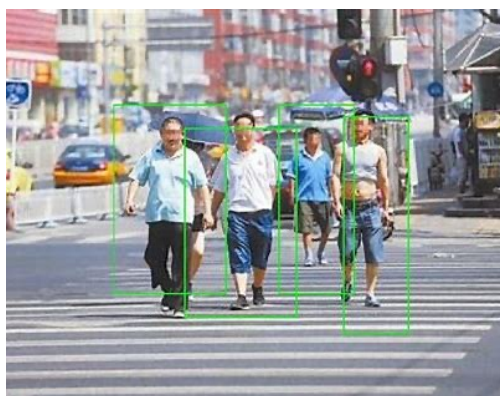
工位睡觉检测

“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”



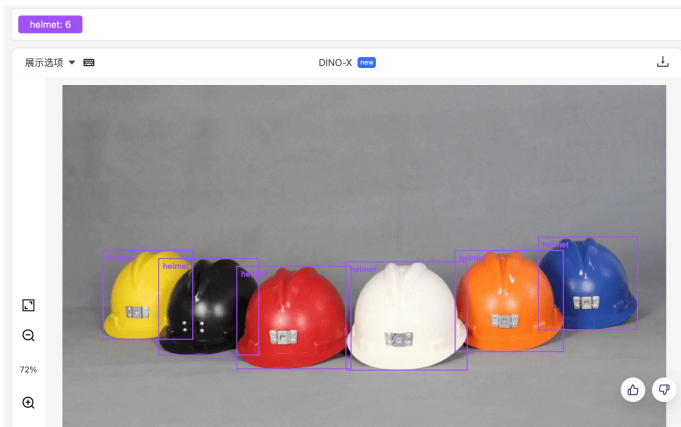
交通管理

“cars that are crushed”

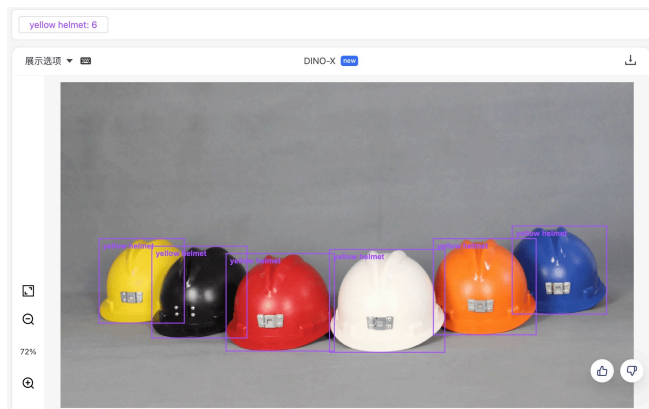
What is the Next Step for Object Detection?

Finding 1: State-of-the-art Open-set detection model (DINO-X) **lack language comprehension capabilities**

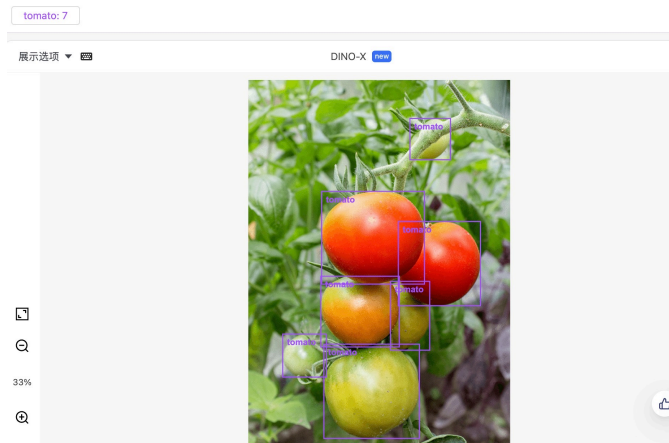
“helmet”



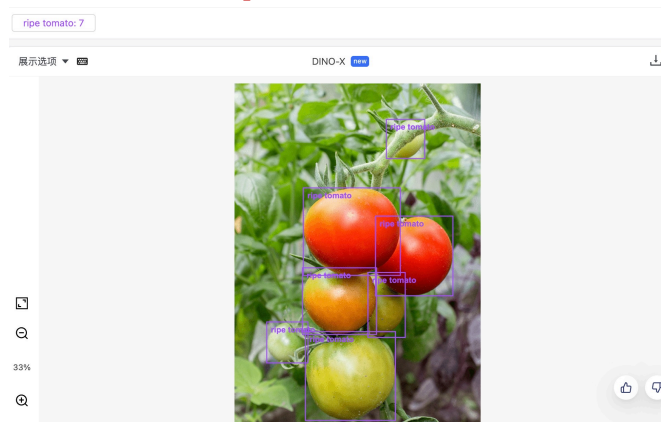
“yellow helmet”



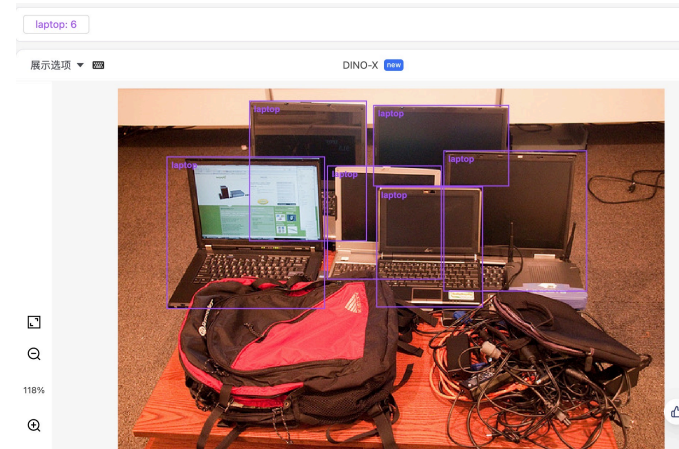
“tomato”



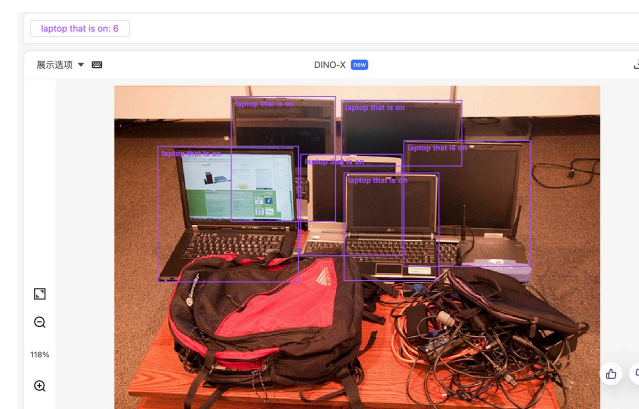
“ripe tomato”



“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**



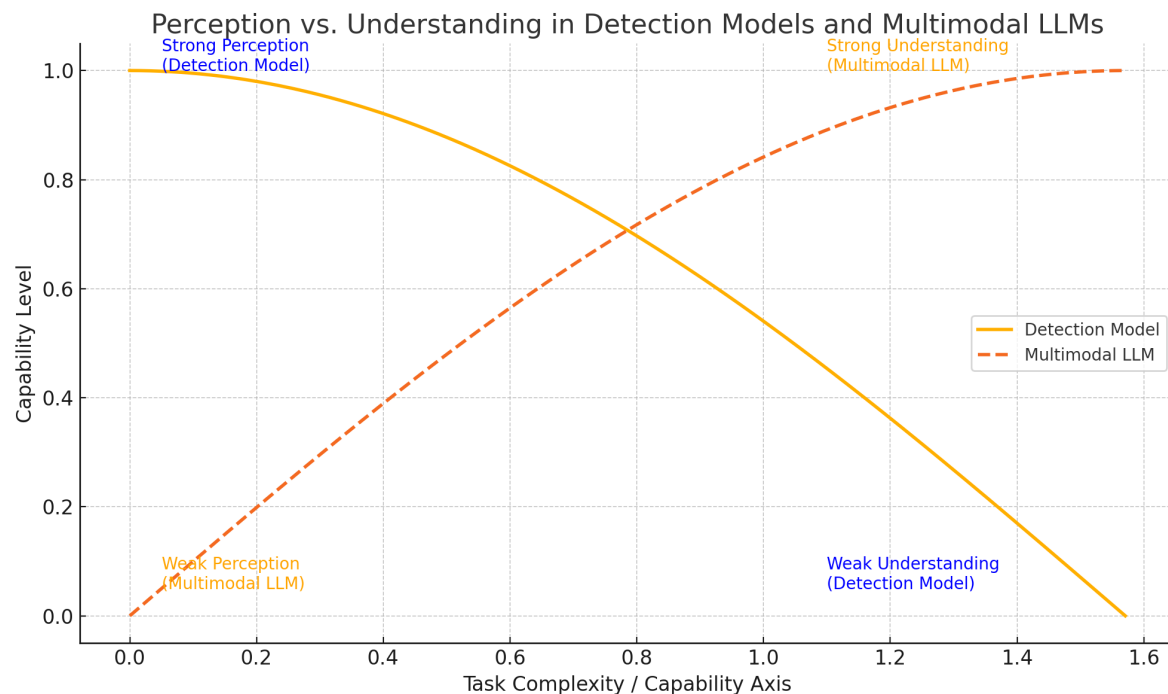
User: Please help me detect person in this image

MLLMs:

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

- coordinate shift
- tiny object detection
- dense object detection

What is the Next Step for Object Detection?



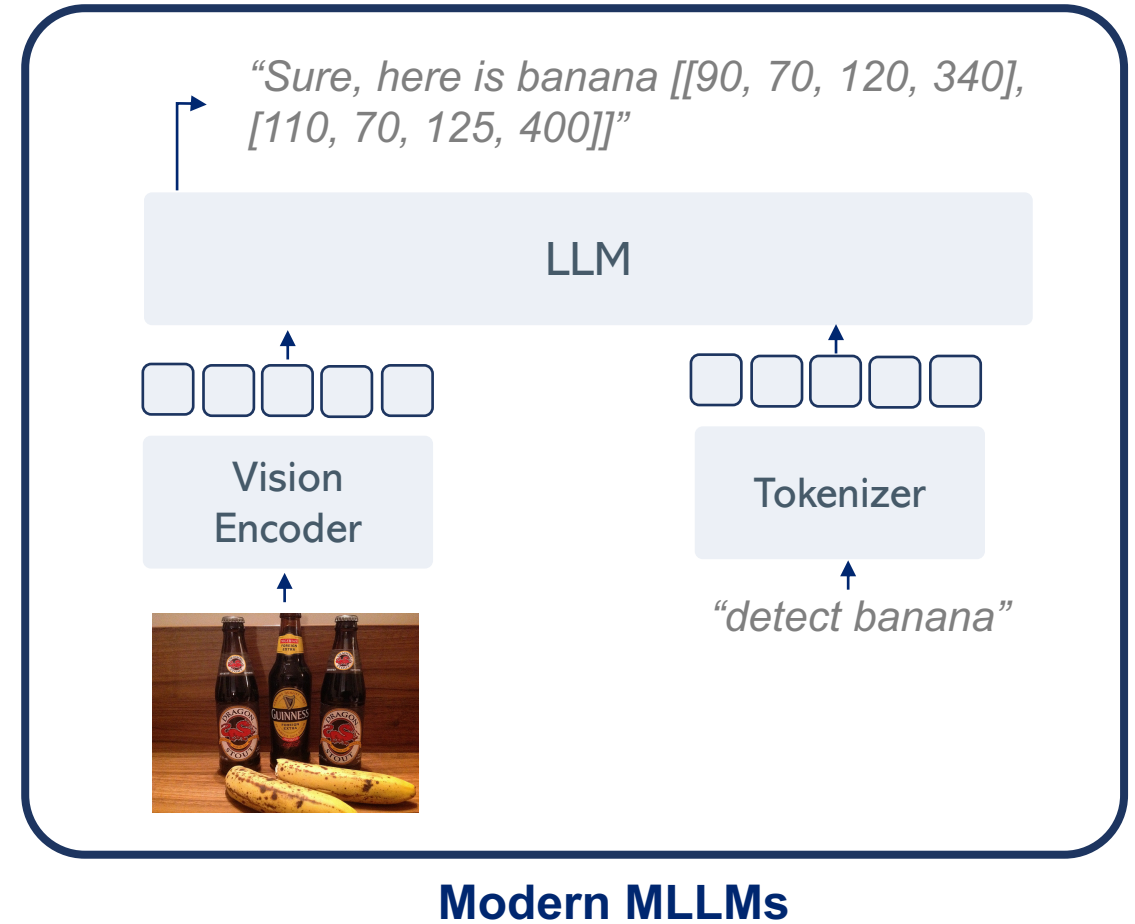
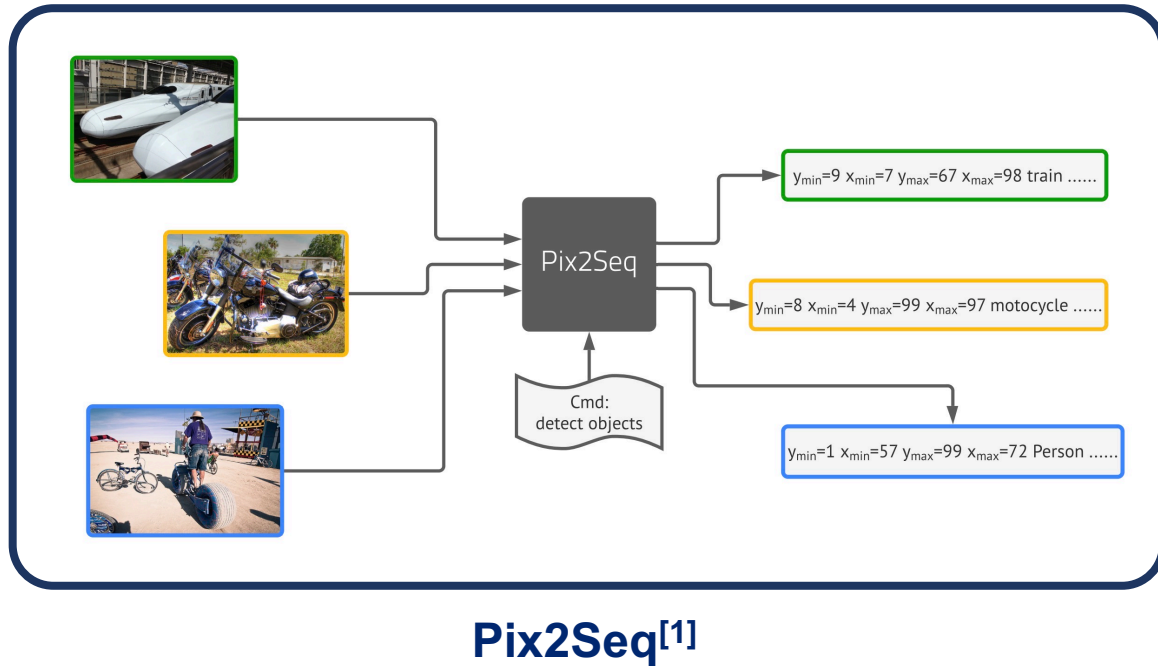
Detection Model: **Strong Perception**, **Weak Understanding**

Multimodal LLMs: **Weak Perception**, **Strong Understanding**

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

Motivation: How MLLMs do Detection?

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



[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

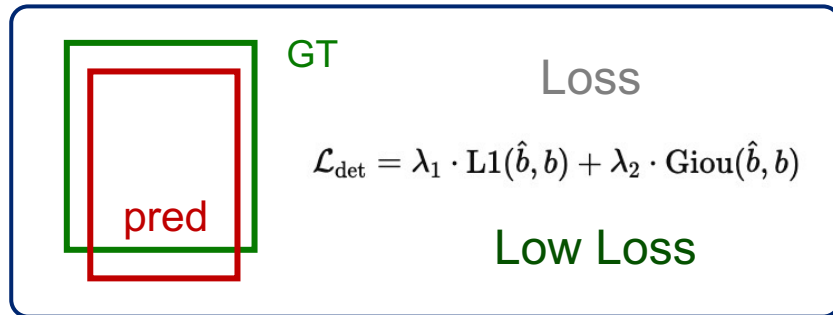


Method	Type	COCO-Val			LVIS-Mini Val					
		P@0.5	R@0.5	mAP	P@0.5	R@0.5	mAP	AP-R	AP-C	AP-F
Faster-RCNN [70]	Closed-set Detection Model	-	-	42.0	-	-	-	-	-	-
DETR [8]		-	-	43.3	-	-	-	-	-	-
Pix2Seq [12]		-	-	43.2	-	-	-	-	-	-
DINO [102]		-	-	49.4	-	-	-	-	-	-
Florence2 [88]	Open-set Detection Model	-	-	43.4	-	-	-	-	-	-
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

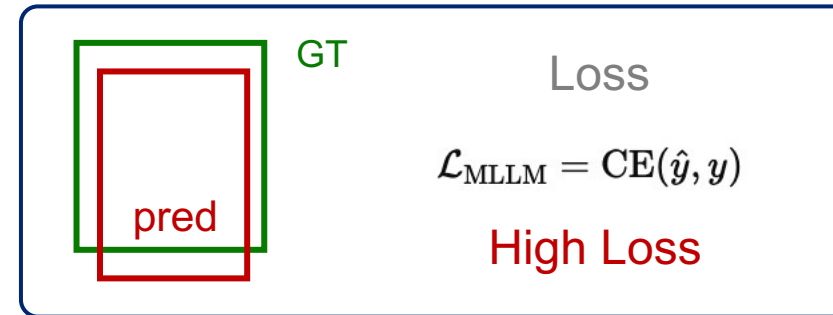
Low Recall Rate

Motivation: What's the Challenge?

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



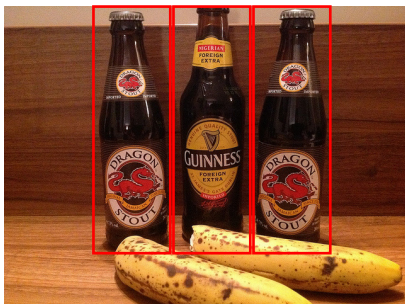
detection model training



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”

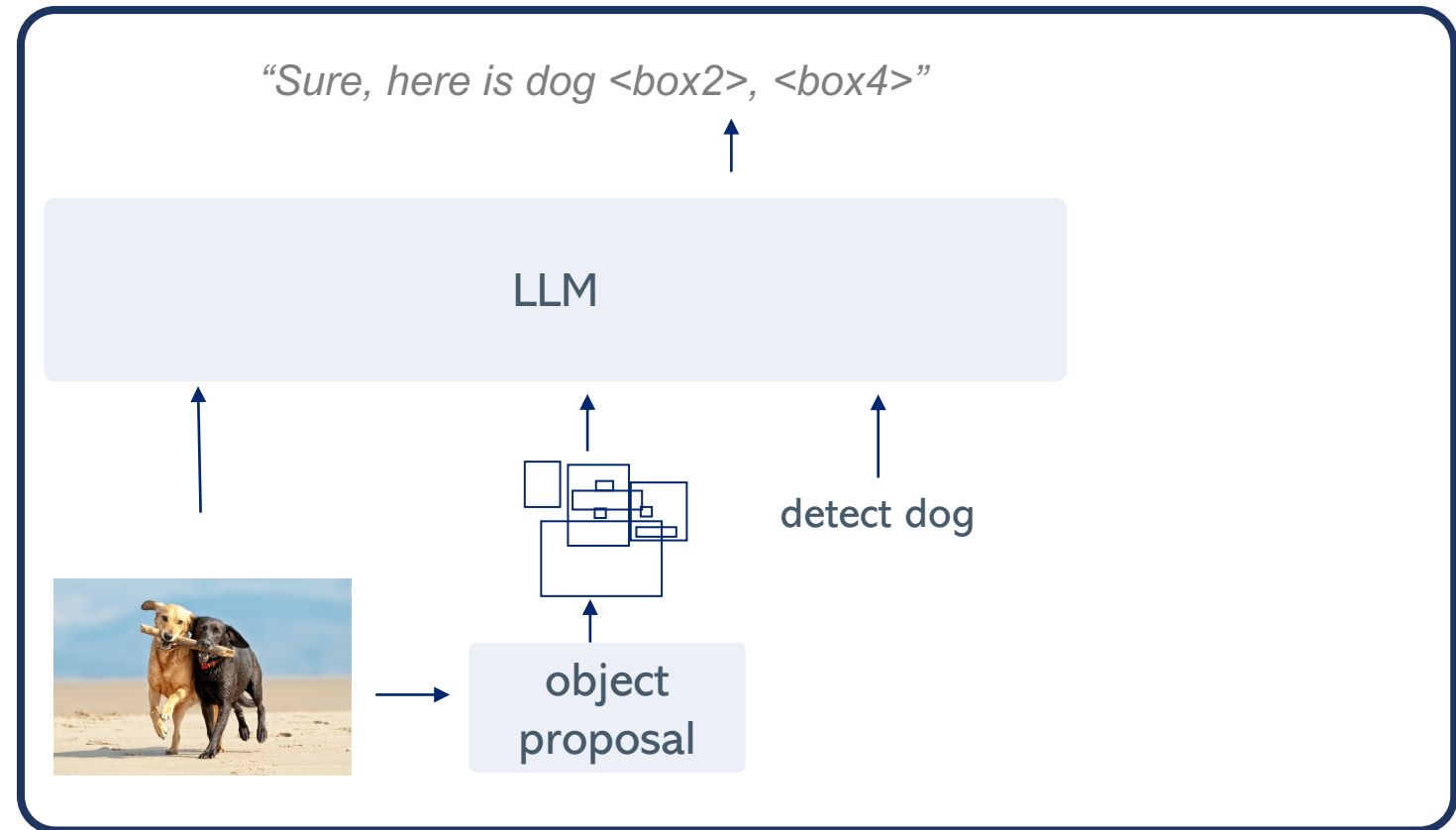
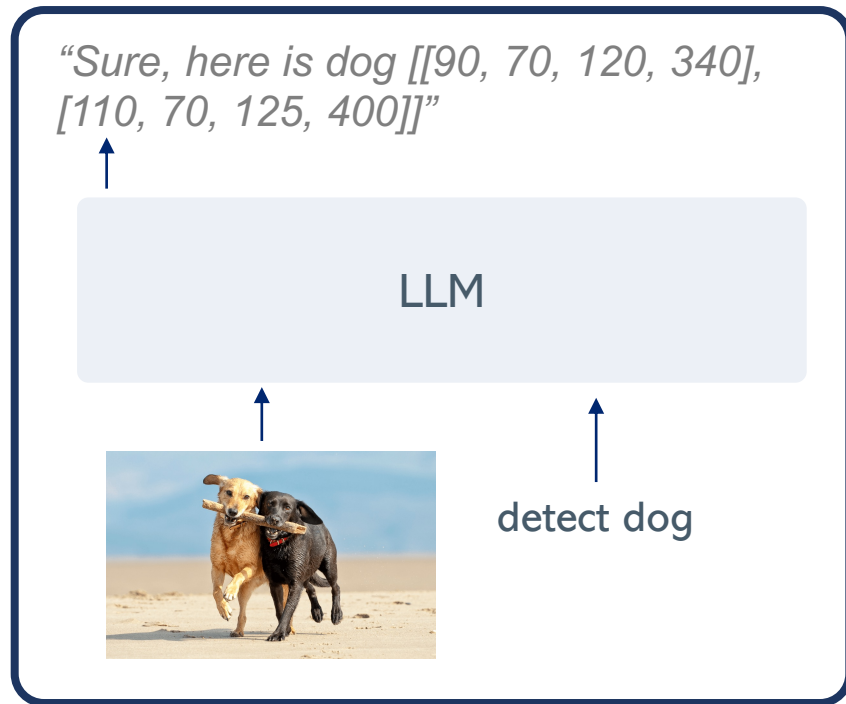
“bottle3, bottle2, bottle1”

“bottle2, bottle1, bottle3”

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

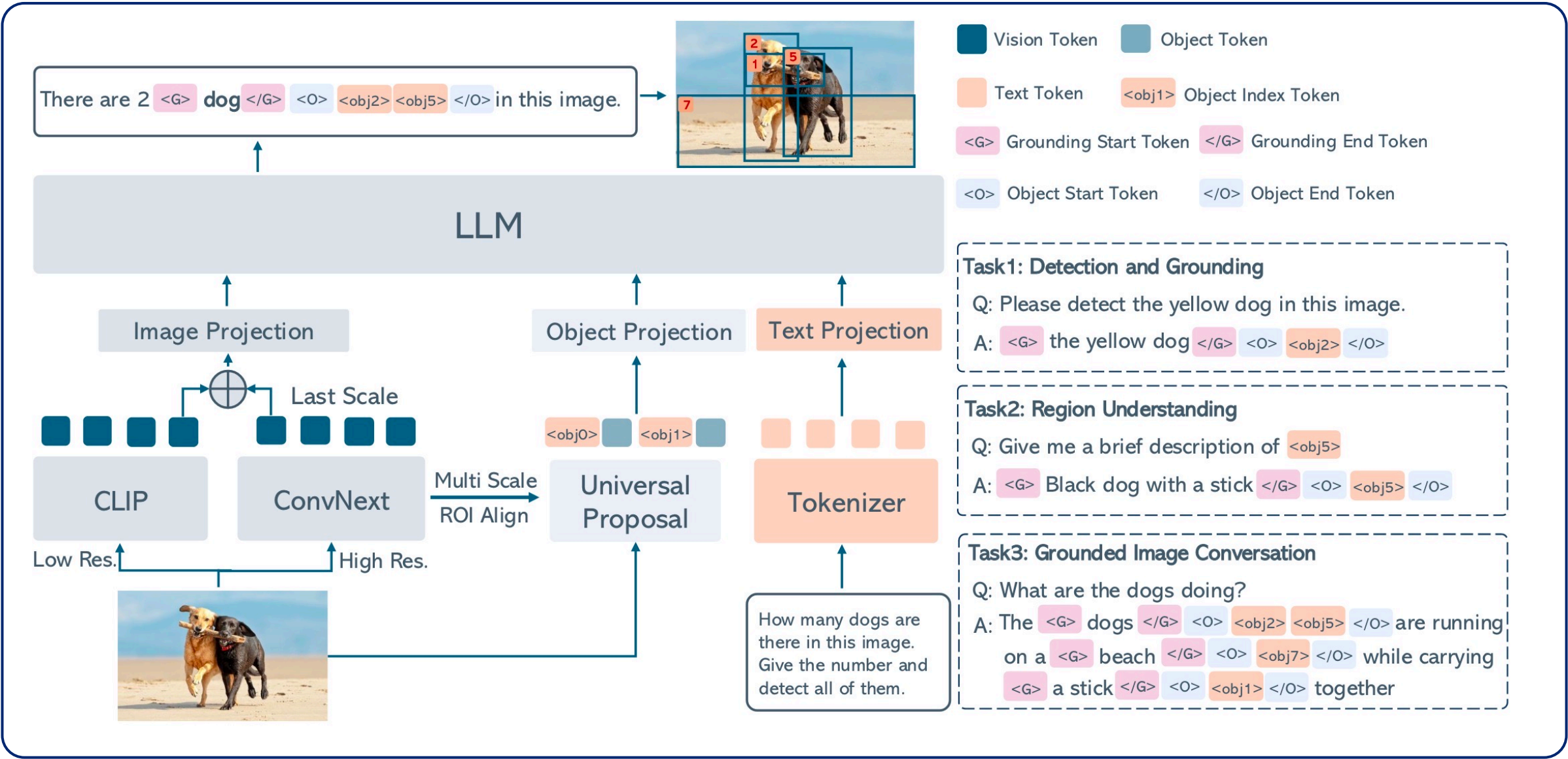
Solution: Retrieval-based Perception MLLM

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



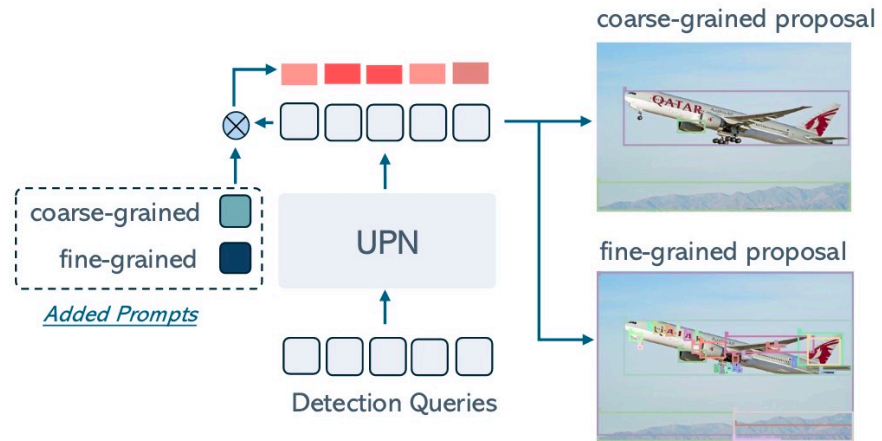
Retrieval based method

ChatRex: Detection-Oriented MLLM



ChatRex: Detection-Oriented MLLM

ChatRex Architecture



Universal Proposal Architecture



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

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>`

Visualization:

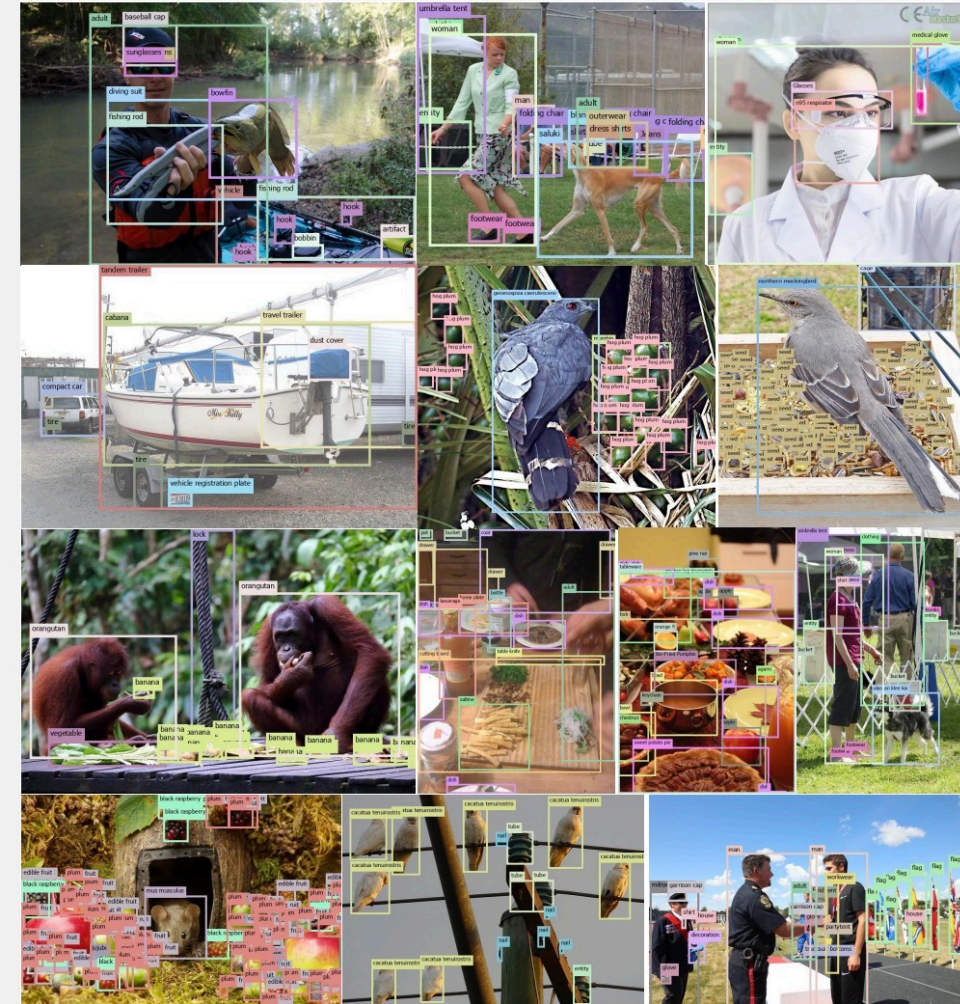


Figure 8: Visualization on Long-tailed Object Detection Task.

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:

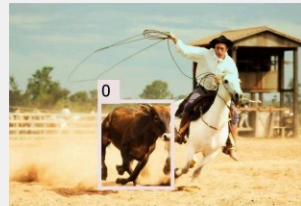
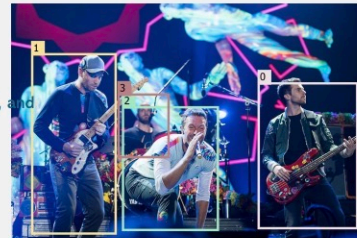


Question: What is <obj0> wearing and what is he doing now?

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

Question: What are the positions of <obj0>, <obj1>, <obj2>, and <obj3> in the band, respectively?

ChatRex: <obj0> is the bassist, <obj1> is the guitarist, <obj2> is the vocalist, and <obj3> is the drummer.



Question: What is <obj0>? And how might <obj0> feel now?

ChatRex: <obj0> is a cow, and it might feel scared or startled by the action of the man on the horse.

Question: What is the price for <obj0> according to this image.

ChatRex: The price for <obj0> is 2.49, as indicated by the sign in front of the cat.



Question: Describe the wearing of <obj0> in detail

ChatRex: The dress of <obj0> is a bright yellow, which stands out against the snowy backdrop. It appears to be a winter coat, designed to provide warmth in cold weather. The coat is long-sleeved and reaches down to the person's knees, suggesting it is made for cold temperatures. The coat's color is vibrant and contrasts with the muted colors of the snowy environment.

Figure 12: Visualization on Region QA Task.

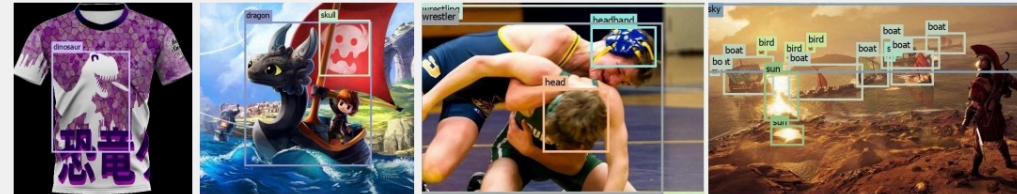
QA Example:

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:

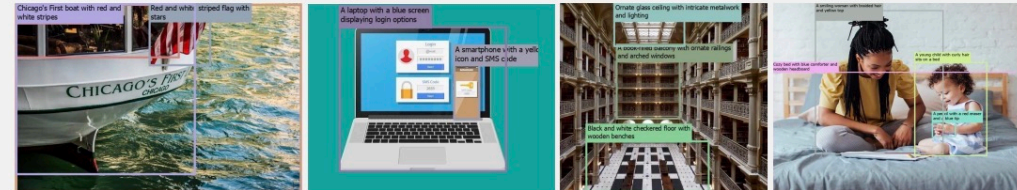
Region Caption in Category Name



Region Caption in Short Phrase



Region Caption in Referring Style



Region Caption in Detail

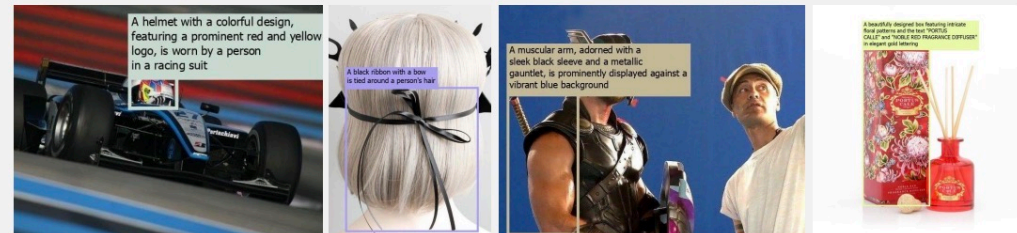


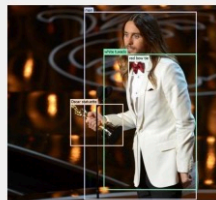
Figure 11: Visualization on Detailed Grounded Image Caption Task.

QA Example:

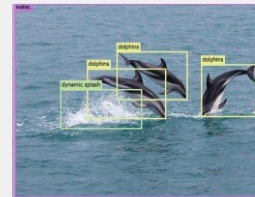
User: Please briefly describe this image and detect all the mentioned objects. Answer with grounded object indexes.

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.

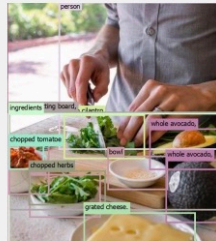
Visualization:



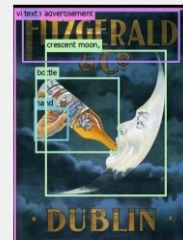
A man in a white tuxedo with a red bow tie is holding an Oscar statuette and standing on a stage with a microphone, while a large, ornate Oscar statue is visible in the background



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



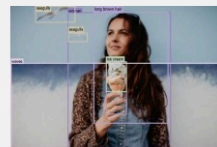
A person is slicing fresh cilantro a wooden cutting board, surrounded by various ingredients including a bowl of chopped tomatoes, a bowl of grated cheese, a bowl of chopped herbs, a whole avocado and a bowl of grated cheese.



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?

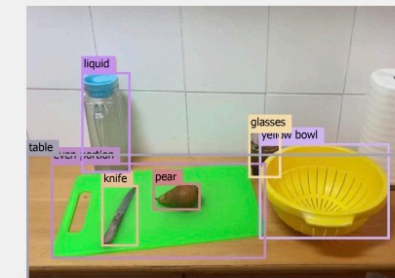
ChatRex: To organize the space, I would sort the items into categories such as stationery, electronics, personal items, and accessories. For stationery, I would group the `<g>pen</g><o><obj23></o>`, `<g>pencil</g><o><obj12></o>`...

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.

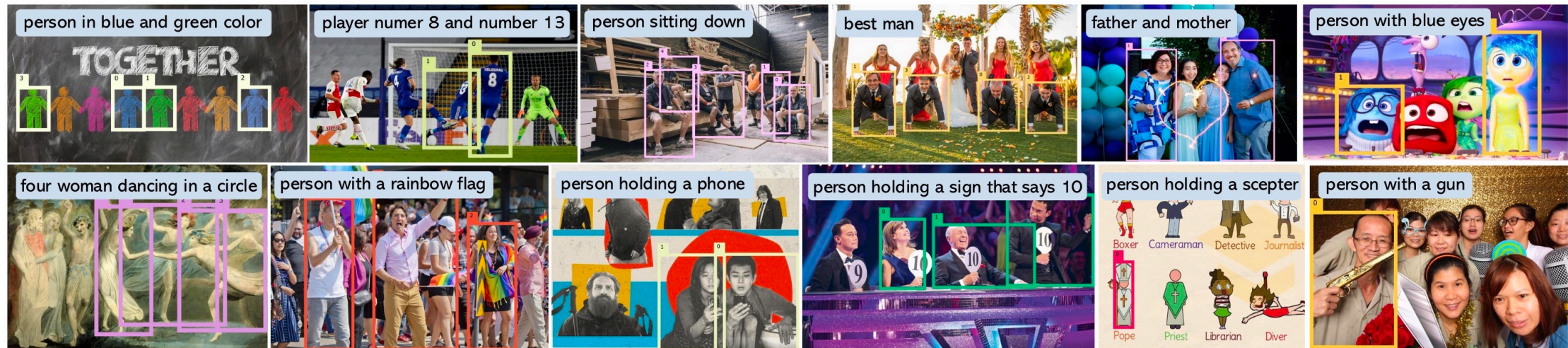
Referring to Any Person

Qing Jiang^{1,2}, Lin Wu^{1,2}, Zhaoyang Zeng¹, Tianhe Ren¹, Yuda Xiong¹
Yihao Chen¹, Liu Qin¹, Lei Zhang^{1,2†}

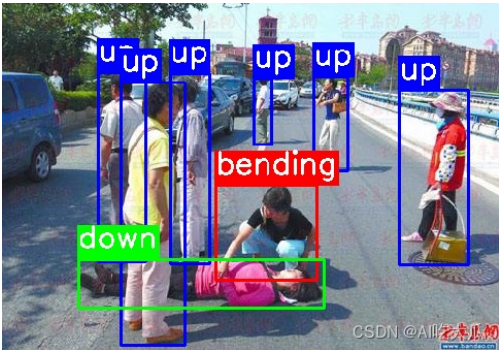
¹International Digital Economy Academy (IDEA)

²South China University of Technology

mountchicken@outlook.com, leizhang@idea.edu.cn



Most Detection Tasks Can be formulated as Referring



摔倒检测

“person fallen”



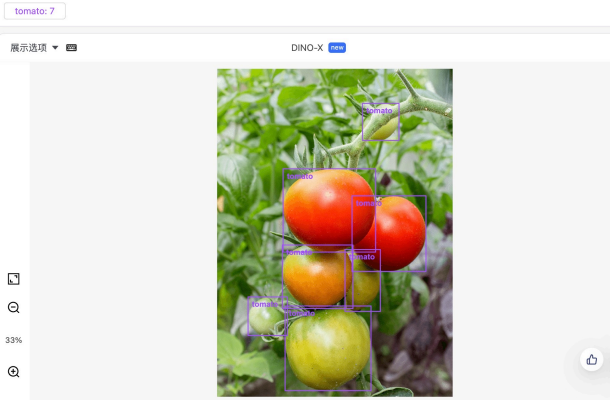
佩戴安全帽检测

“person that are not wearing helmet”



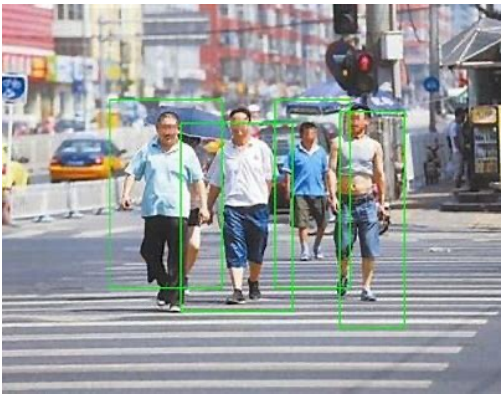
工位睡觉检测

“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

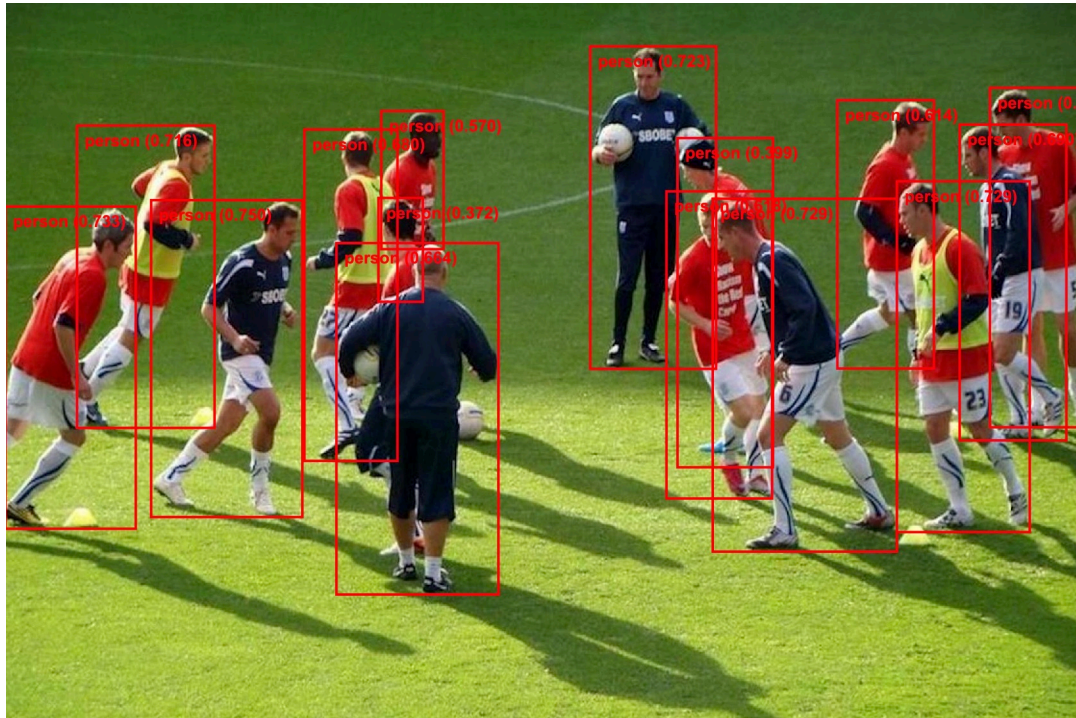
“person that are smoking”



交通管理

“cars that are crushed”

Referring V.S. Detection



Detection: “person”



Referring: “person who is holding two footballs”

Referring V.S. Detection

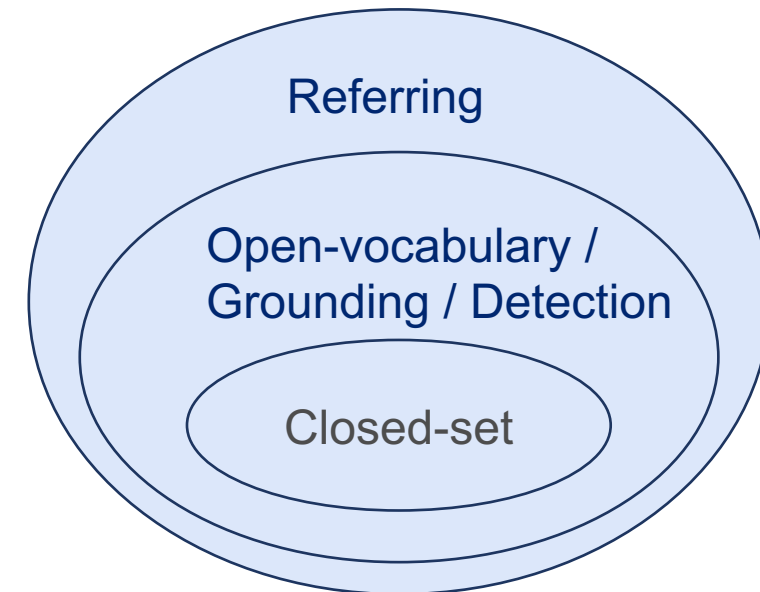
Detection: Category name e.g. man

Referring: Category name +



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



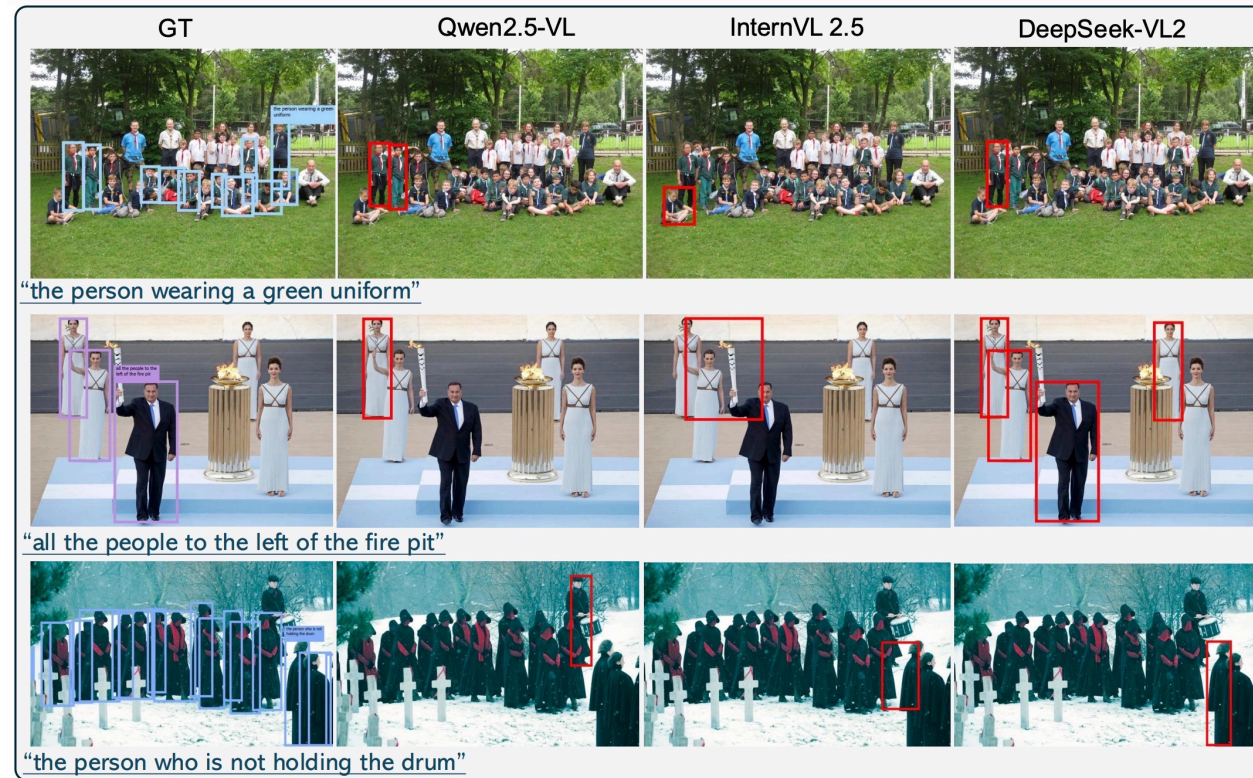
Motivation: Current SOTA models lack usability

Datasets	InternVL2.5	Qwen2.5-VL	Qwen2.5-VL
	78B	72B	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



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



Low Performance in real-world scenarios

HumanRef Dataset

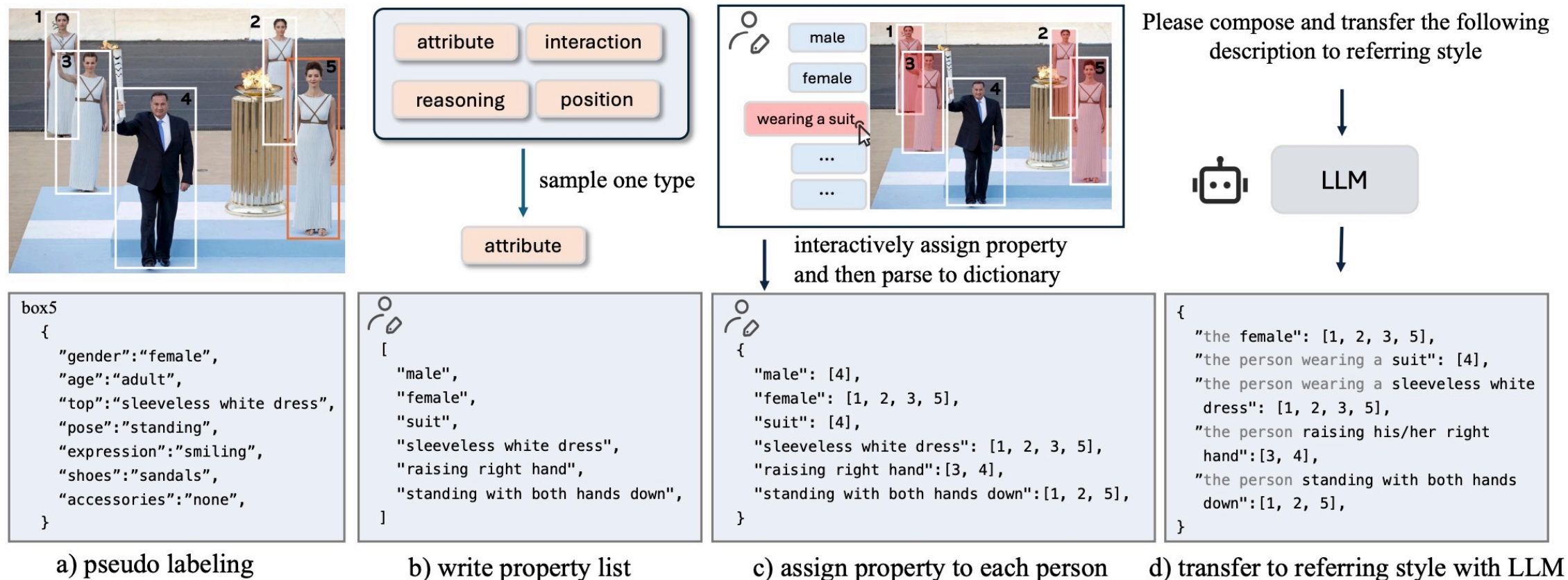
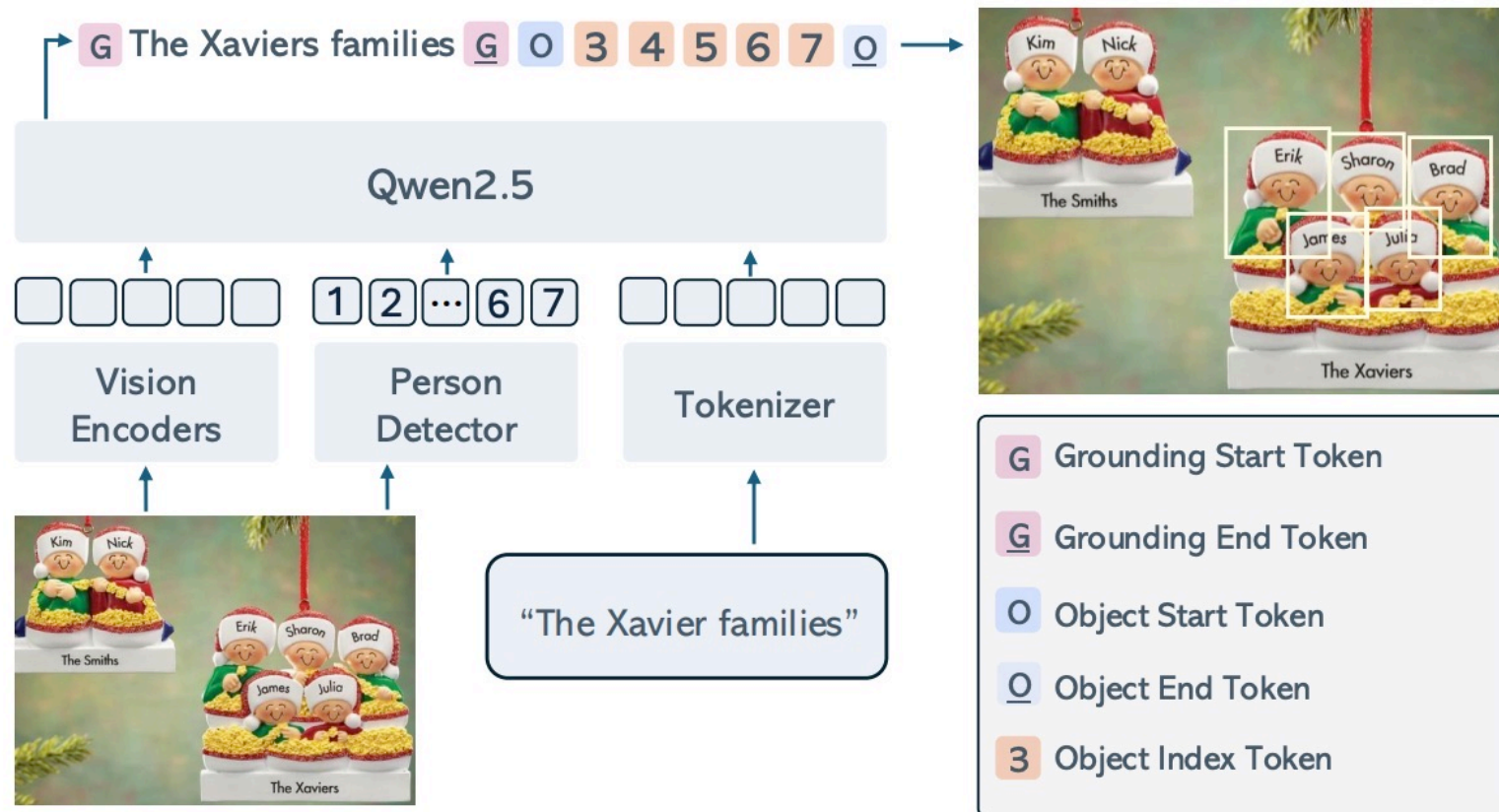


Figure 3. Overview of the manual annotation pipeline of the HumanRef dataset.

RexSeek

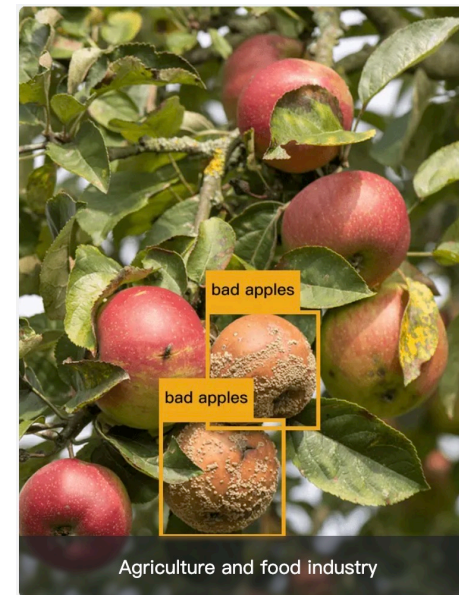
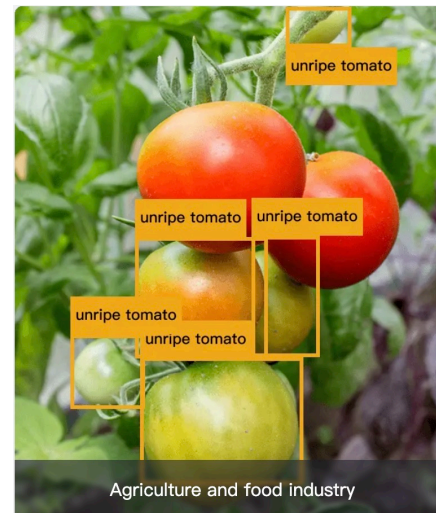
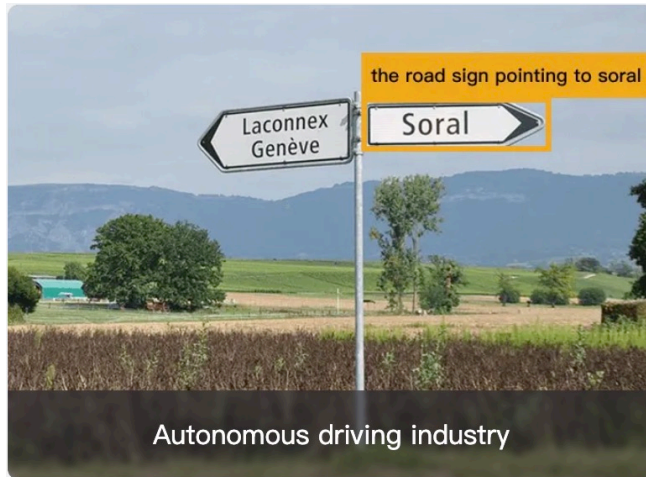


1. Strong perception capability
2. Strong language comprehension

Applications: Referring Any Person



Applications: Any Object Referring



Rex-Thinker: Grounded Object Referring via Chain-of-Thought Reasoning

Qing Jiang^{1,2*}, **Xingyu Chen^{3*}**, **Zhaoyang Zeng¹**, **Junzhi Yu³**, **Lei Zhang^{1,2†}**

¹International Digital Economy Academy (IDEA)

²South China University of Technology

³Peking University

{jiangqing, chenxingyu, leizhang}@idea.edu.cn

Live Reasoning Demo

Watch Rex-Thinker analyze and reason step-by-step

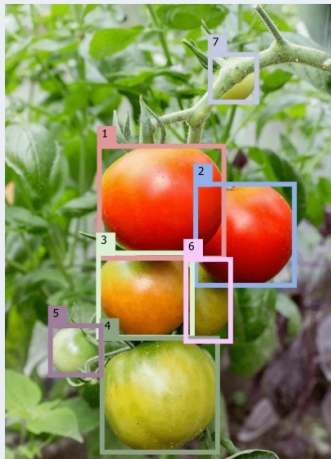


Analyzing image...

Rex-Thinker AI Reasoning

Human:

Box hint



Answer



Question: Please detect **ripe tomato**..

Answer:

<think>OK, the user needs us to detect the ripe tomato. In order to accomplish this task, I need to break it down into the following steps:

- Step 1: Determine the ripeness of each tomato based on its color. Let's start our analysis:

Step 1: Determine the ripeness of each tomato based on its color

I see seven tomatoes in this image, let's look at their colors and put a **!** if it is a ripe tomato and **×** if it is not.

- Tomato 1: It has a deep red color, indicating it is a ripe tomato. **!**
- Tomato 2: It also has a deep red color, indicating it is a ripe tomato. **!**
- Tomato 3: It has an orange-brown color, which is not the typical red color of a fully-ripe tomato. **×**
- Tomato 4: It is green in color, so it is not a ripe tomato. **×**
- Tomato 5: It is green in color, so it is not a ripe tomato. **×**
- Tomato 6: It is green in color, so it is not a ripe tomato. **×**
- Tomato 7: It is green in color, so it is not a ripe tomato. **×**

Summarize and Re-Check answer

Let's now recheck our answer and put **✓** for the target tomato and **×** for others

- Tomato 1: It is a ripe tomato. **✓**
- Tomato 2: It is a ripe tomato. **✓**
- Tomato 3: It is not a ripe tomato. **×**
- Tomato 4: It is not a ripe tomato. **×**
- Tomato 5: It is not a ripe tomato. **×**
- Tomato 6: It is not a ripe tomato. **×**
- Tomato 7: It is not a ripe tomato. **×** **</think>**

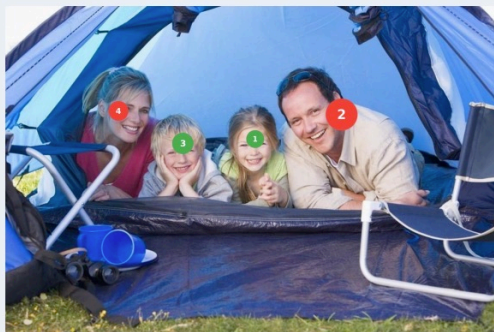
<answer>[{"bbox_2d": [235, 358, 565, 657], "label": "ripe tomato"}, {"bbox_2d": [481, 455, 744, 726], "label": "ripe tomato"}]**</answer>**

Planning

Action

Summarization

Set-of-marks
Image



Meta Information

Question Referring: the person between the two adults

Answer: Person 1, Person 3 Number of Person: 4

Position Order: [Person4, Person 3, Person 1, Person 2]

System Prompt (Partial)

Given the image and the description below, output a detailed analysis of how you found the object(s) matching the description. You should conduct an analysis of what you see and how each object relates to the provided description and put a ✗ or ✓ at the end.... Example: Example 1,2 ...



GPT-4o

1. Planning:

OK, the user needs us to detect the person between the two adults. To accomplish this task, I need to break it down into the following steps:

- Step 1: Identify the two adults in the image.
- Step 2: Determine the person located between the two adults.

2. Action:

Step 1: Identify the two adults in the image.

I see four people in this image. Let's look at their age - related attributes and put a ⚠ if they are adults and ✗ if they are not.

- Person 1: A young girl with long hair, not an adult. → ✗
- Person 2: A man with sunglasses on his head, appears to be an adult. → ⚠
- Person 3: A young boy with short hair, not an adult. → ✗
- Person 4: A woman with her hair tied back, appears to be an adult. → ⚠

So Person 2 and Person 4 are the two adults.

Step 2: Determine the person located between the two adults.

The order from left to right is [person 4, person 3, person 1, person 2].

Between Person 4 (the woman on the left - hand side) and Person 2 (the man on the right - hand side), there are two people: Person 3 and Person 1.

3. Summarization:

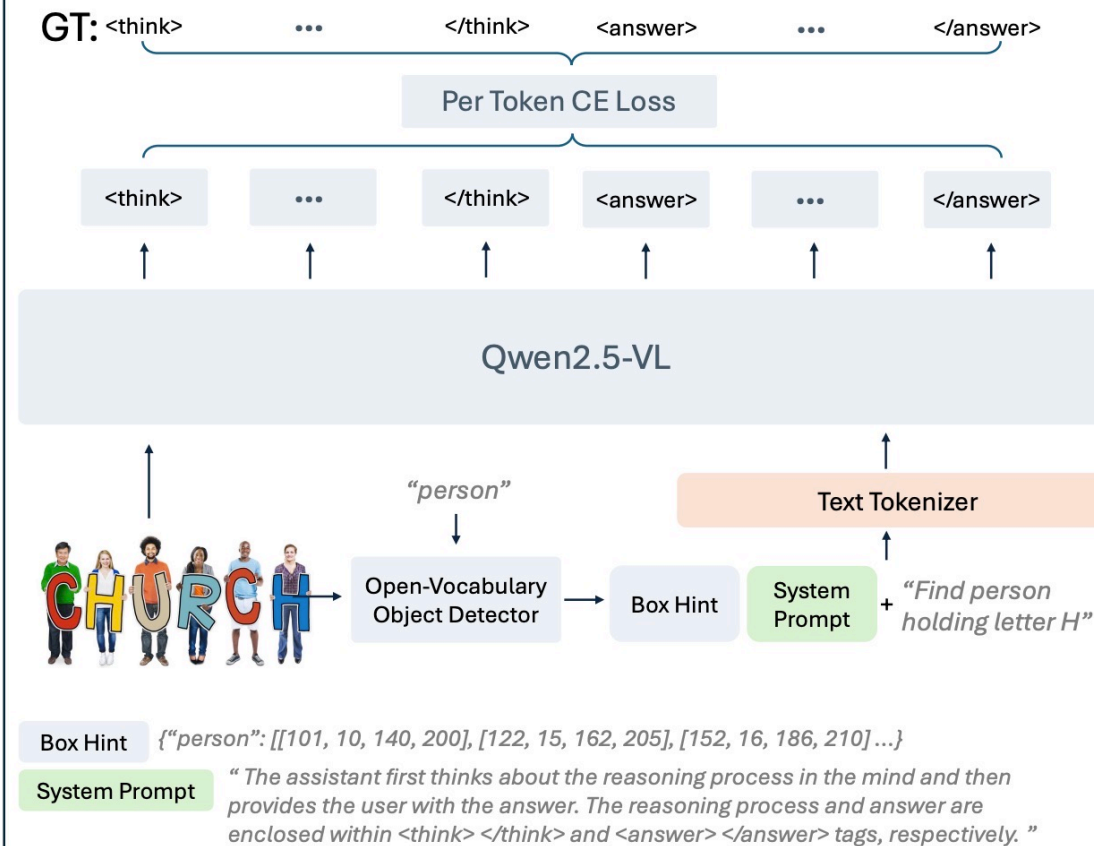
Summarize and Re-Check answer

Let's now recheck our answer and put ✓ for the target people and ✗ for others:

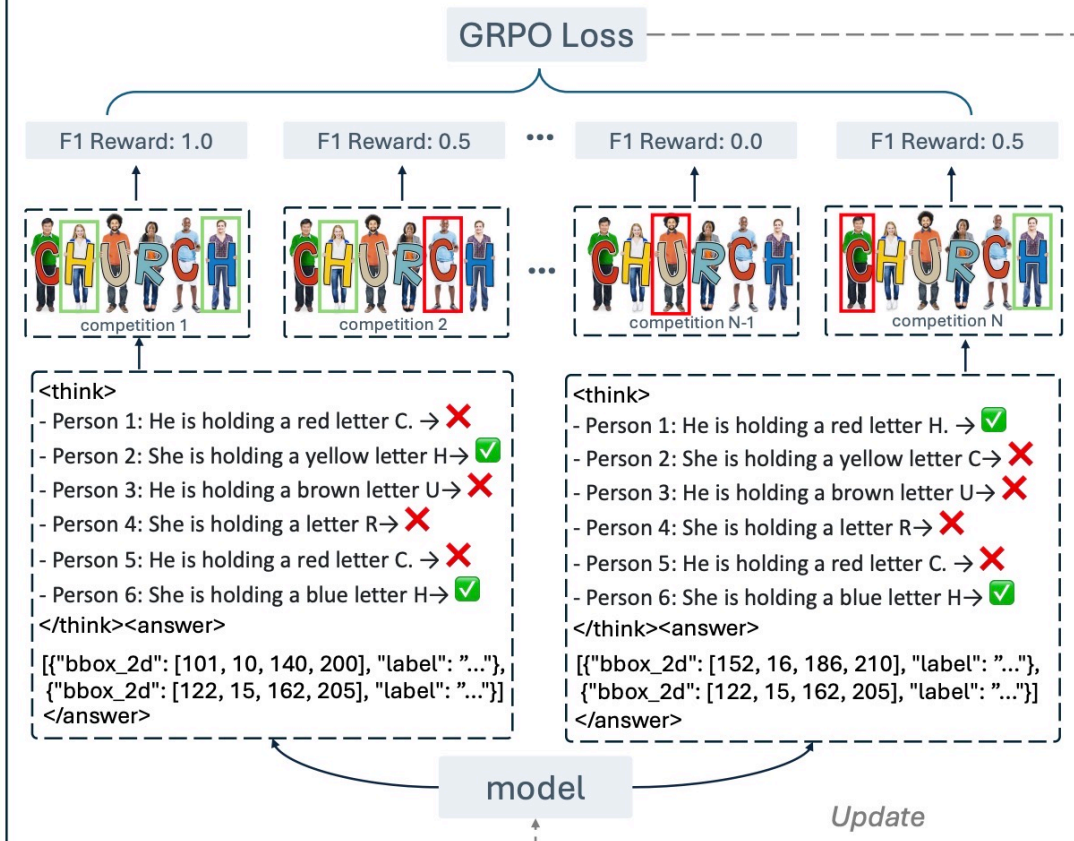
- Person 1: Located between the two adults. → ✓
- Person 2: One of the adults, not the person between the adults. → ✗
- Person 3: Located between the two adults. → ✓
- Person 4: One of the adults, not the person between the adults. → ✗

Model: Qwen2.5-VL with SFT + GRPO Training

SFT Cold Start



RL Post Training



Method	Attribute			Position			Interaction			Reasoning			Celebrity			Average			Rejection Score
	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	
DINOX [46]	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	36.0
InternVL-2.5-8B [8]	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	54.9
Ferret-7B [66]	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	2.0
Groma-7B [36]	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	0.0
ChatRex-7B [22]	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	0.0
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	7.1
DeepSeek-VL2-small [57]	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	3.1
Molmo-7B-D [11]	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	68.6
RexSeek-7B [21]	<u>87.2</u>	<u>86.8</u>	<u>81.5</u>	<u>86.1</u>	<u>86.3</u>	<u>83.8</u>	84.8	<u>84.6</u>	80.7	87.8	<u>84.7</u>	<u>81.5</u>	83.4	86.5	84.2	<u>85.9</u>	<u>85.8</u>	<u>82.3</u>	54.1
Rex-Thinker-Plain	83.0	88.7	81.4	82.5	83.9	81.3	80.1	85.6	<u>80.2</u>	80.5	82.2	77.3	86.7	88.7	86.8	82.6	85.8	81.4	53.5
Rex-Thinker-CoT	86.6	87.7	<u>82.7</u>	86.5	87.0	<u>84.3</u>	79.6	81.7	77.2	85.7	83.8	80.3	87.6	89.5	87.2	85.2	85.9	<u>82.3</u>	67.3
Rex-Thinker-GRPO	88.5	88.7	84.1	87.2	87.1	84.6	<u>81.5</u>	83.5	79.1	<u>87.7</u>	85.4	82.3	88.0	<u>89.3</u>	87.2	86.6	86.8	83.5	68.2

Rex-Thinker-Plain: Training without CoT

Rex-Thinker-CoT: Training with CoT

Rex-Thinker-GRPO: Training with CoT and GRPC

- Conclusion 1: CoT can effectively improve mode answer questions about objects that are not present
- Conclusion 2: Reinforcement learning fine-tuning



Question: Please detect **Kobe Bryant**.

Answer:

<think>OK, the user needs us to detect Kobe Bryant. To accomplish this task, I need to break it down into the following steps:

- Step 1: Identify each person based on their appearance and basketball-related attributes.
- Step 2: Check if any of the identified people match the identity of Kobe Bryant.

Step 1: Identify each person based on their appearance and basketball-related attributes.

- Person 1: A man wearing a white Dallas-branded basketball jersey, holding a basketball. He has short hair and a beard.
- Person 2: A man wearing a green Boston-branded basketball jersey with the number 0. He is in a shooting pose.
- Person 3: A man wearing a white Denver Nuggets-branded basketball jersey with the number 15. He is in a shooting pose.
- Person 4: A man wearing a black San Antonio Spurs-branded basketball jersey. He has a muscular build.
- Person 5: A man wearing a blue and yellow basketball jersey. He has a beard and is smiling.
- Person 6: Not clearly visible in terms of distinct features related to identification.
- Person 7: Partially visible, no clear distinct features for identification.
- Person 8: A man wearing a purple and yellow hoodie, not in a basketball-playing pose.
- Person 9: A man wearing a blue Thunder-branded basketball jersey with the number 2. He has a headband.

Step 2: Check if any of the identified people match the identity of Kobe Bryant.

Kobe Bryant is a well-known basketball player, but **none of the people in the image have the exact appearance and context that can confirm them as Kobe Bryant...**

Summarize and Re-Check answer

Let's now recheck our answer and put for the target person and for others

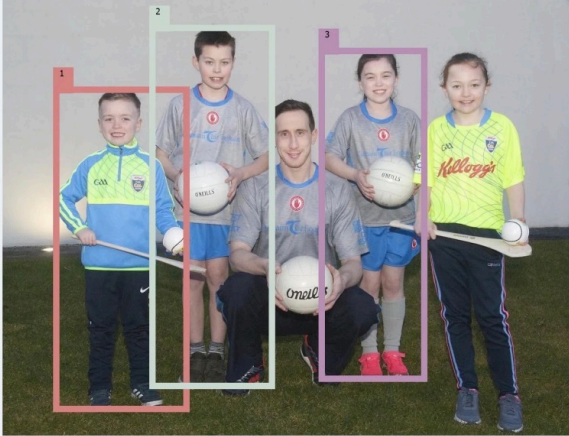
- Person 1-9: Not Kobe Bryant → </think>

<answer>There is no such thing in the image.</answer>

Ablation: Cold start training is important

With Cold Start	Attribute			Position			Interaction			Reasoning			Celebrity			Average			Rejection
	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	R	P	DF1	Score
No	81.4	85.8	78.1	80.2	80.2	77.5	79.6	82.6	78.0	77.6	75.0	70.6	87.3	86.5	84.8	81.2	82.0	77.8	66.4
Yes	88.5	88.7	84.1	87.2	87.1	84.6	81.5	83.5	79.1	87.7	85.4	82.3	88.0	89.3	87.2	86.6	86.8	83.5	68.2

Table 5: Ablation on the impact of CoT-based cold start on final performance after GRPO training.



Question: Please detect the person holding a small ball

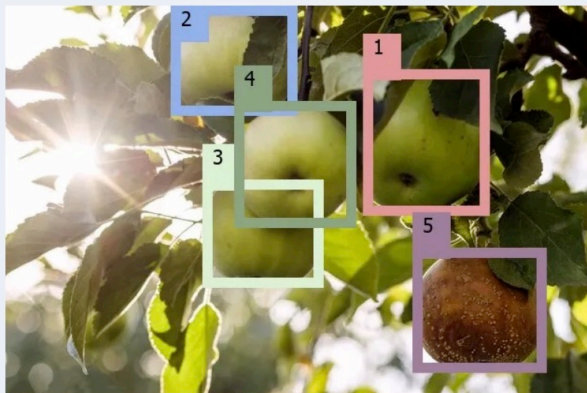
Answer:
<think>
There is **one person** holding a small ball in the image.
This person is standing on the left side of the group.
</think>
<answer>
[100, 174, 369, 821],[290, 51, 538, 775],[625, 97, 839, 761]
</answer>

- Conclusion 3: CoT-based SFT cold start helps the model learn the reasoning format, enabling it to perform reasoning in the desired format during GRPO.

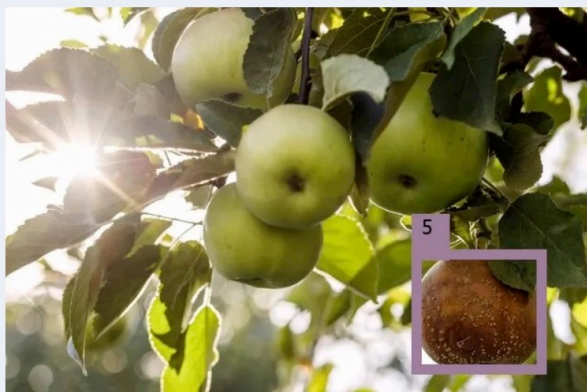
GRPO without CoT SFT as cold start

Visualization Results: Generalized to Any Object

Box hint



Answer



Question: Please detect **apple with disease**, and tell me what kind of disease it may be..

Answer:

<think>OK, the user needs us to detect the apple with disease and tell what kind of disease it may be. In order to accomplish this task, I need to break it down into the following steps:

- Step 1: Examine each apple to determine if it has any visible signs of disease.
- Step 2: If an apple is found to have a disease, identify the type of the disease. Let's start our analysis:

Step 1: Examine each apple to determine if it has any visible signs of disease

I see five apples in this image. Let's look at them one-by-one and put a ⚠ if there are signs of disease and ✖ if there are none.

- Apple 1: It appears to be a healthy green apple with no visible signs of disease. ✖
- Apple 2: It is also a healthy-looking green apple with no obvious signs of disease. ✖
- Apple 3: This apple looks like a healthy green apple with no clear signs of disease. ✖
- Apple 4: Another healthy green apple with no visible signs of disease. ✖
- Apple 5: **This apple has a brownish spot on its surface, which indicates that it has a disease.** ⚠

Step 2: Identify the type of the disease

The brownish spot on Apple 5 suggests that it may have a fungal or bacterial infection, which are common types of diseases for apples.

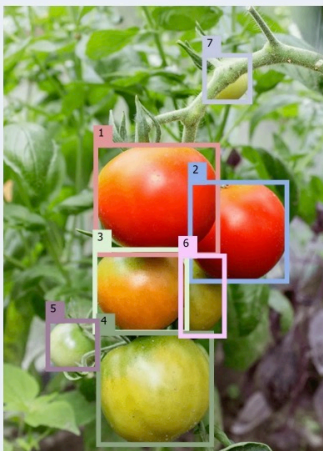
Summarize and Re-Check answer

Let's now recheck our answer and put ✅ for the target apple and ✖ for others:

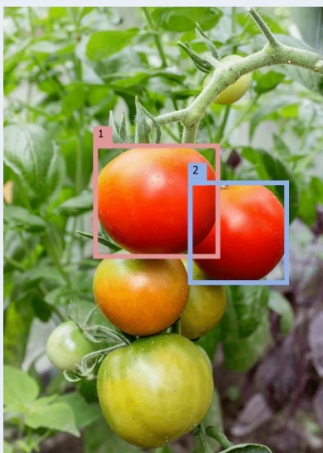
- Apple 1: It is a healthy apple with no signs of disease. ✖
- Apple 2: It is a healthy apple with no signs of disease. ✖
- Apple 3: It is a healthy apple with no signs of disease. ✖
- Apple 4: It is a healthy apple with no signs of disease. ✖
- Apple 5: It has a brownish spot, indicating a possible fungal or bacterial infection. ✅ **</think>**

<answer[{"bbox_2d": [429, 261, 570, 395], "label": "apple with disease"}]></answer>

Box hint



Answer



Question: Please detect **ripe tomato**..

Answer:

<think>OK, the user needs us to detect the ripe tomato. In order to accomplish this task, I need to break it down into the following steps:

- Step 1: Determine the ripeness of each tomato based on its color. Let's start our analysis:

Step 1: Determine the ripeness of each tomato based on its color

I see seven tomatoes in this image, let's look at their colors and put a  if it is a ripe tomato and  if it is not.

- Tomato 1: It has a deep red color, indicating it is a ripe tomato. 
- Tomato 2: It also has a deep red color, indicating it is a ripe tomato. 
- Tomato 3: It has an orange-brown color, which is not the typical red color of a fully-ripe tomato. 
- Tomato 4: It is green in color, so it is not a ripe tomato. 
- Tomato 5: It is green in color, so it is not a ripe tomato. 
- Tomato 6: It is green in color, so it is not a ripe tomato. 
- Tomato 7: It is green in color, so it is not a ripe tomato. 

Summarize and Re-Check answer

Let's now recheck our answer and put  for the target tomato and  for others

- Tomato 1: It is a ripe tomato. 
- Tomato 2: It is a ripe tomato. 
- Tomato 3: It is not a ripe tomato. 
- Tomato 4: It is not a ripe tomato. 
- Tomato 5: It is not a ripe tomato. 
- Tomato 6: It is not a ripe tomato. 
- Tomato 7: It is not a ripe tomato.  **</think>**

<answer[{"bbox_2d": [235, 358, 565, 657], "label": "ripe tomato"}, {"bbox_2d": [481, 455, 744, 726], "label": "ripe tomato"}]**</answer>**

Box hint



Answer



Question: Please detect **damaged container**.

Answer:

<think>OK, the user needs us to detect the damaged container. In order to accomplish this task, I need to break it down into the following steps:

- Step 1: Examine each of the 31 containers to determine if they have any visible signs of damage.

Step 1: Step 1: Examine each of the 31 containers to determine if they have any visible signs of damage.

- Container 1: It appears to be in good condition with no visible signs of damage. ✗
- Container 2: There is no visible sign of damage on this container. ✗
- Container 3: **This container has a large hole in its side, indicating significant damage.** ⚠
- Container 4: No visible signs of damage. ✗

...

- Container 31: No visible signs of damage. ✗

Summarize and Re-Check answer

Let's now recheck our answer and put ✓ for the target container and ✗ for others

- Container 1: No visible signs of damage → ✗
- Container 2: No visible signs of damage → ✗
- Container 3: Has a large hole, is damaged → ✓
- Container 4: No visible signs of damage → ✗

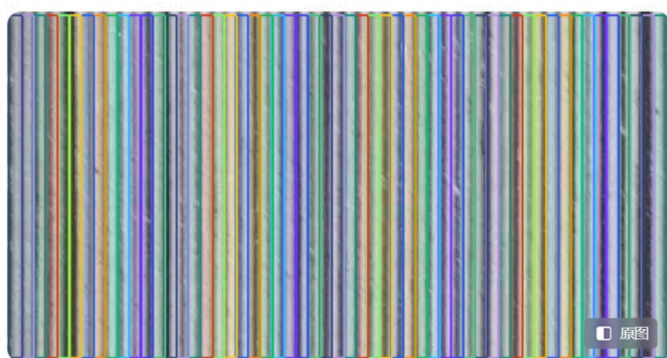
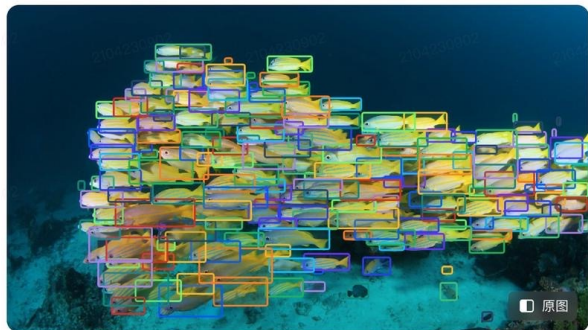
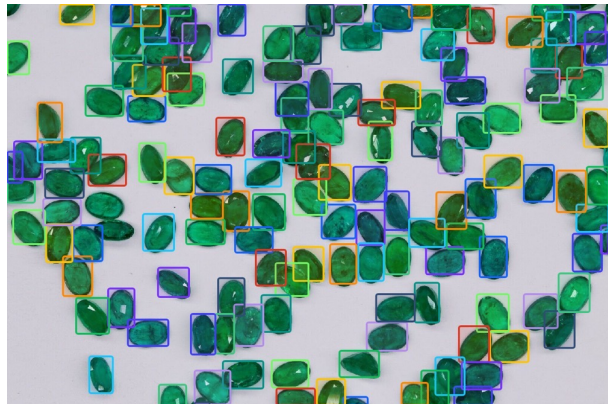
...

- Container 30: No visible signs of damage → ✗
- Container 31: No visible signs of damage → ✗

</think>

<answer>[{"bbox_2d": [288, 183, 699, 459], "label": "damaged container"}]``

SEED-VL-1.5 (Dense Object Detection)

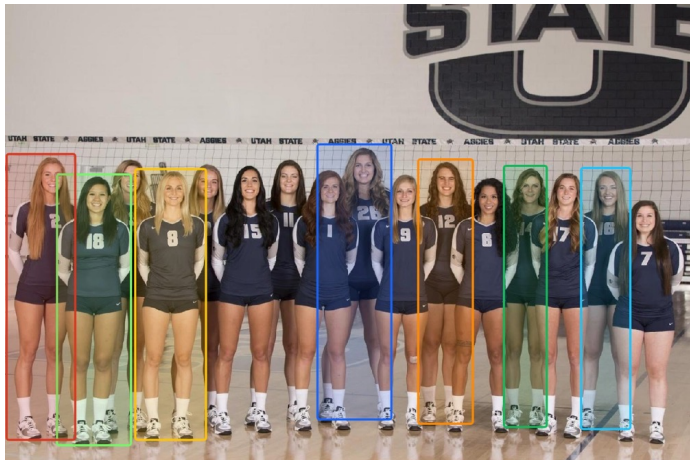


Question:帮我框选图中所有的鸽子，以
<bbox>x1 y1 x2 y2</bbox>的形式表示

Abswer:

```
<bbox>3 10 22 987</bbox> <bbox>21 10 40 987</bbox> <bbox>39 10 58 987</bbox> <bbox>57 10 76 987</bbox> <bbox>75 10 94 987</bbox> <bbox>93 10 112 987</bbox> <bbox>111 10 130 987</bbox> <bbox>129 10 148 987</bbox> <bbox>147 10 166 987</bbox> <bbox>165 10 184 987</bbox> <bbox>183 10 202 987</bbox> <bbox>201 10 220 987</bbox> <bbox>219 10 238 987</bbox> <bbox>237 10 256 987</bbox> <bbox>255 10 274 987</bbox> <bbox>273 10 292 987</bbox> <bbox>291 10 310 987</bbox> <bbox>309 10 328 987</bbox> <bbox>327 10 346 987</bbox> <bbox>345 10 364 987</bbox> <bbox>363 10 382 987</bbox> <bbox>381 10 400 987</bbox> <bbox>399 10 418 987</bbox> <bbox>417 10 436 987</bbox> <bbox>435 10 454 987</bbox> <bbox>453 10 472 987</bbox> <bbox>471 10 490 987</bbox> <bbox>489 10 508 987</bbox> <bbox>507 10 526 987</bbox> <bbox>525 10 544 987</bbox> <bbox>543 10 562 987</bbox> <bbox>561 10 580 987</bbox> <bbox>579 10 598 987</bbox> <bbox>597 10 616 987</bbox> <bbox>615 10 634 987</bbox> <bbox>633 10 652 987</bbox> <bbox>651 10 670 987</bbox> <bbox>669 10 688 987</bbox> <bbox>687 10 706 987</bbox> <bbox>705 10 724 987</bbox> <bbox>723 10 742 987</bbox> <bbox>741 10 760 987</bbox> <bbox>759 10 778 987</bbox> <bbox>777 10 796 987</bbox> <bbox>795 10 814 987</bbox> <bbox>813 10 832 987</bbox> <bbox>831 10 850 987</bbox> <bbox>849 10 868 987</bbox> <bbox>867 10 886 987</bbox> <bbox>885 10 904 987</bbox> <bbox>903 10 922 987</bbox> <bbox>921 10 940 987</bbox> <bbox>939 10 958 987</bbox> <bbox>957 10 976 987</bbox> <bbox>975 10 994 987</bbox>
```


SEED-VL-1.5 (Referring)



帮我框选图中身着衣服为偶数的人，以
<bbox>x1 y1 x2 y2</bbox>的形式表示



帮我框选图中穿红色衣服的人，以<bbox>x1 y1 x2 y2</bbox>的形式表示

- The next step for object detection is moving toward multimodal large language models.

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>

RexThinker

- Paper: <https://arxiv.org/abs/2506.04034>
- Code: <https://github.com/IDEA-Research/Rex-Thinker>



Taming Multimodal LLM for Object Perception

Thanks!