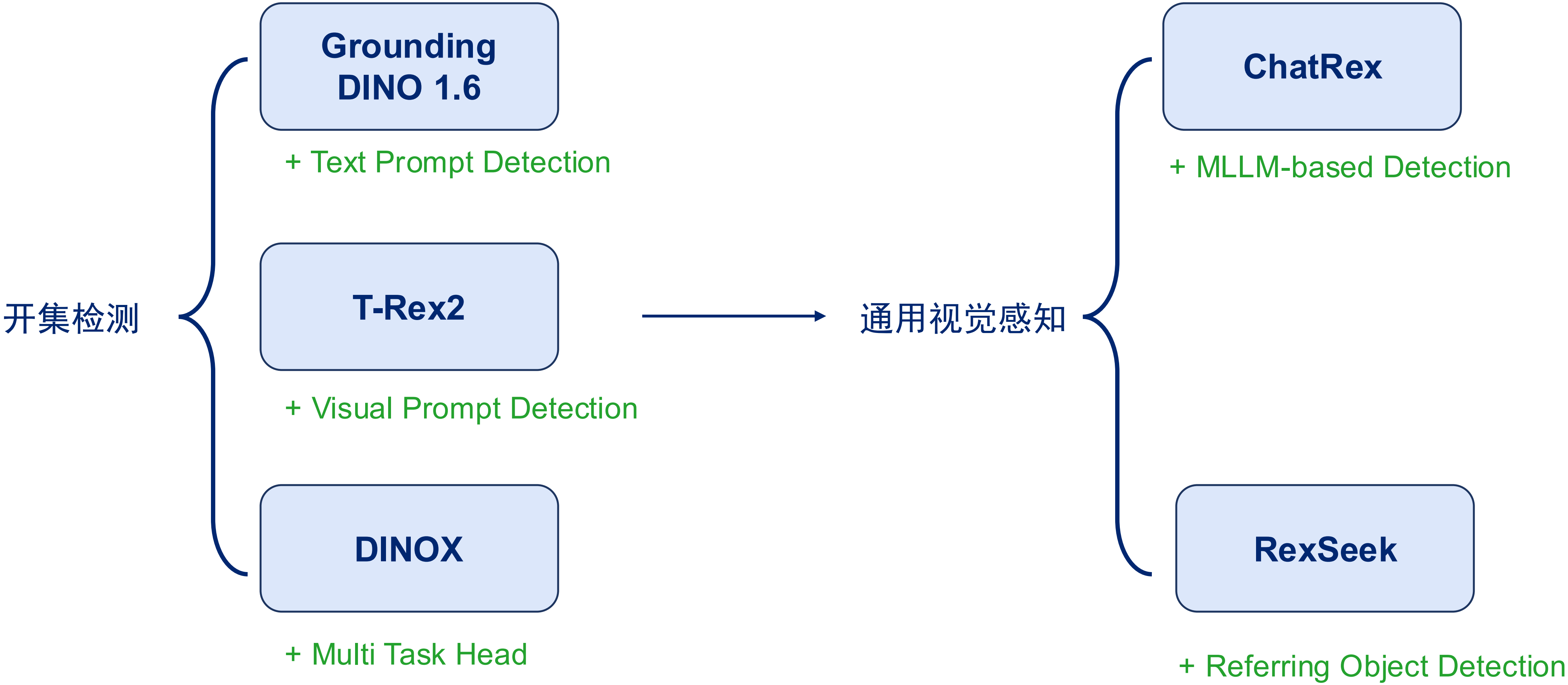
从开集检测迈向通用视觉感知

蒋擎

6-27



开集检测

Grounding DINO 1.6

+ Text Prompt Detection

T-Rex2

+ Visual Prompt Detection

DINOX

+ Multi Task Head

通用视觉感知

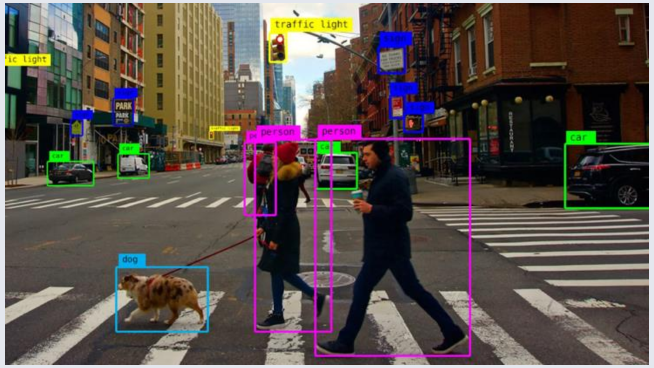
ChatRex

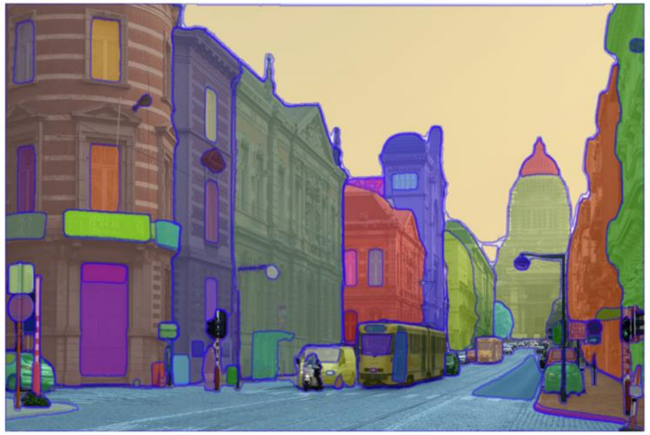
+ MLLM-based Detection

RexSeek

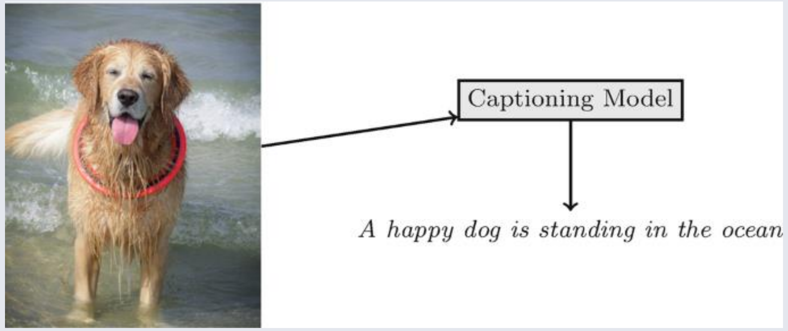
+ Referring Object Detection

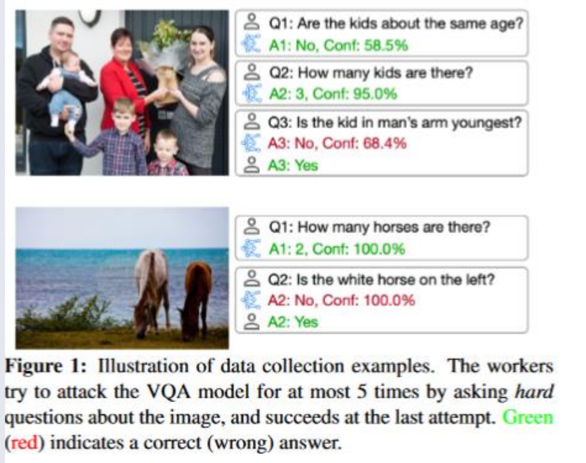
计算机视觉三大任务





感知 (Perception)



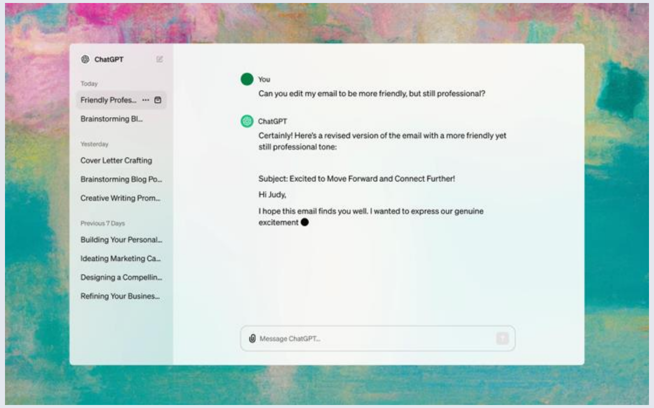


Q1: Are the kids about the same age?
A1: No, Conf: 58.5%
Q2: How many kids are there?
A2: 3, Conf: 95.0%
Q3: Is the kid in man's arm youngest?
A3: No, Conf: 68.4%
A3: Yes

Q1: How many horses are there?
A1: 2, Conf: 100.0%
Q2: Is the white horse on the left?
A2: No, Conf: 100.0%
A2: Yes

Figure 1: Illustration of data collection examples. The workers try to attack the VQA model for at most 5 times by asking *hard* questions about the image, and succeeds at the last attempt. Green (red) indicates a correct (wrong) answer.

理解 (Understanding)



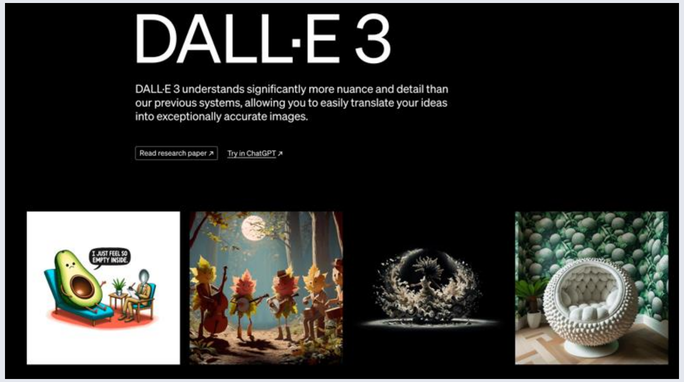
You
Can you edit my email to be more friendly, but still professional?

ChatGPT
Certainly! Here's a revised version of the email with a more friendly yet still professional tone:

Subject: Excited to Move Forward and Connect Further!

Hi Judy,

I hope this email finds you well. I wanted to express our genuine excitement ●

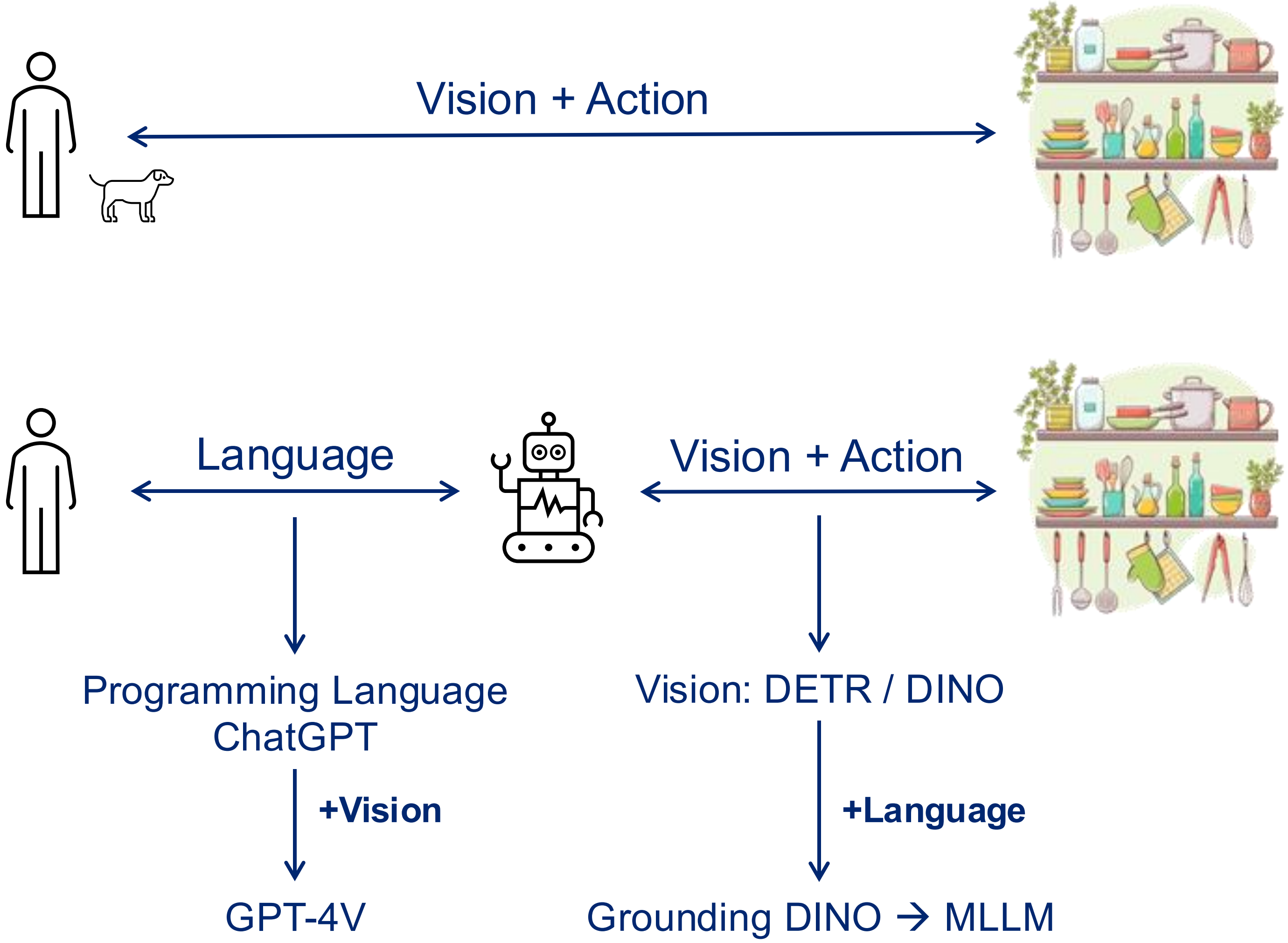


DALL·E 3

DALL·E 3 understands significantly more nuance and detail than our previous systems, allowing you to easily translate your ideas into exceptionally accurate images.

生成 (Generation)

视觉感知是机器和物理世界交互的基础



Vision + Action

Language

Vision + Action

Programming Language
ChatGPT

Vision: DETR / DINO

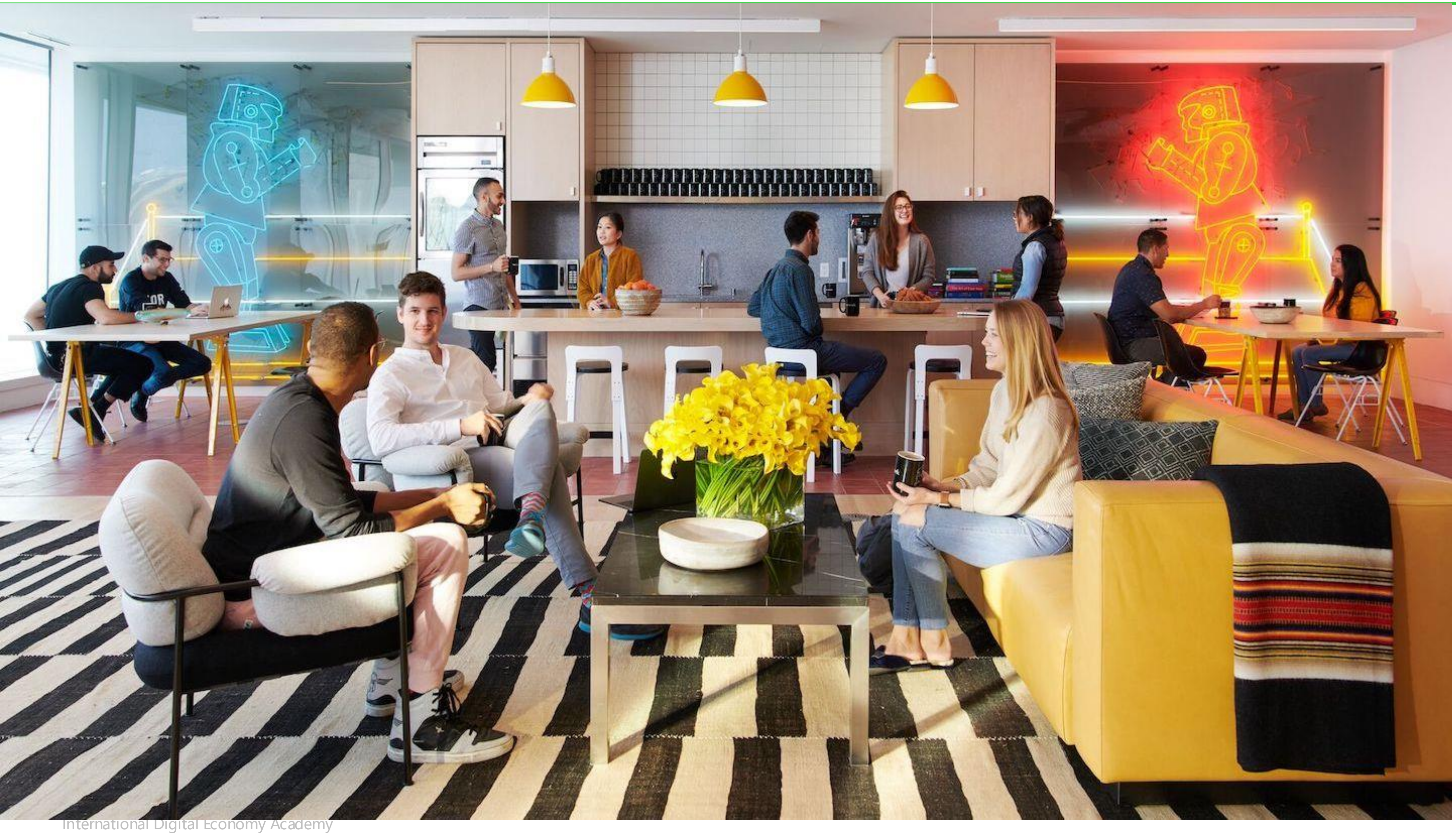
+Vision

+Language

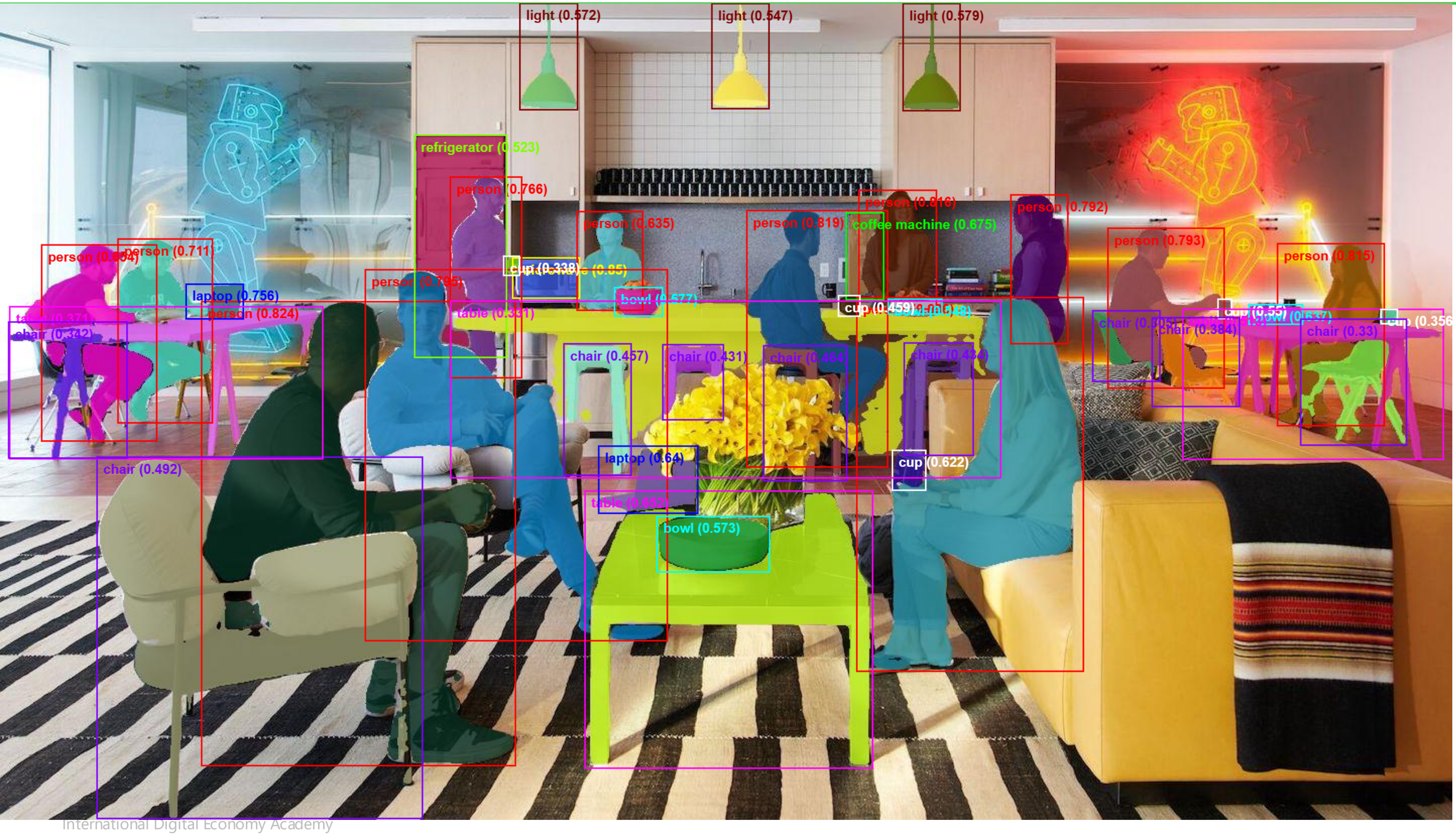
GPT-4V

Grounding DINO → MLLM

什么是视觉感知?

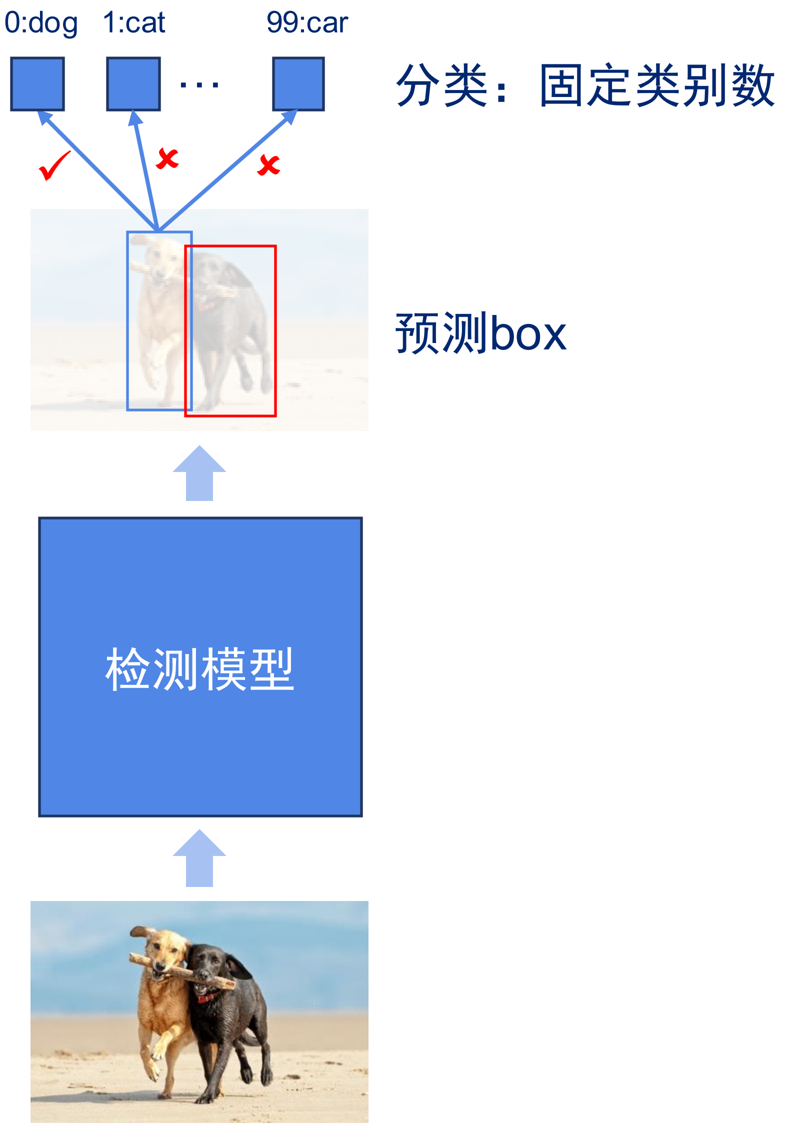


什么是视觉感知？以物体检测为例



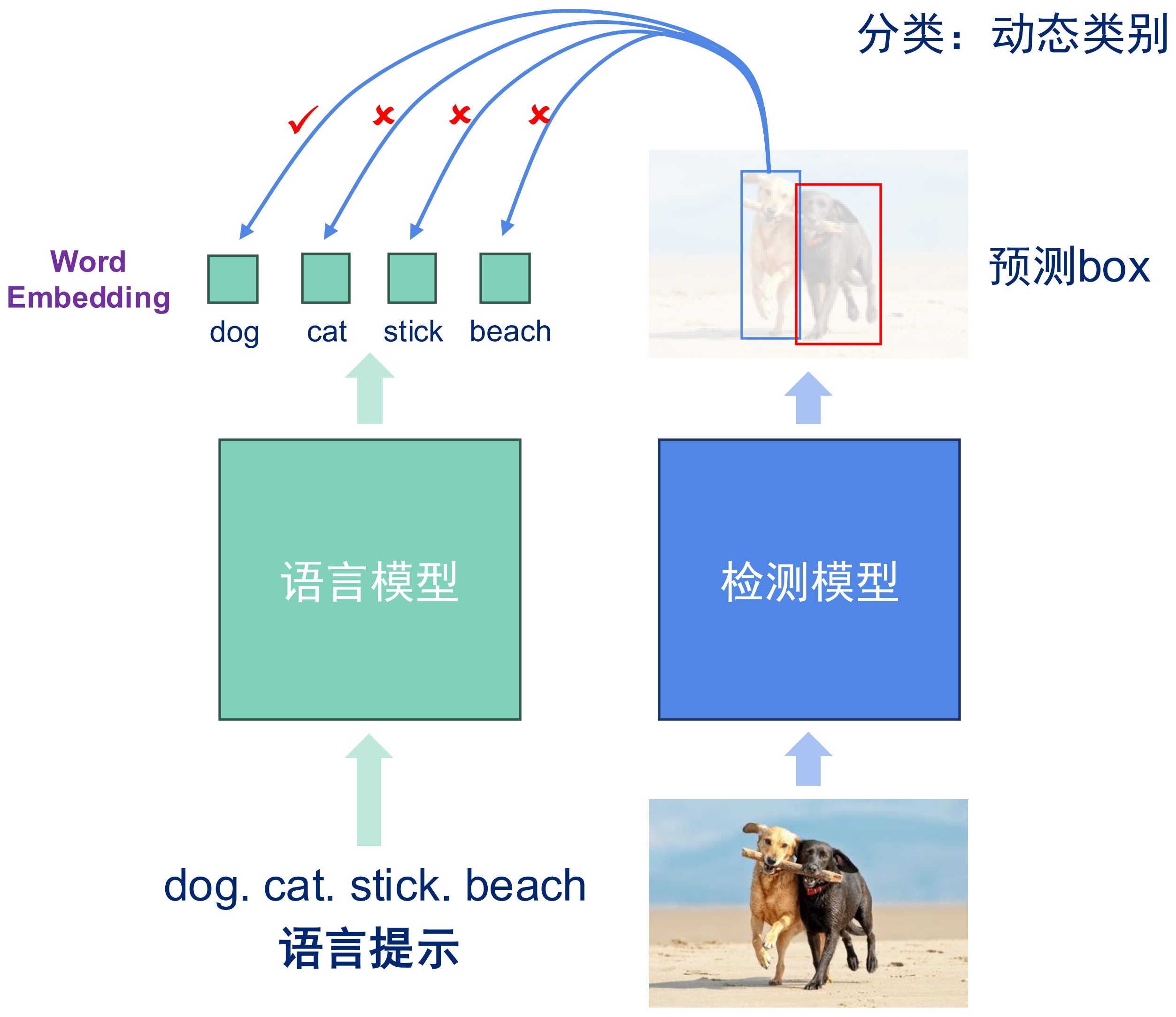
person. cup.
bowl. light.
chair.
coffee machine.
microwave.
refrigerator.
laptop. robot.
table

物体检测范式的迁移：闭集检测 vs. 开集检测



分类：固定类别数

预测box



分类：动态类别

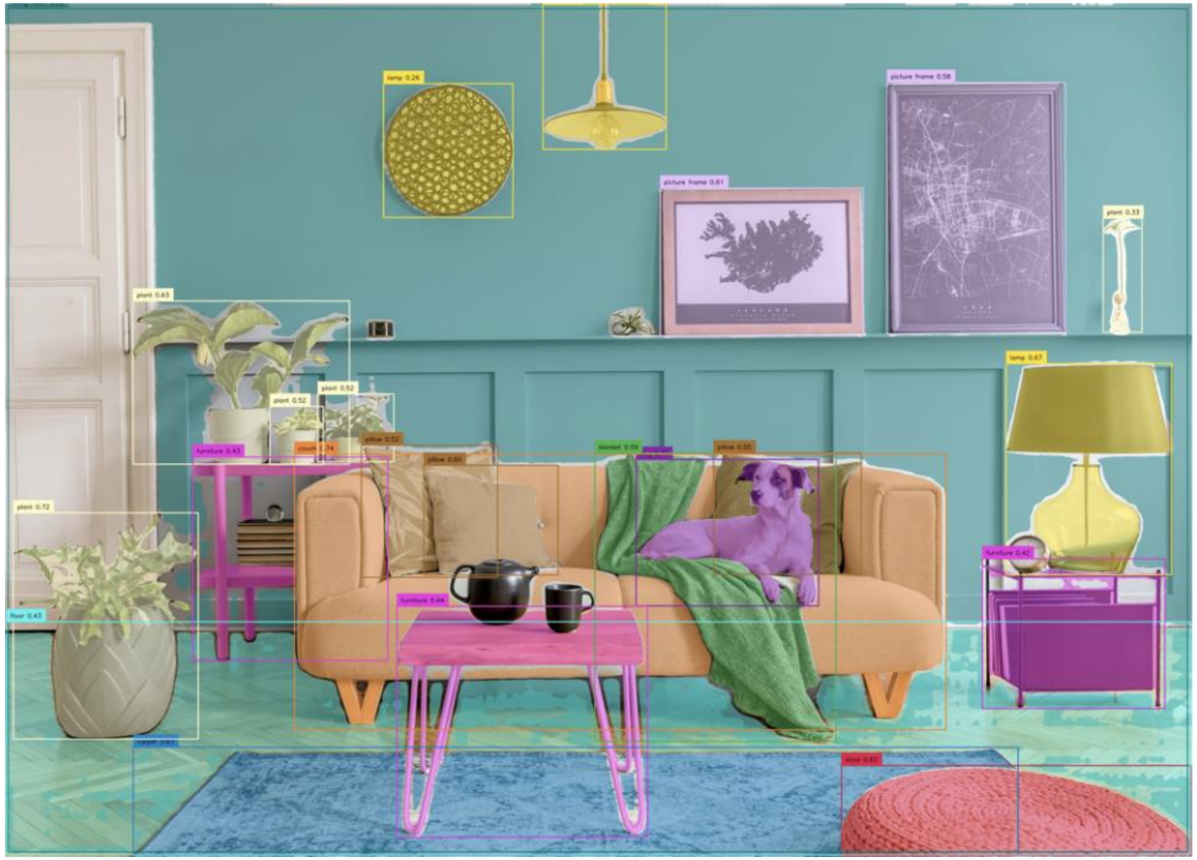
预测box

dog. cat. stick. beach

语言提示

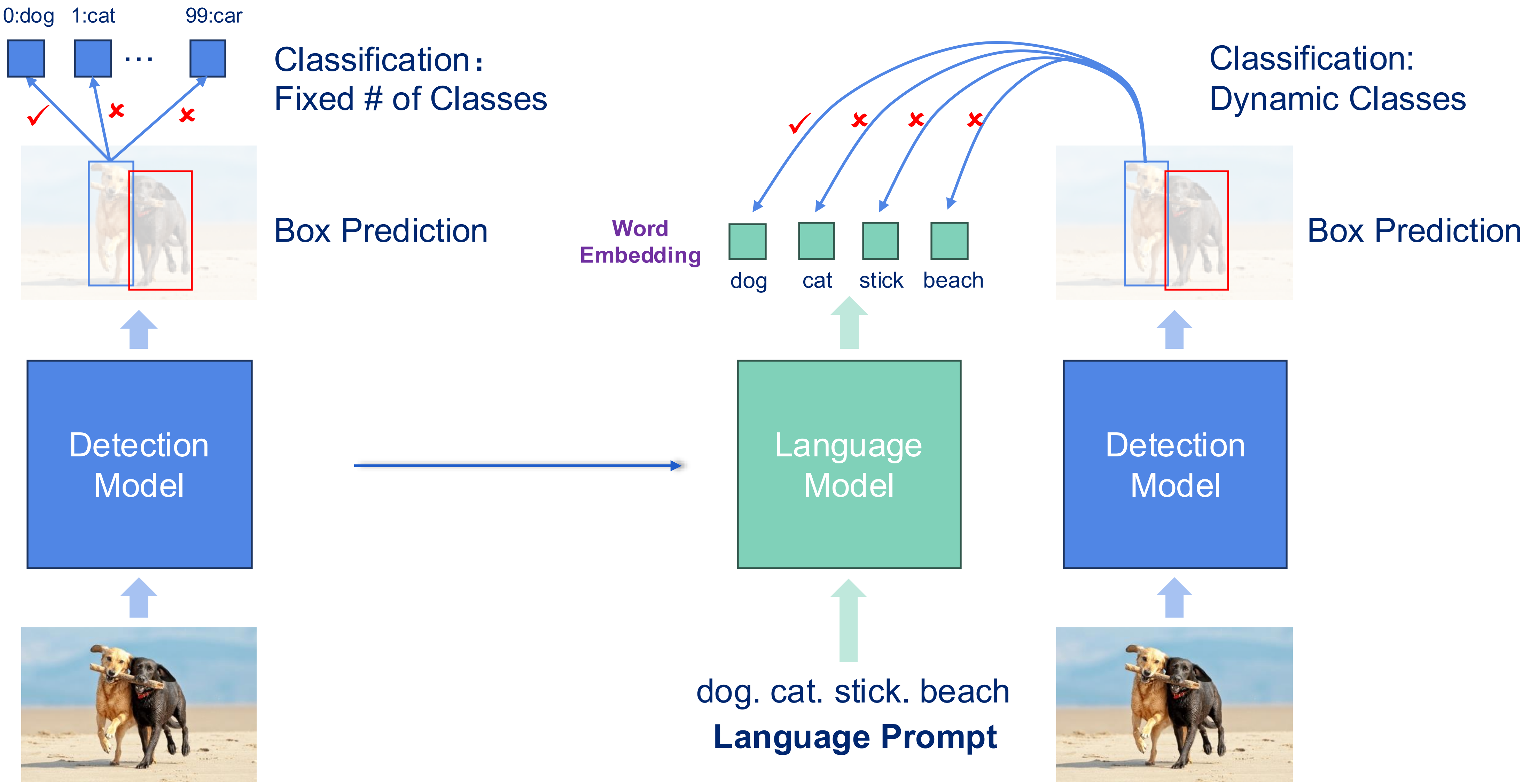
开集检测的目标

- 给定一张图片和任意的提示（文本提示，视觉提示）
- 模型能够根据提示检测出任意的物体，而不需要微调



“armchair, blanket, lamp, carpet, couch, dog, floor, furniture, gray, green, living room, picture frame, pillow, plant, room, sit, stool, wood floor”

基于文本提示的开集检测模型

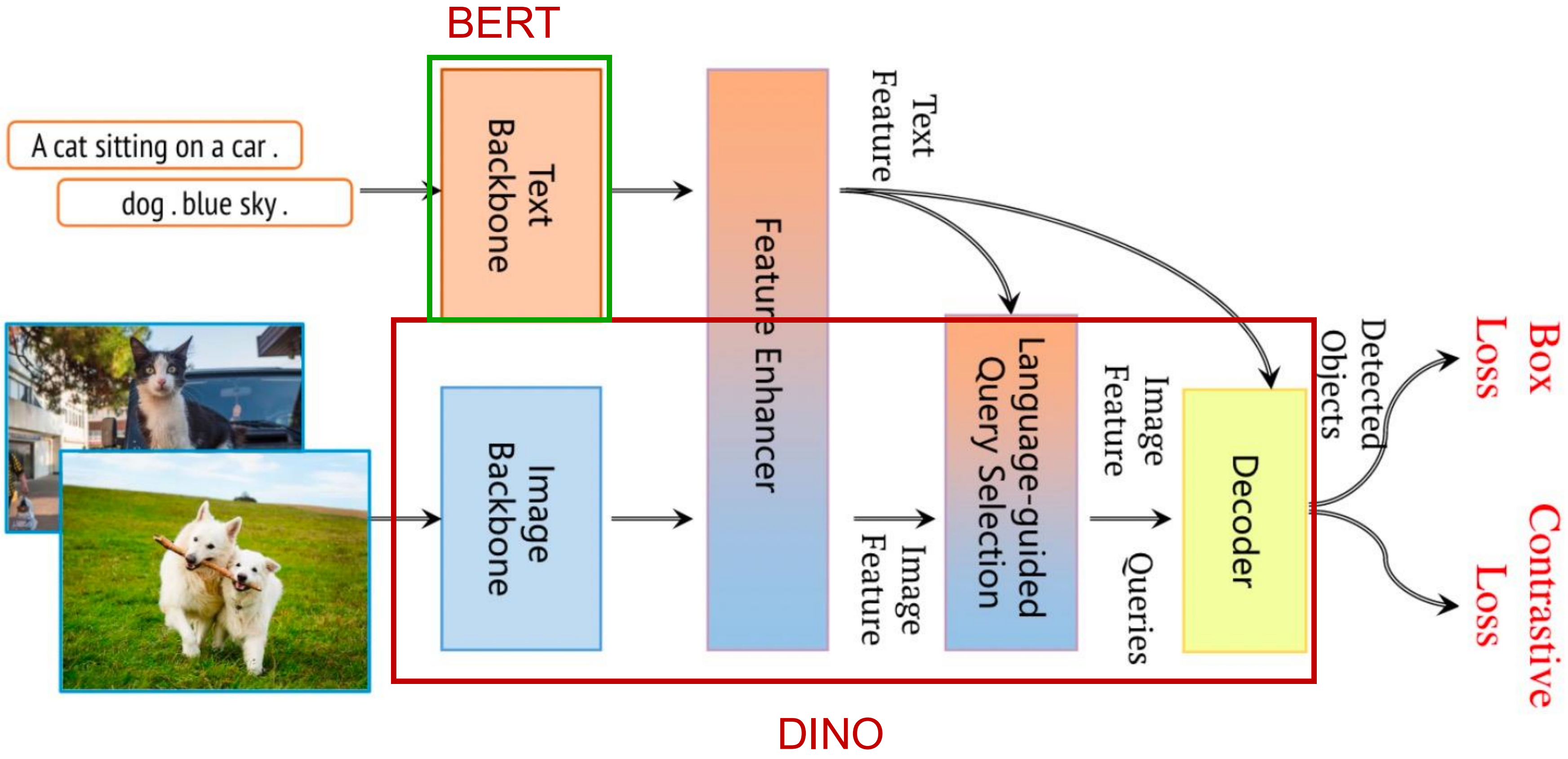


dog. cat. stick. beach

Language Prompt

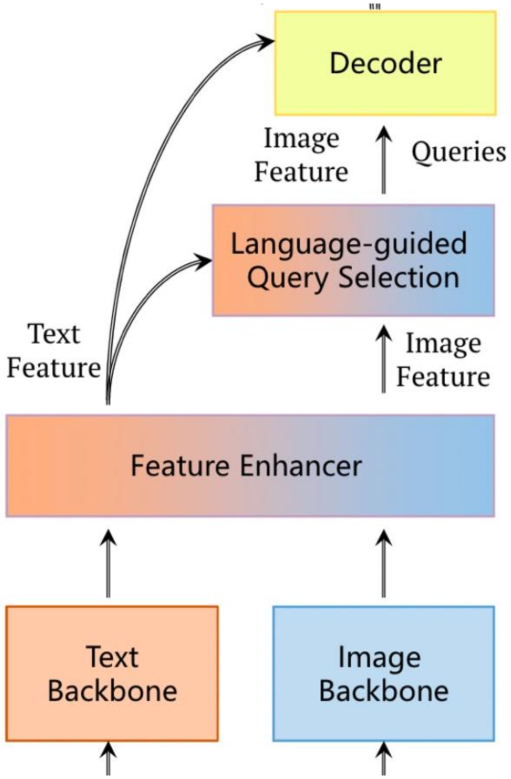
Grounding DINO 1.5: Advance the “Edge” of Open-Set Object Detection

Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection



(a) Model Framework

Pro V.S. Edge: Overall Architecture



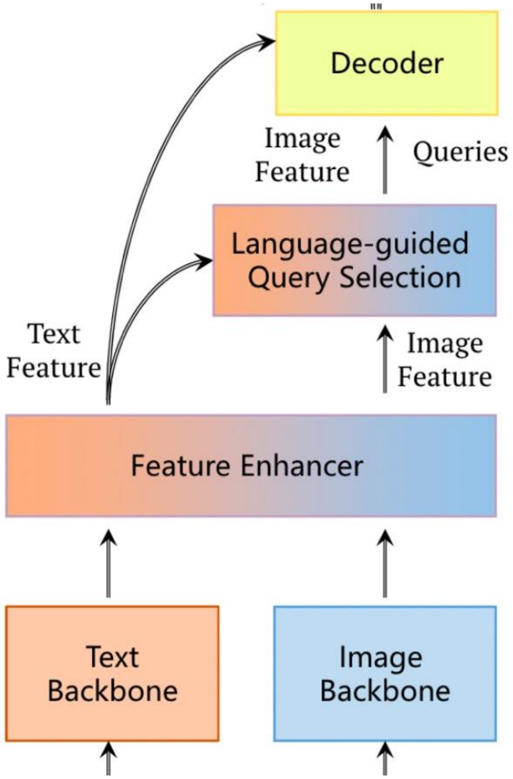
6 x Transformer Decoder

6 x Transformer Encoder, Multi-Scale

BERT-Base

ViT-L

1.5 Pro



6 x Transformer Decoder

1 x Transformer Encoder, Multi Scale

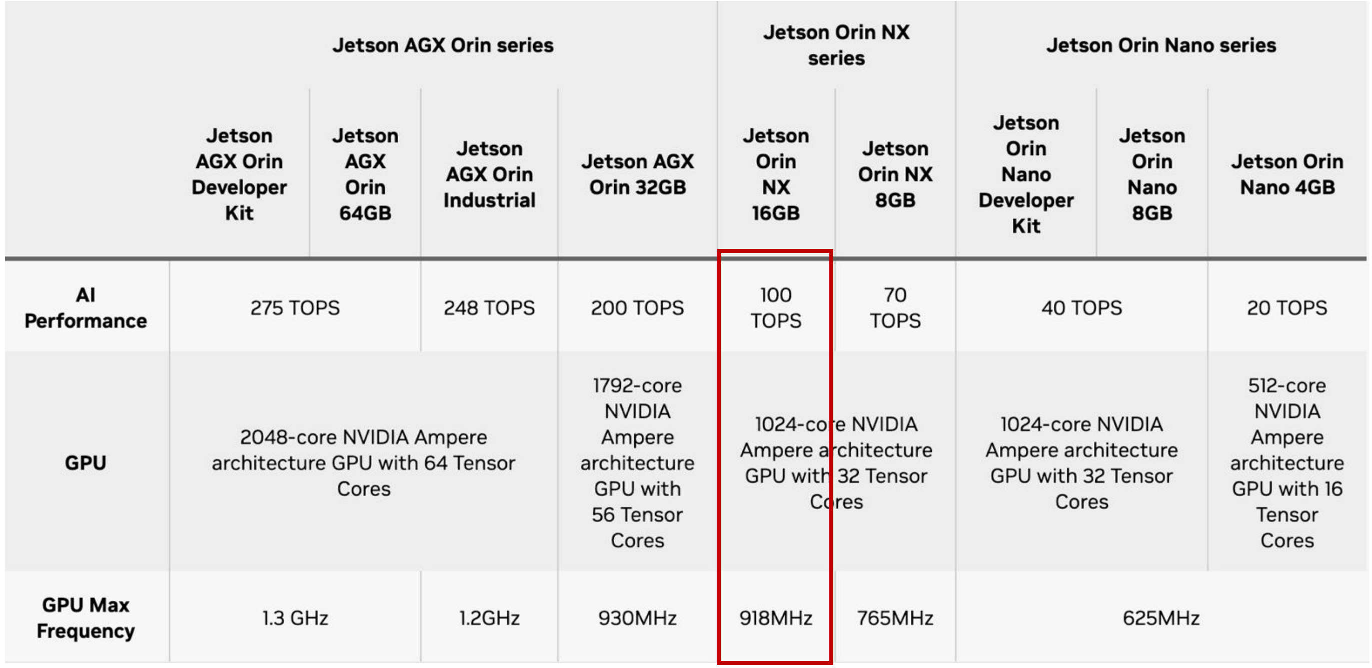
BERT-Base

EfficientViT-L1

1.5 Edge

Cai, Han, et al. "Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction". ICCV 2023.

边缘计算设备部署 (NVIDIA Orin NX)



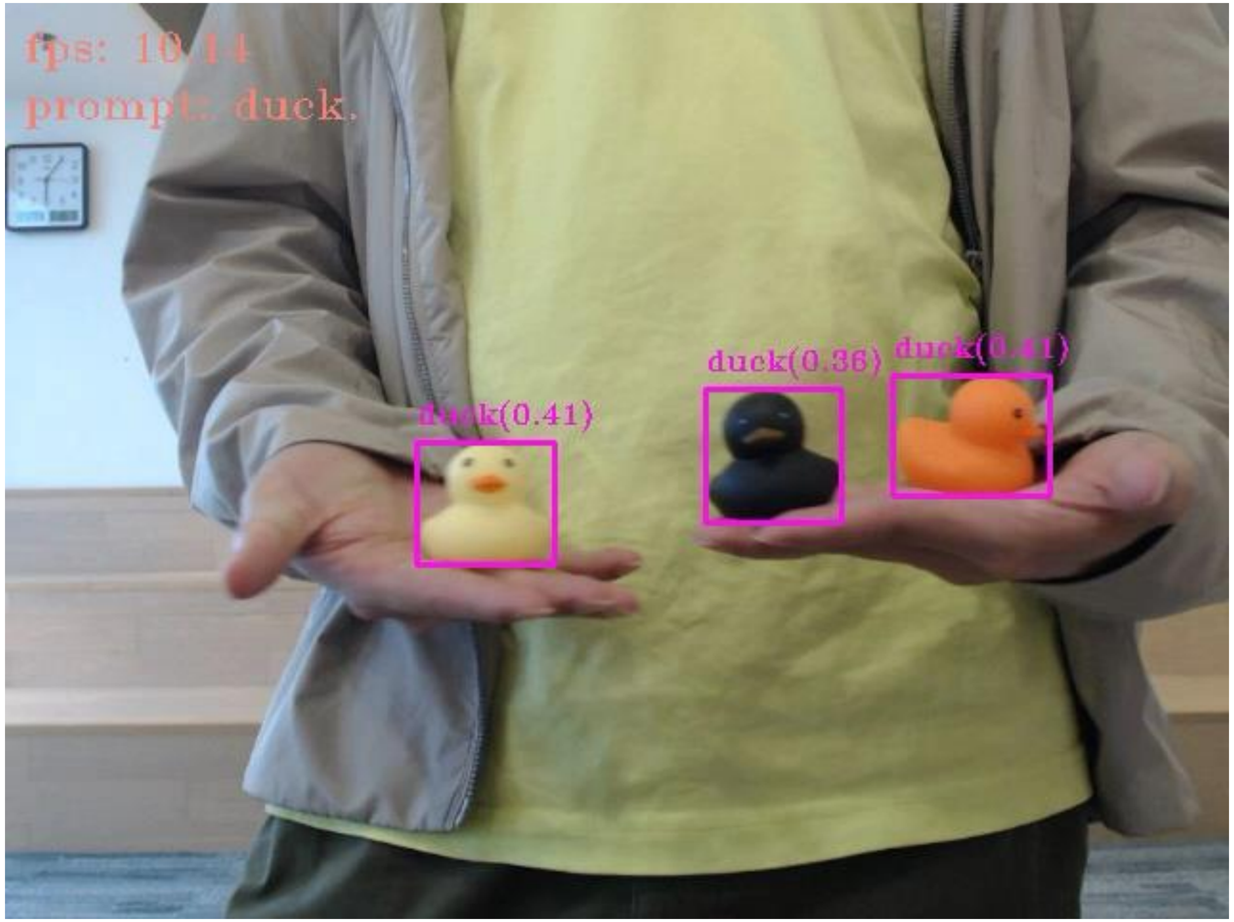
	Jetson AGX Orin series				Jetson Orin NX series		Jetson Orin Nano series		
	Jetson AGX Orin Developer Kit	Jetson AGX Orin 64GB	Jetson AGX Orin Industrial	Jetson AGX Orin 32GB	Jetson Orin NX 16GB	Jetson Orin NX 8GB	Jetson Orin Nano Developer Kit	Jetson Orin Nano 8GB	Jetson Orin Nano 4GB
AI Performance	275 TOPS		248 TOPS	200 TOPS	100 TOPS	70 TOPS	40 TOPS		20 TOPS
GPU	2048-core NVIDIA Ampere architecture GPU with 64 Tensor Cores			1792-core NVIDIA Ampere architecture GPU with 56 Tensor Cores	1024-core NVIDIA Ampere architecture GPU with 32 Tensor Cores		1024-core NVIDIA Ampere architecture GPU with 32 Tensor Cores		512-core NVIDIA Ampere architecture GPU with 16 Tensor Cores
GPU Max Frequency	1.3 GHz		1.2GHz	930MHz	918MHz	765MHz	625MHz		

Specification	Orin NX	RTX 3090
CUDA Cores	1024 cores	10496 cores
Tensor Cores	32 cores	328 cores
GPU Max Freq.	918MHZ	1695MHZ
TOPS	100 TOPS	~285TOPS

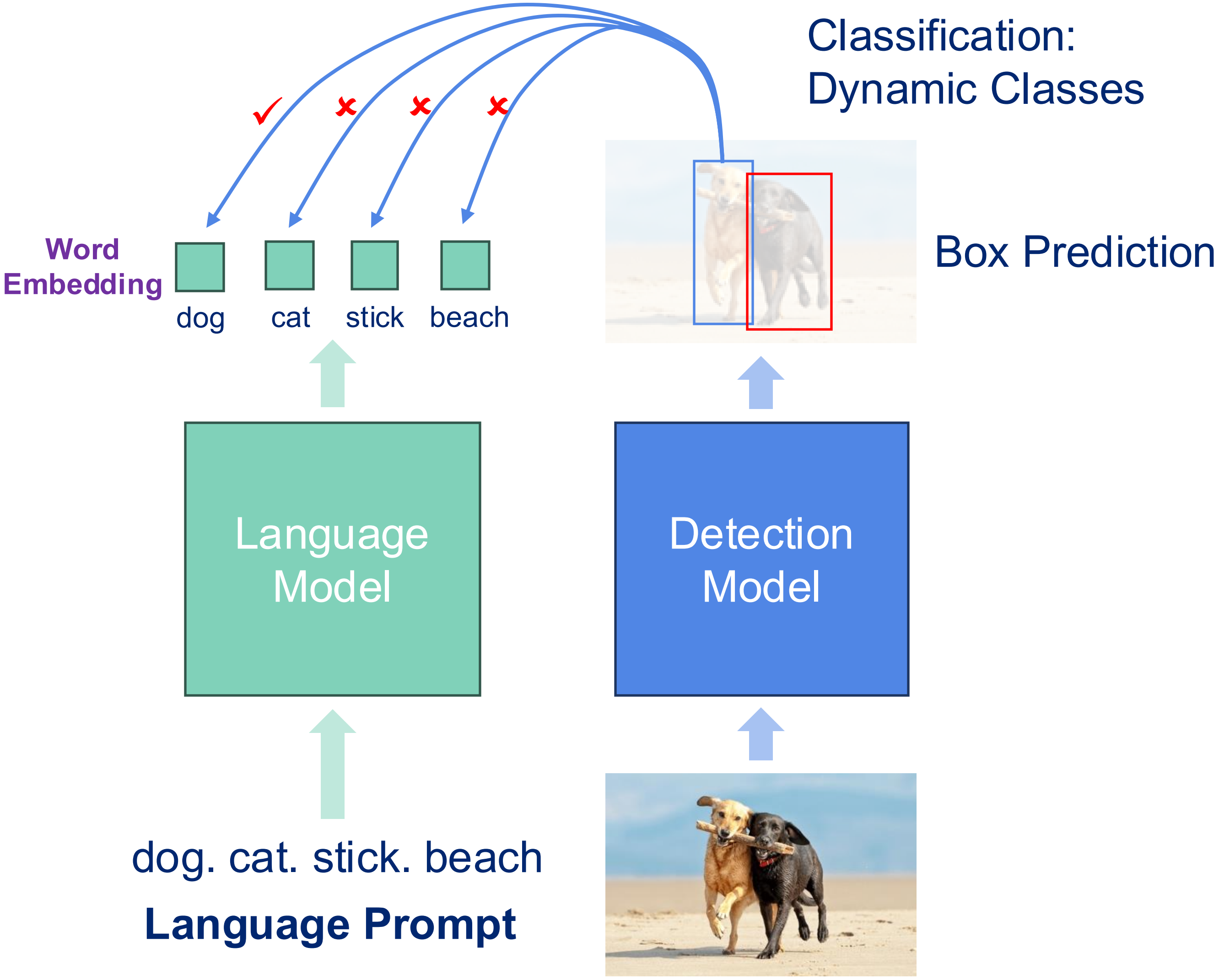
https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/

边缘计算设备部署 (NVIDIA Orin NX)

[redacted]



文本提示 v.s. 视觉提示



Classification:
Dynamic Classes

✓ ✗ ✗ ✗

Word Embedding

dog cat stick beach

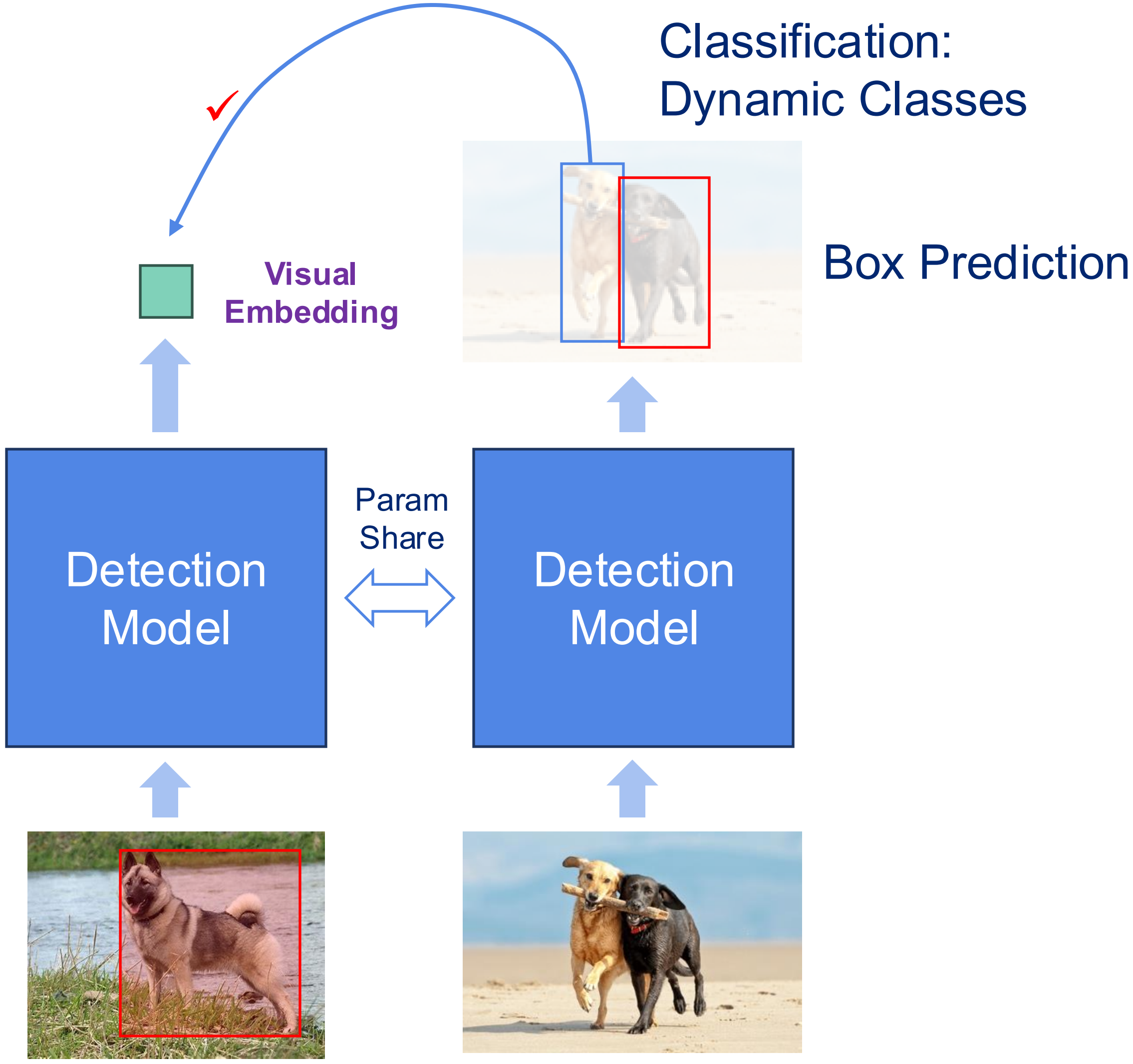
Box Prediction

Language Model

Detection Model

dog. cat. stick. beach

Language Prompt



Classification:
Dynamic Classes

✓

Visual Embedding

Box Prediction

Detection Model

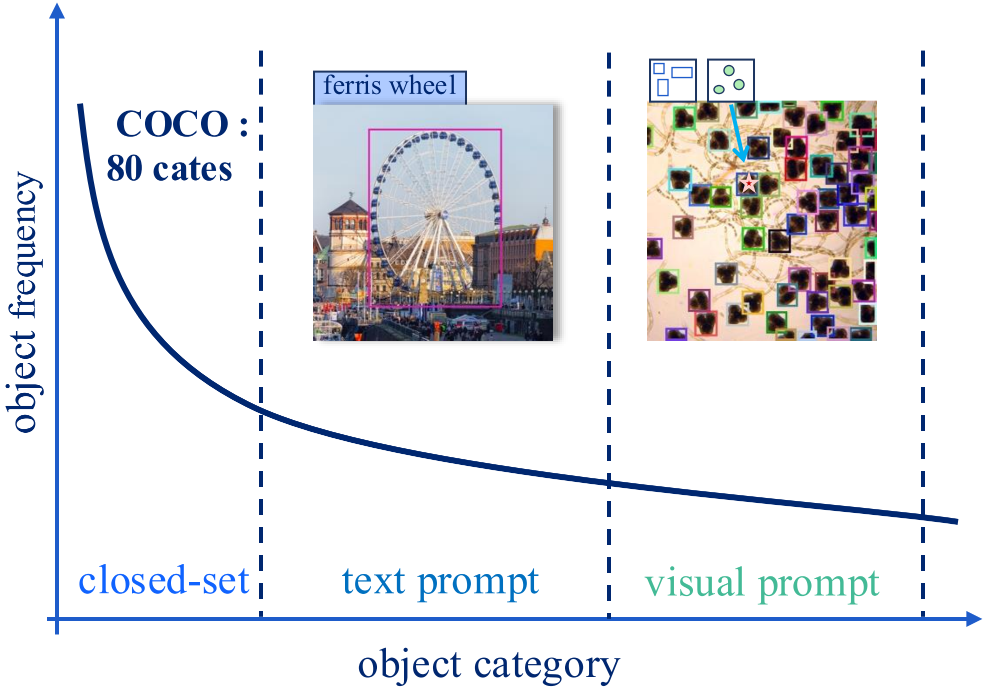
Param Share

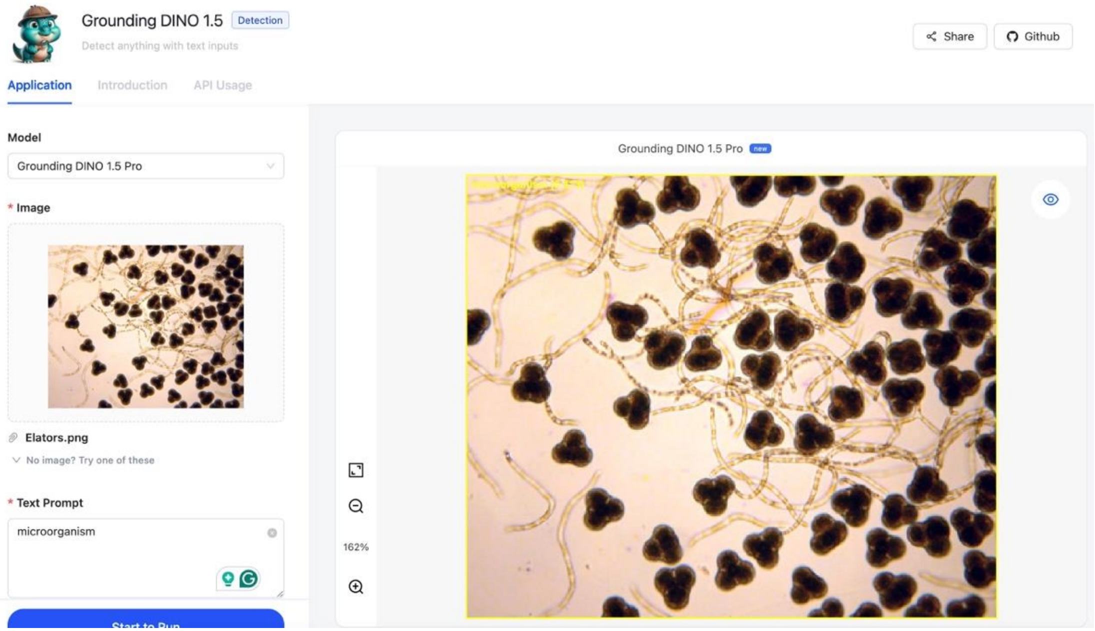
Detection Model

Visual Prompt

基于文本提示的方法面临的困境

- 可以使用自然语言描述待检测物体
- 需要进行文本与视觉模态的对其，受长尾数据短缺的影响
- 存在大量物体无法用语言进行描述





Model
Grounding DINO 1.5 Pro

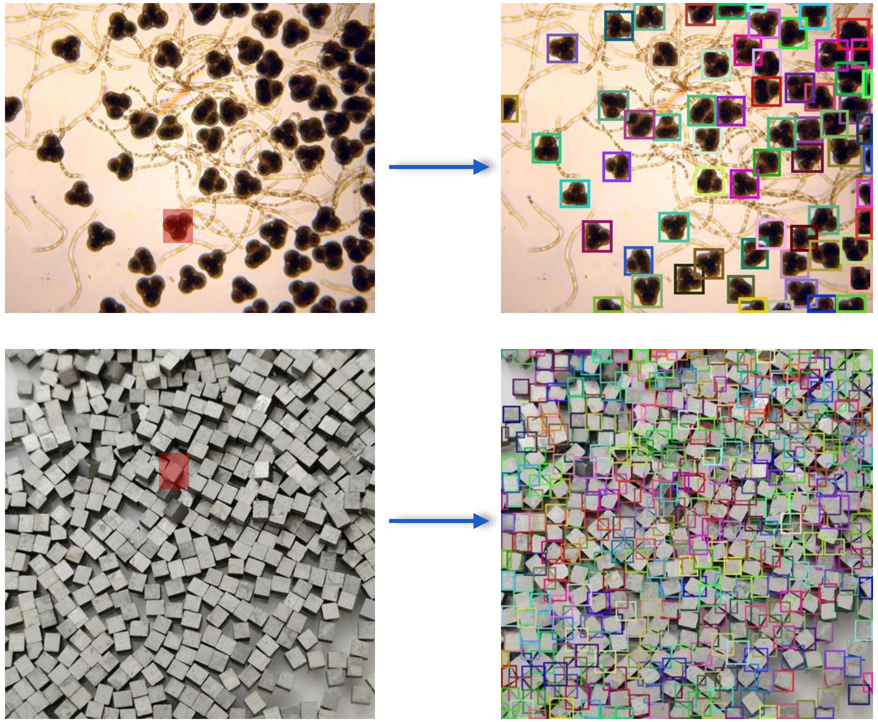
* Image
Elators.png

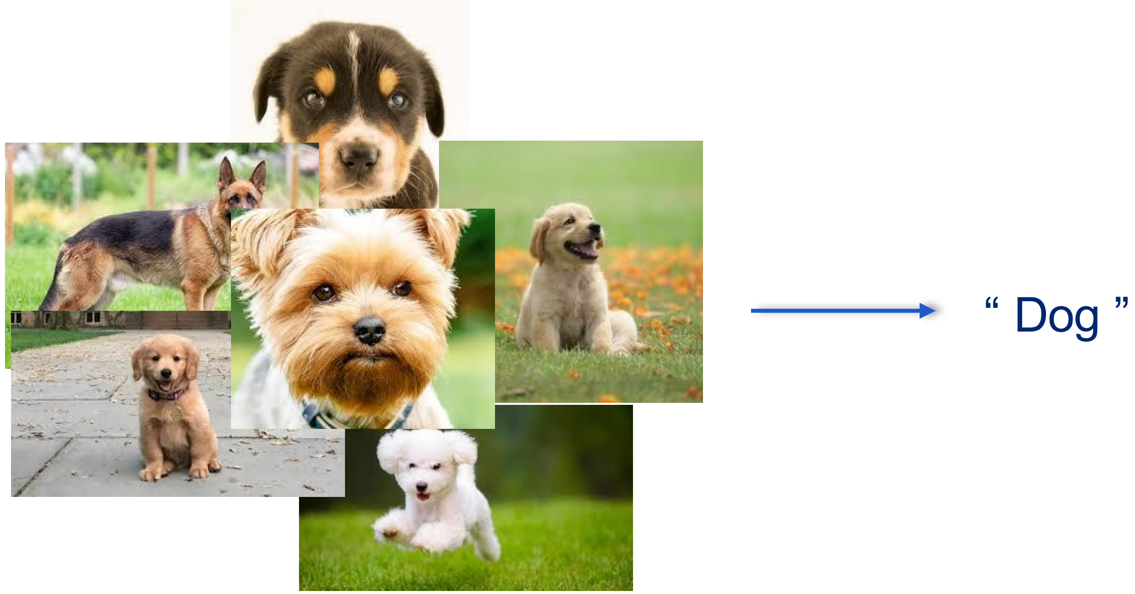
* Text Prompt
microorganism

基于视觉提示的方法面临的困境

- 可以通过视觉样例来表示待检测物体

- 难以很好的表征通用概念

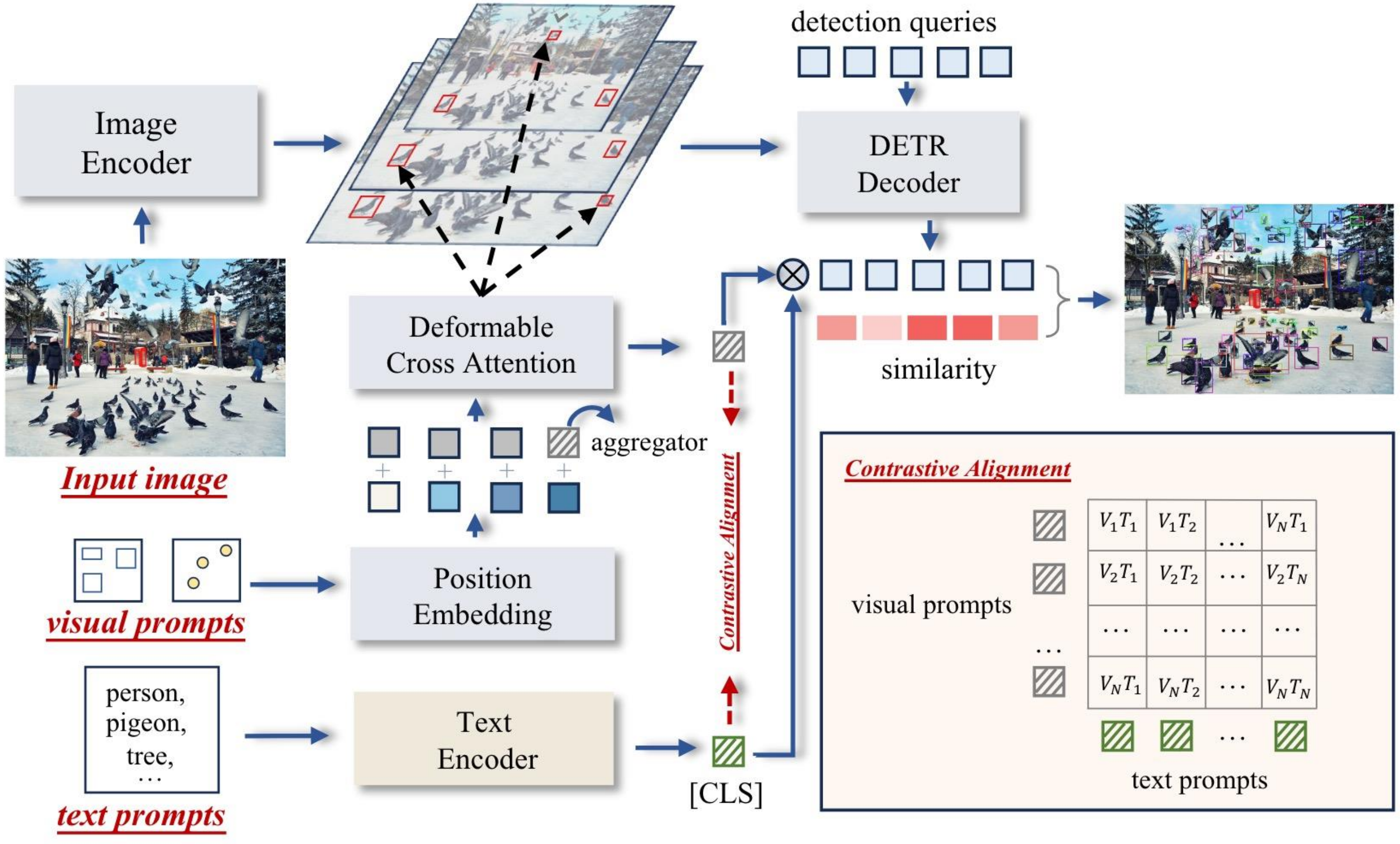




需要大量的样本来表示一个通用的概念

T-Rex2: 视觉提示与文本提示的融合

T-Rex2: 视觉提示与文本提示的融合



DINO-based End-to-End model

Visual Prompt Encoder: Deformable Cross Attention

$$B = \mathrm{Linear}(\mathrm{PE}(b_1, \ldots b_K); \theta_B) : \mathbb{R}^{K\times 4D} \rightarrow \mathbb{R}^{K\times D}$$

$$P = \mathrm{Linear}(\mathrm{PE}(p_1, \ldots p_K); \theta_P) : \mathbb{R}^{K\times 2D} \rightarrow \mathbb{R}^{K\times D}$$

$$Q = \begin{cases} \mathrm{Linear}\left(\mathrm{CAT}\left([C; C'], [B; B']\right); \varphi_B\right), \mathrm{box} \\ \mathrm{Linear}\left(\mathrm{CAT}\left([C; C'], [P; P']\right); \varphi_P\right), \mathrm{point} \end{cases}$$

$$Q'_j = \begin{cases} \mathrm{MSDeformAttn}(Q_j, b_j, \{\boldsymbol{f}_i\}_{i=1}^{L}), \mathrm{box} \\ \mathrm{MSDeformAttn}(Q_j, p_j, \{\boldsymbol{f}_i\}_{i=1}^{L}), \mathrm{point} \end{cases}$$

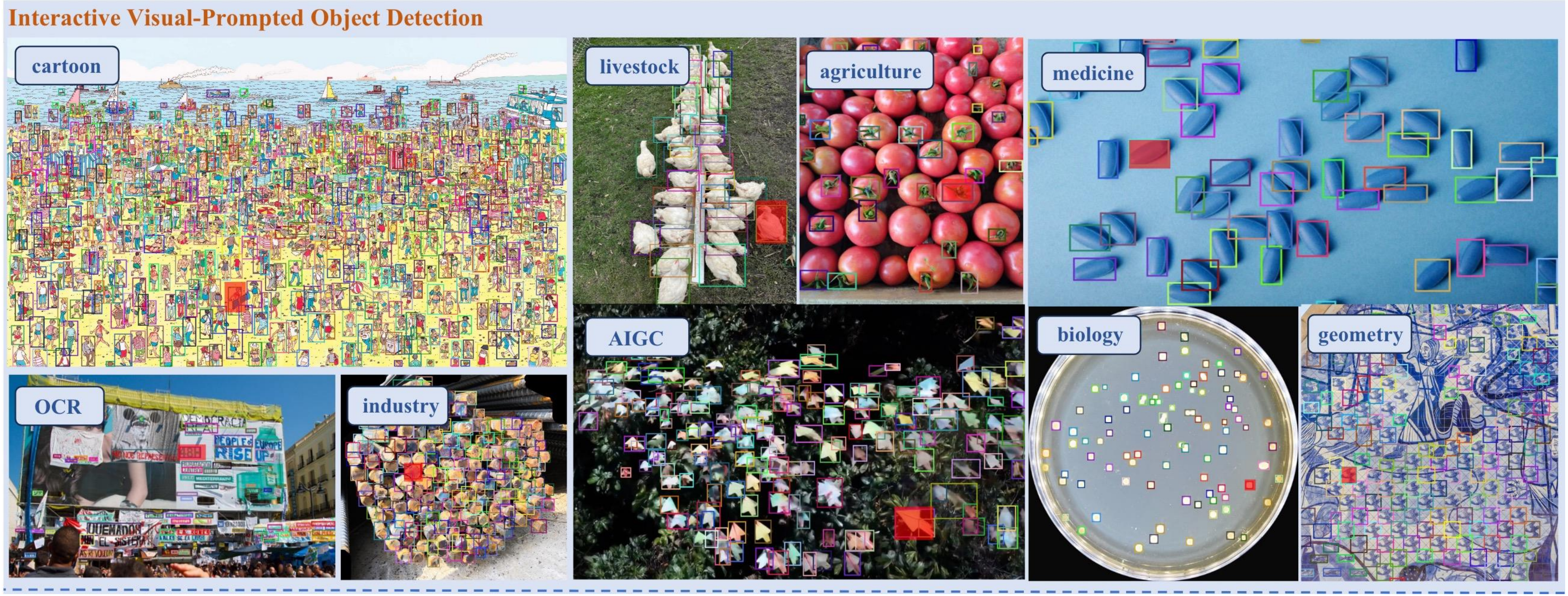
$$V = \mathrm{FFN}(\mathrm{SelfAttn}(Q'))[-1]$$

Text Prompt Encoder: CLIP

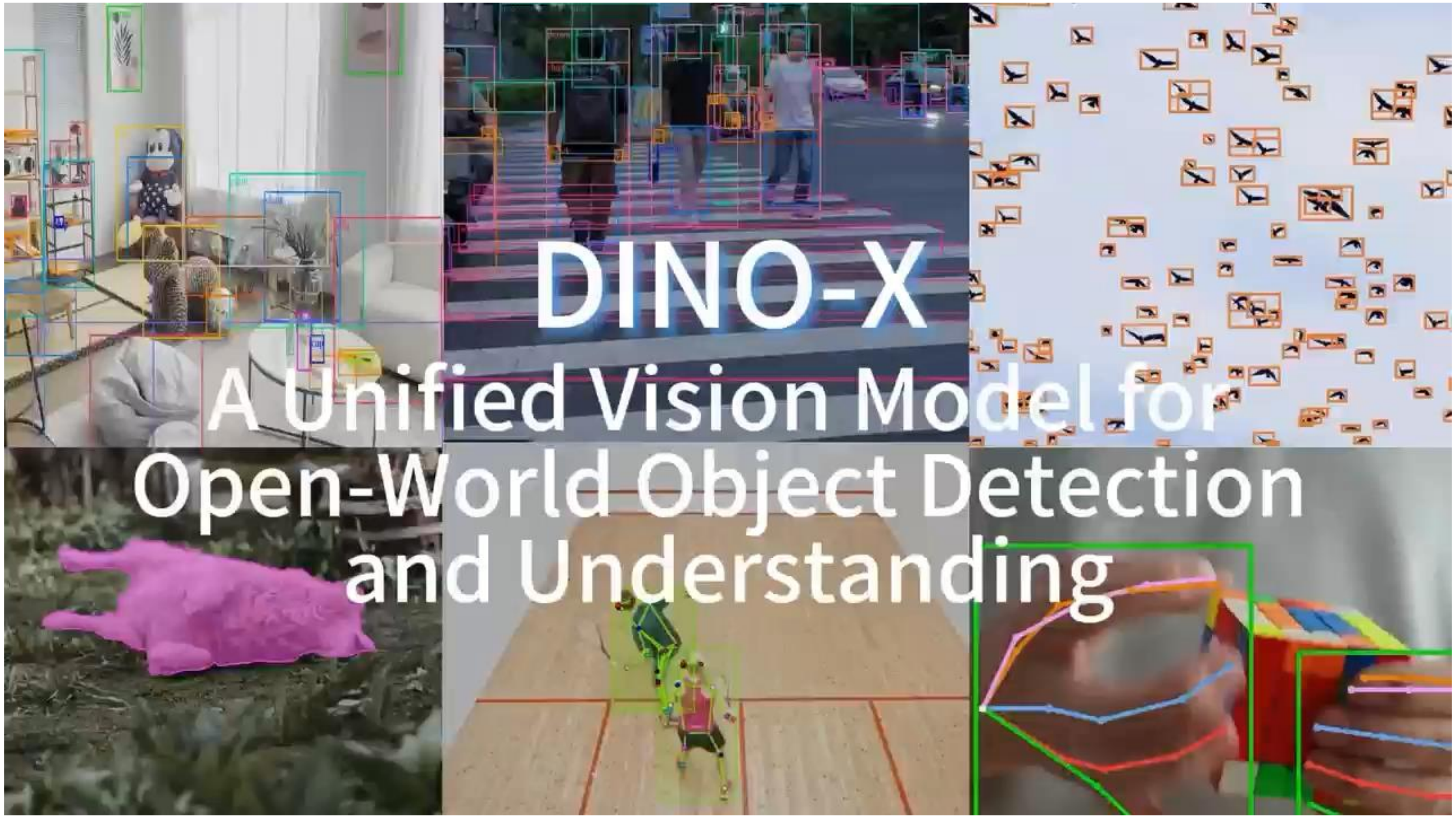
Modality Alignment: Contrastive Learning

$$\mathcal{L}_{align} = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(v_i \cdot t_i)}{\sum_{j=1}^{K} \exp(v_i \cdot t_j)}$$

T-Rex2 对于密集物体检测性能极佳



DINOX：集成更多视觉任务


DINO-X
A Unified Vision Model for Open-World Object Detection and Understanding

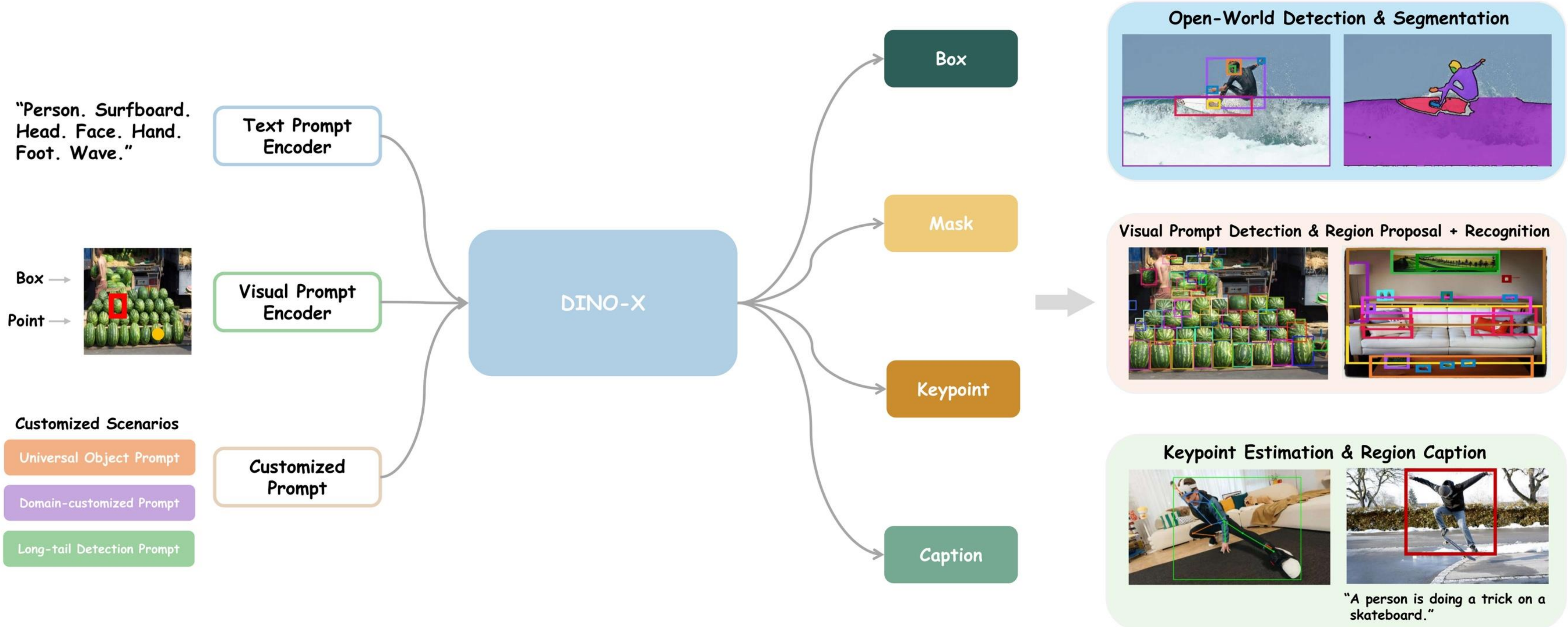
输入形式

- 文本提示
- 视觉提示
- 万物提示

输出形式

- 检测框
- 分割
- 关键点
- 语言描述

DINOX：模型结构



“Person. Surfboard. Head. Face. Hand. Foot. Wave.”

Text Prompt Encoder

Visual Prompt Encoder

Customized Scenarios

- Universal Object Prompt
- Domain-customized Prompt
- Long-tail Detection Prompt

Customized Prompt

DINO-X

Box

Mask

Keypoint

Caption

Open-World Detection & Segmentation

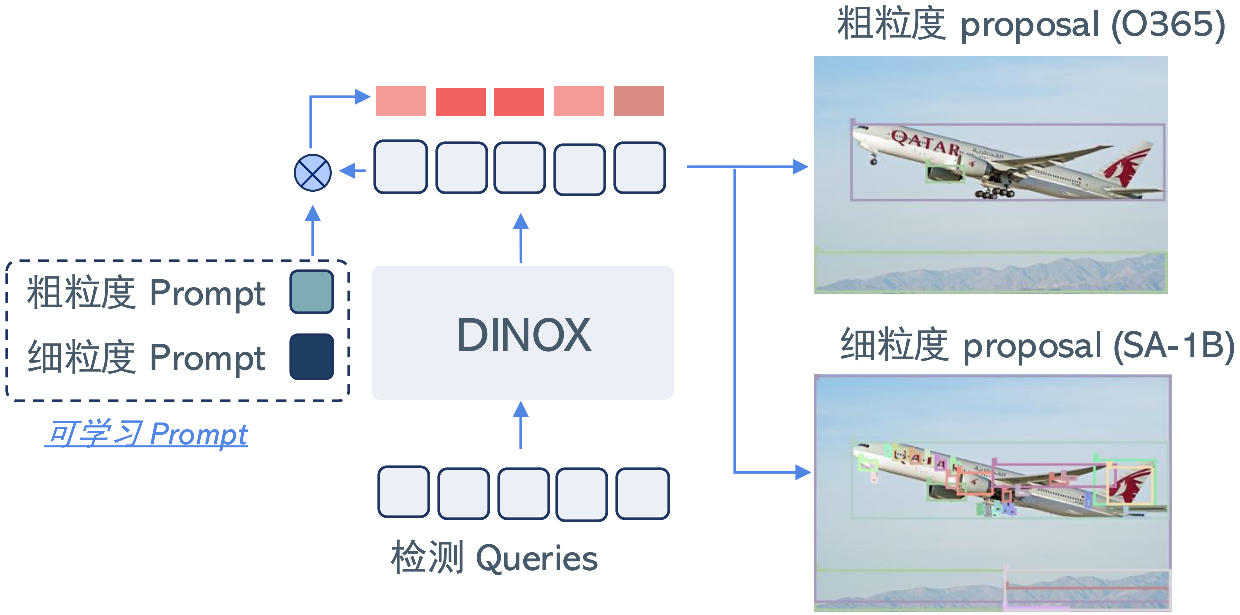
Visual Prompt Detection & Region Proposal + Recognition

Keypoint Estimation & Region Caption

“A person is doing a trick on a skateboard.”

万物提示工作流 (Universal Proposal + TinyLM)

Universal Proposal



粗粒度 Prompt

细粒度 Prompt

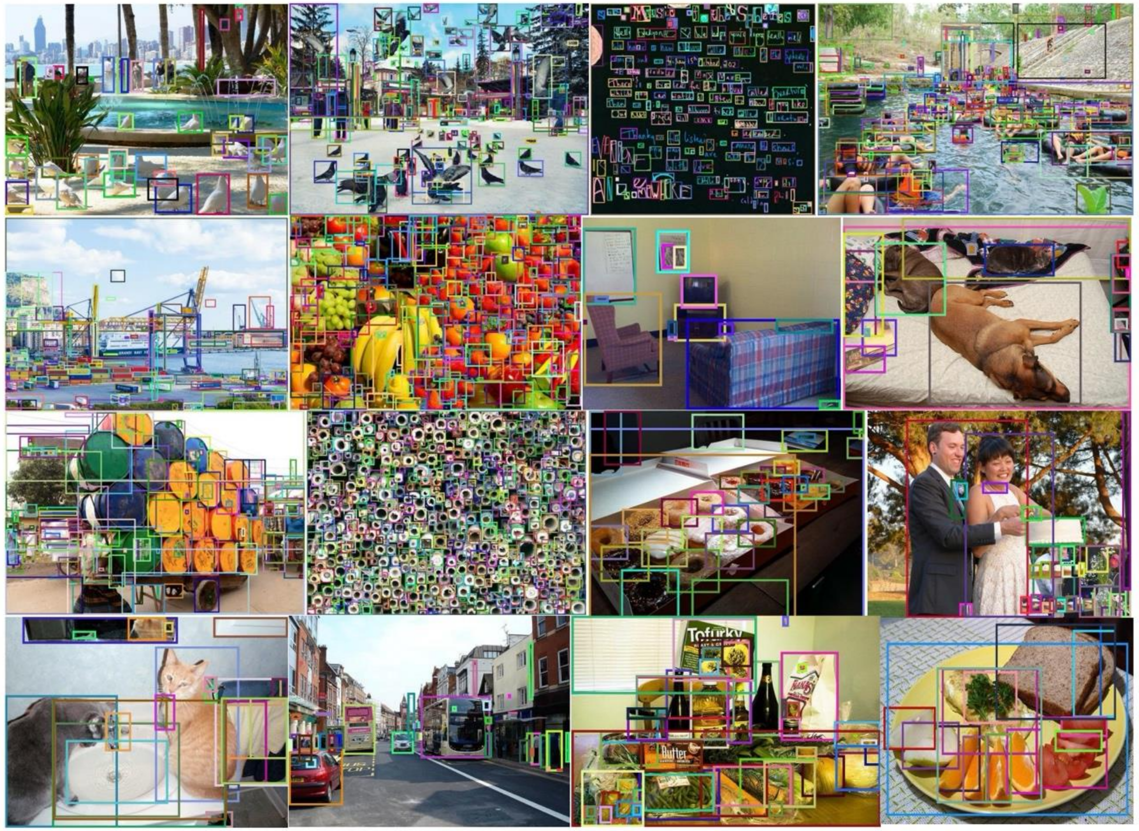
可学习 Prompt

DINOX

检测 Queries

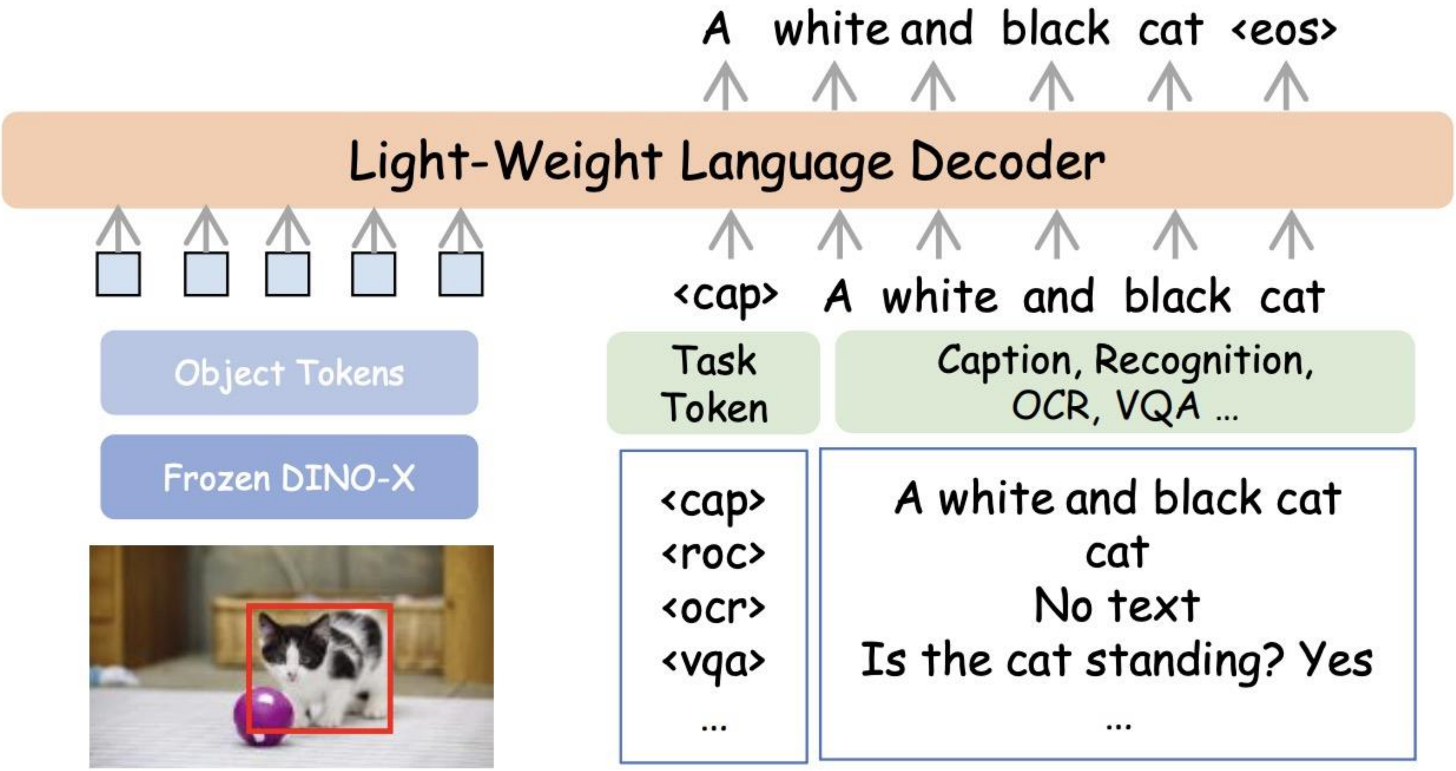
粗粒度 proposal (O365)

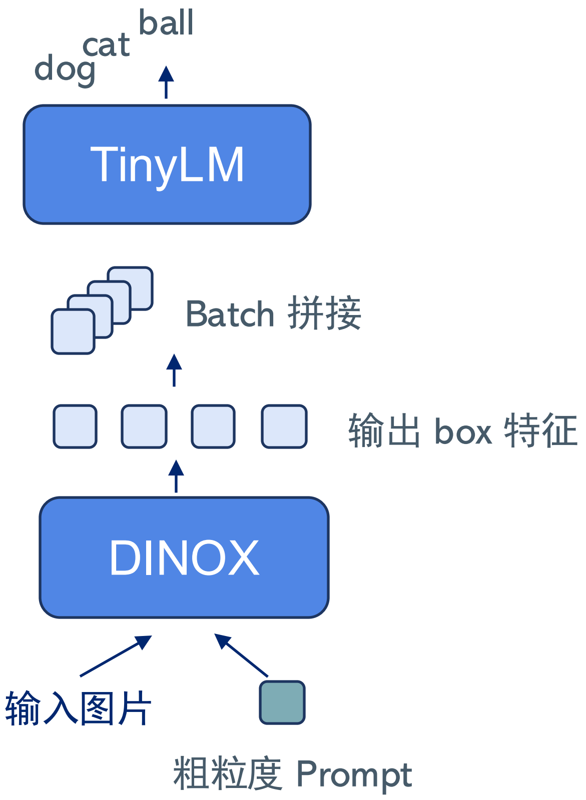
细粒度 proposal (SA-1B)



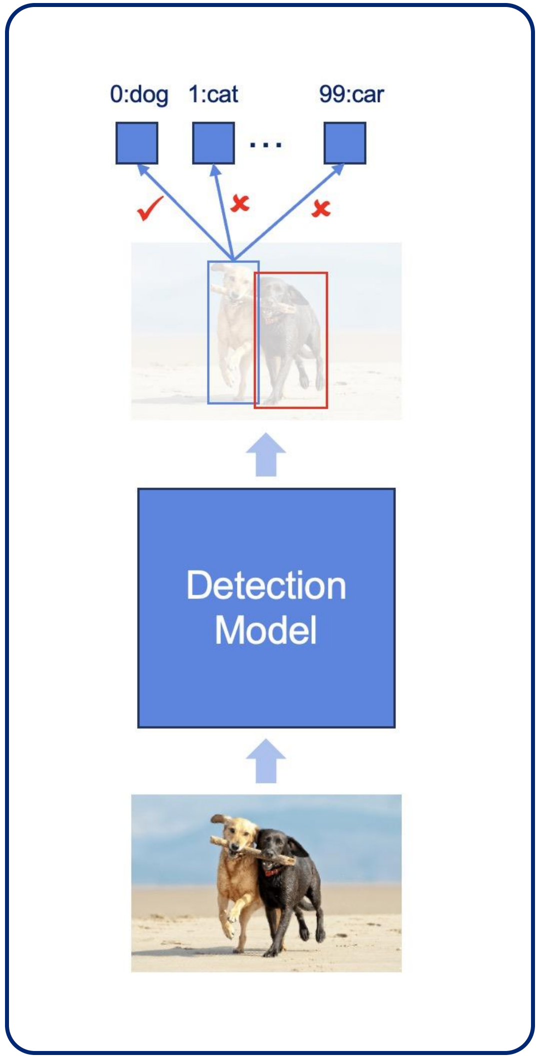
万物提示工作流（Universal Proposal + TinyLM）

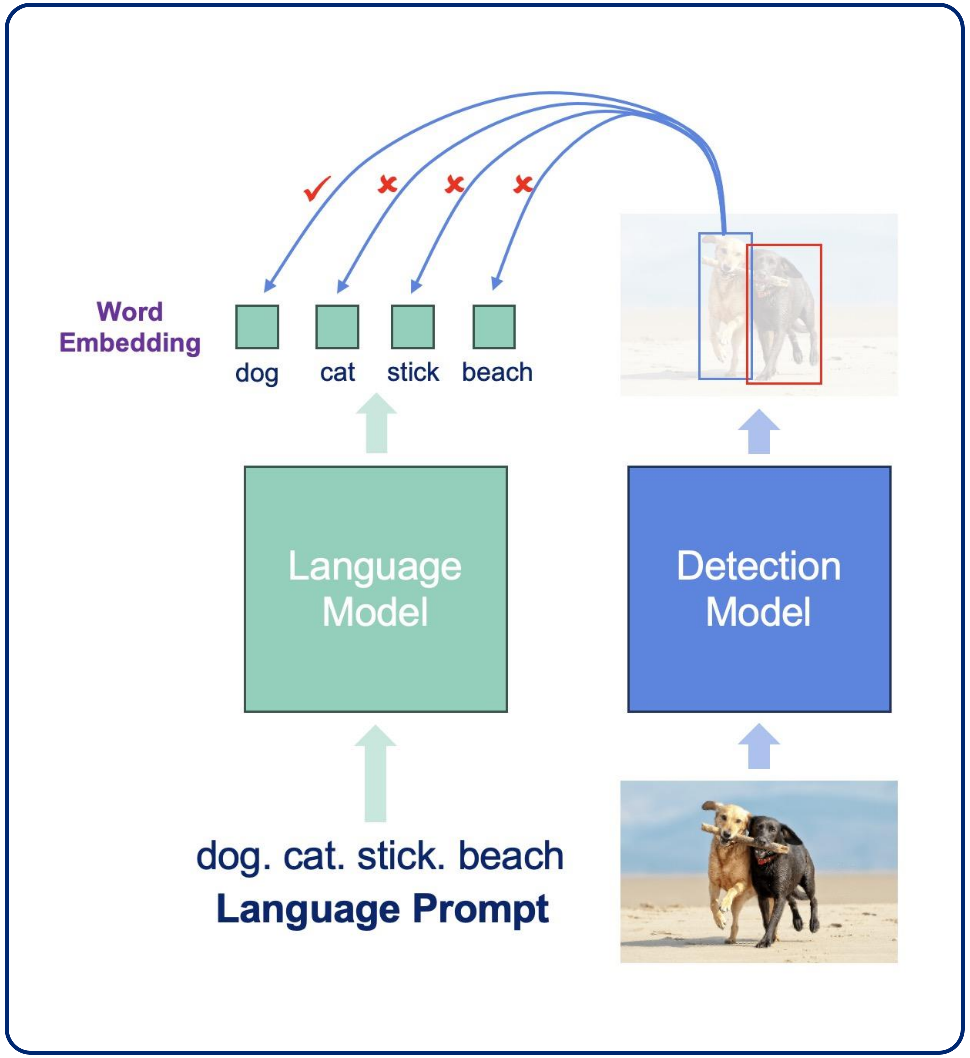
TinyLM

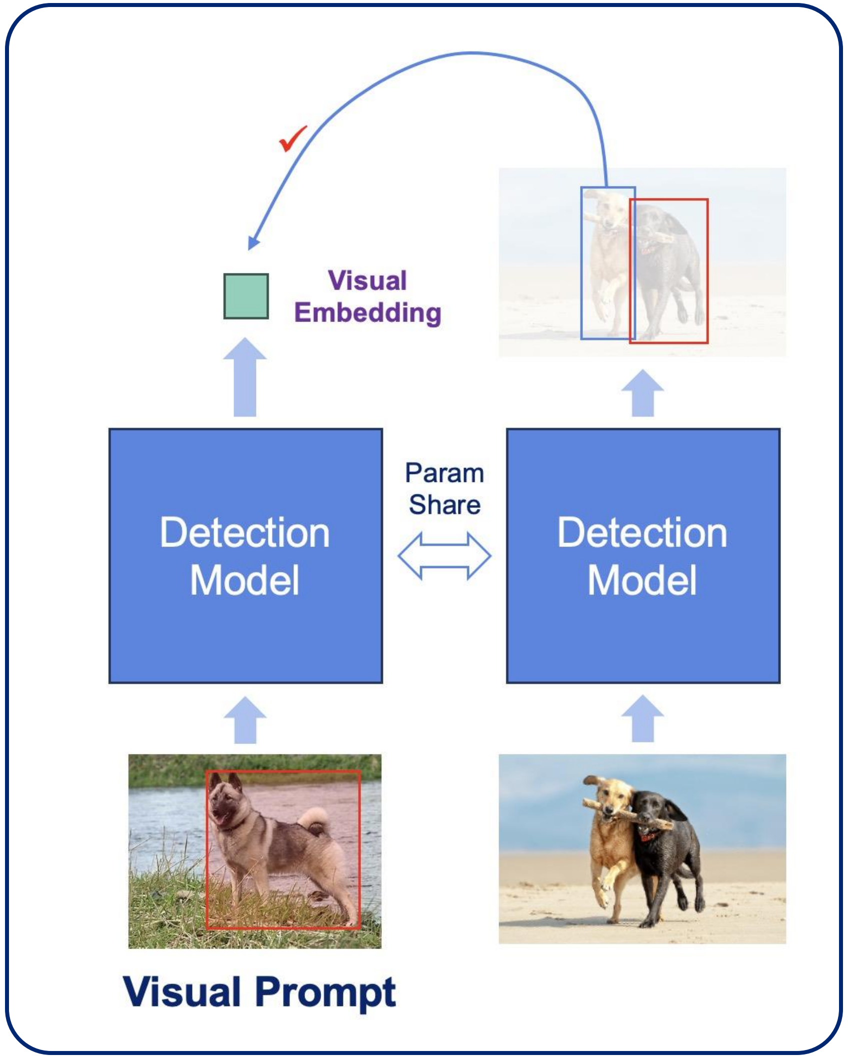

A white and black cat <eos>
Light-Weight Language Decoder
<cap> A white and black cat
Object Tokens
Frozen DINO-X
Task Token
Caption, Recognition, OCR, VQA ...
<cap>
<roc>
<ocr>
<vqa>
...
A white and black cat
cat
No text
Is the cat standing? Yes
...



目标检测下一步是什么?







闭集检测 → 文本提示开集检测 → 视觉提示开集检测

DETR **Grounding DINO** **T-Rex**

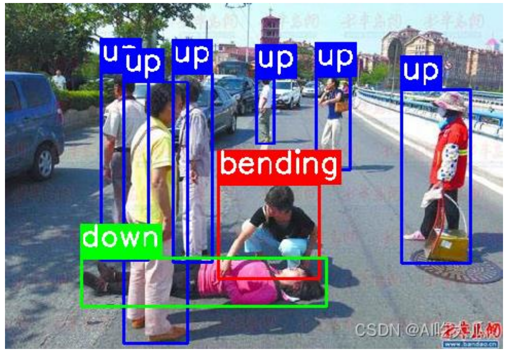
基于多模态大语言模型的目标检测模型

蒋擎

7-11

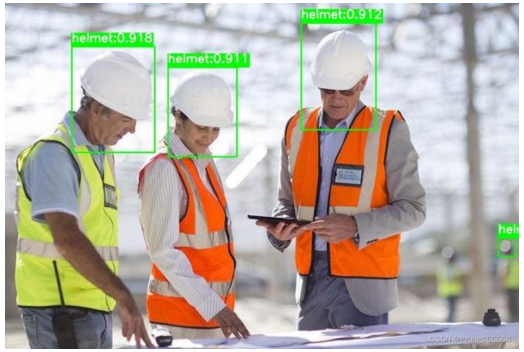
目标检测下一步是什么?

大量的可检测实体都可以用文本表示



摔倒检测

“person fallen”



佩戴安全帽检测

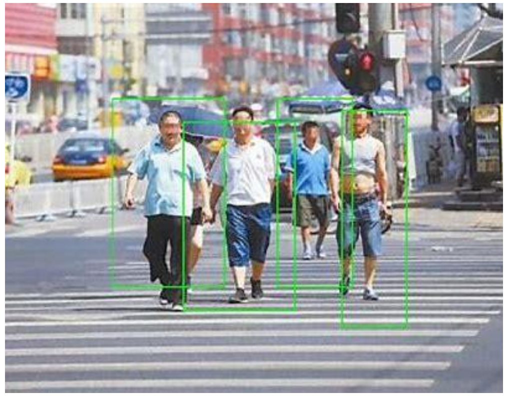
“person that are not wearing helmet”

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”

目标检测下一步是什么?

大量的可检测事件都可以用文本表示

"Incidents of street insecurity"

"Home invasion"

"inappropriate nursing"

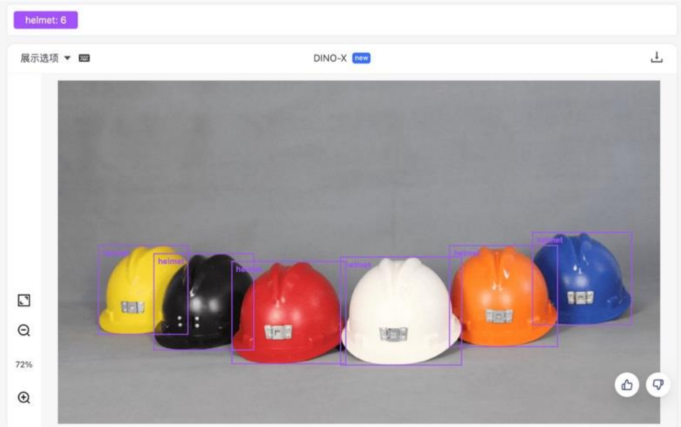
"Childcare"

"Traffic security"

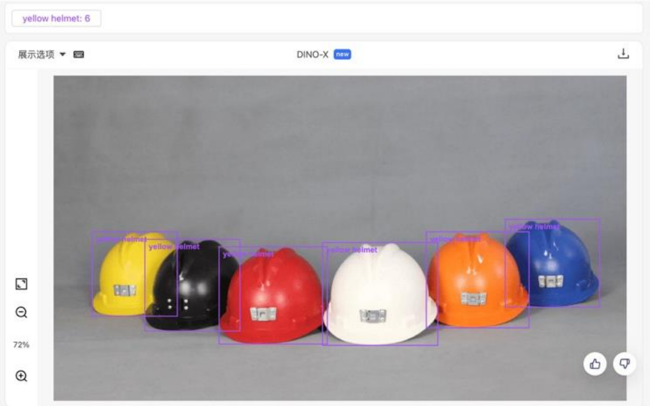
目标检测下一步是什么?

发现 1: SOTA 的开集检测模型缺乏语言理解能力

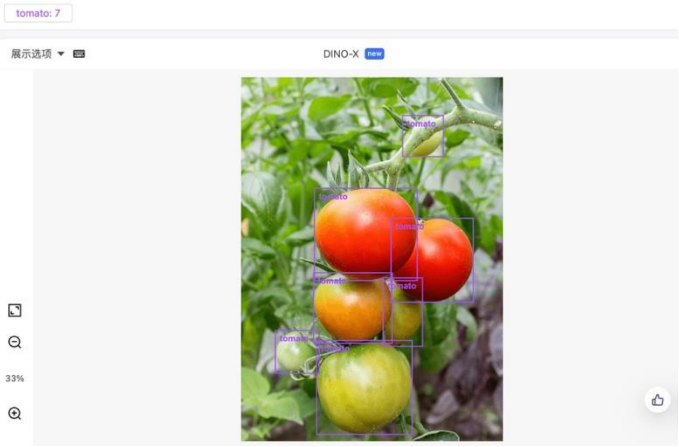
“helmet”



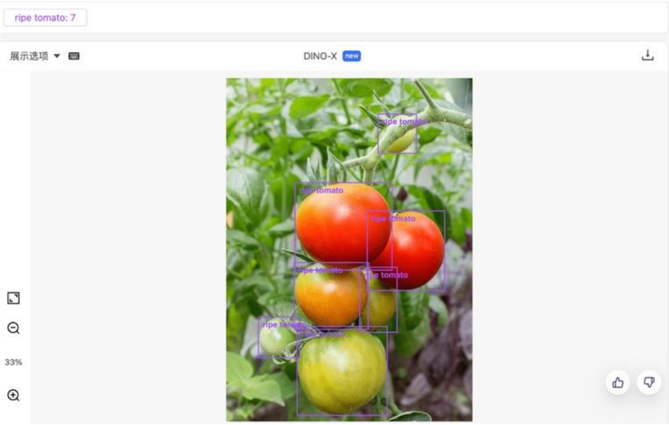
“yellow helmet”



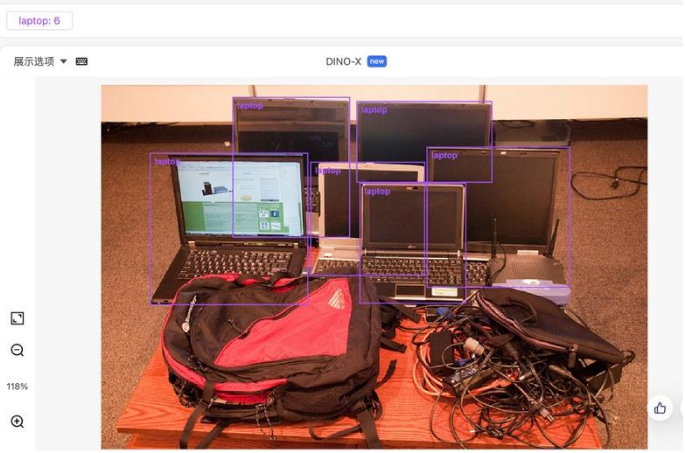
“tomato”



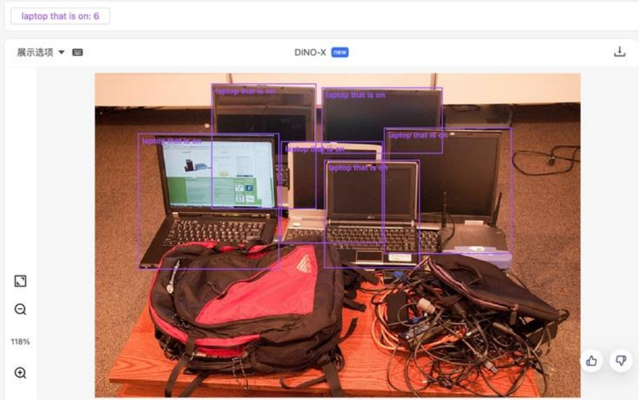
“ripe tomato”



“laptop”

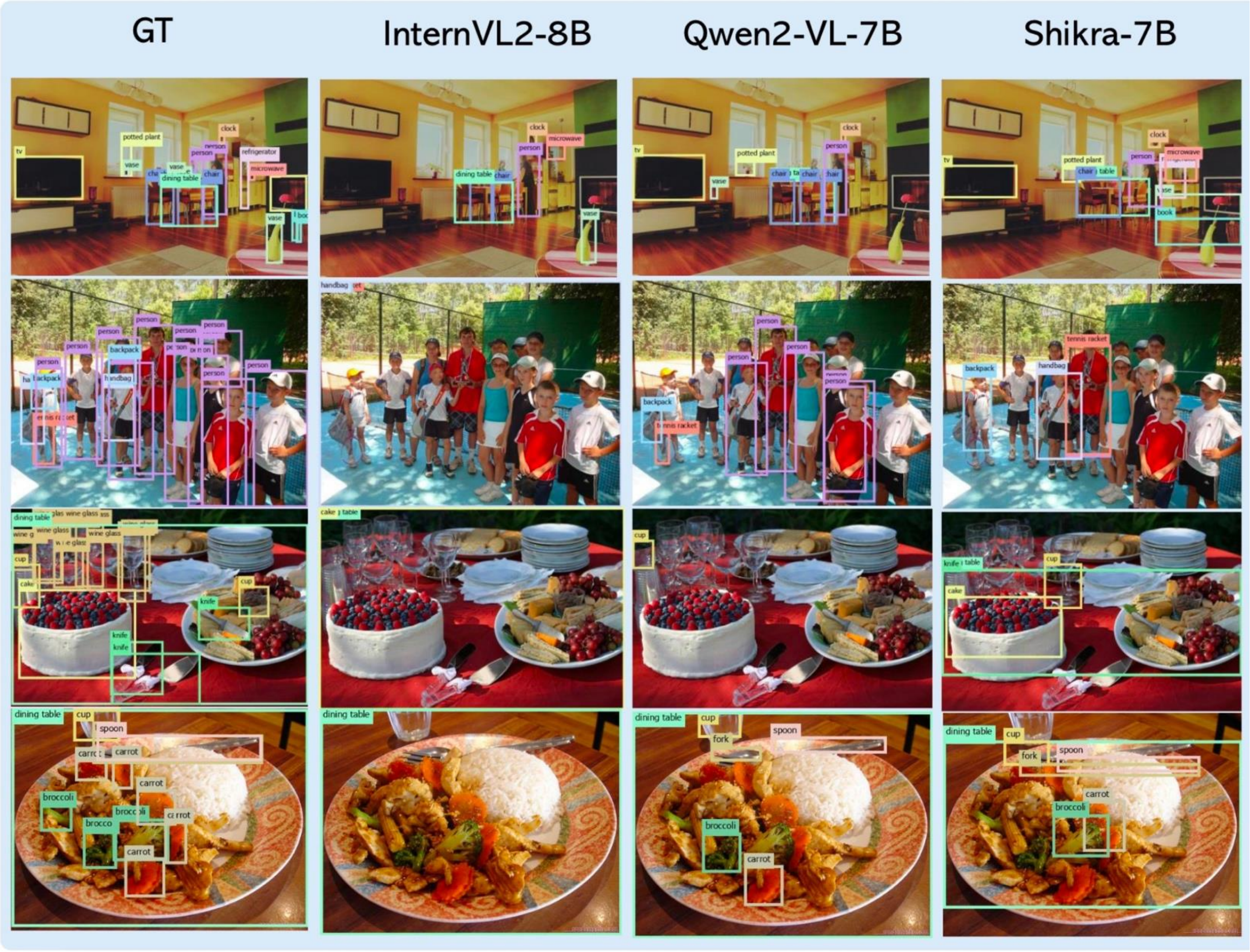


“laptop that is on”



目标检测下一步是什么?

发现 2: SOTA 的多模态大语言模型缺乏细粒度的感知能力



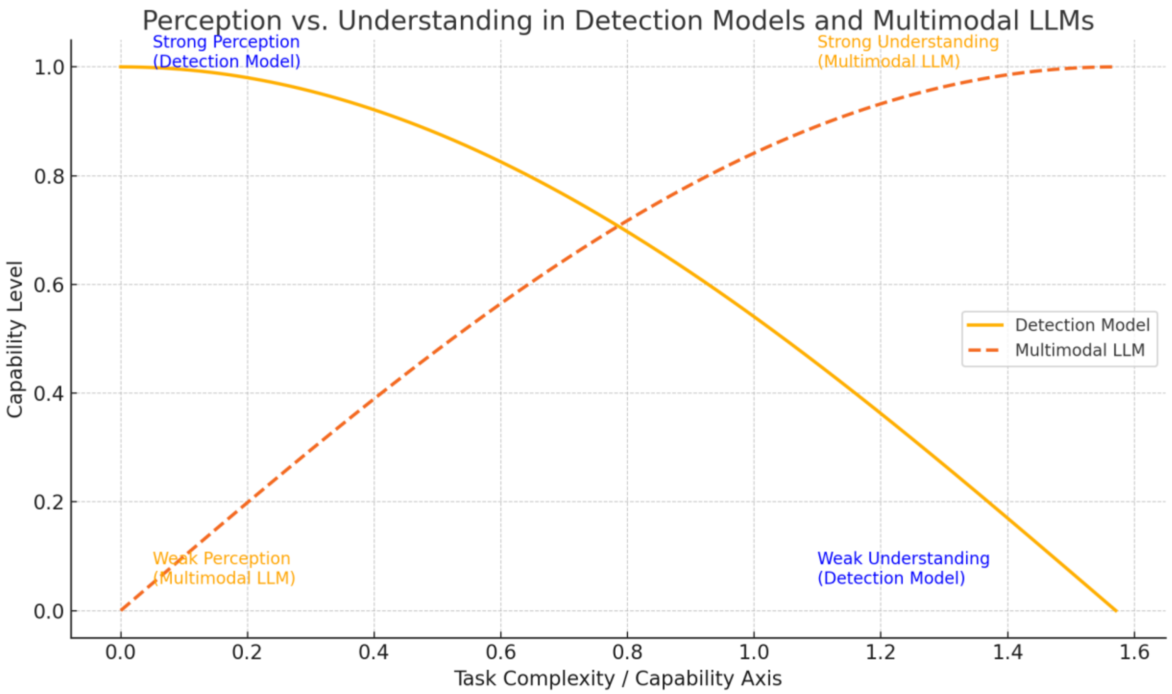
User: Please help me detect person in this image

MLLMs:

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

- coordinate shift
- tiny object detection
- dense object detection

目标检测下一步是什么?



检测模型：强感知, 弱理解

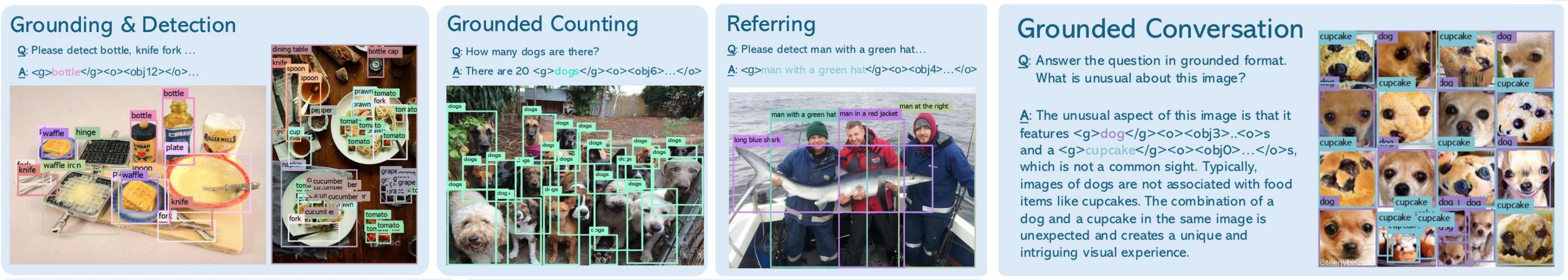
多模态大语言模型：弱感知, 强理解

下一步: 构建一个同时俱备强感知和强理解的多模态模型

ChatRex: Taming Multimodal LLM for Joint Perception and Understanding

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



Grounding & Detection

Q: Please detect bottle, knife fork …

A: <g>bottle</g><o><obj12></o>…

Grounded Counting

Q: How many dogs are there?

A: There are 20 <g>dogs</g><o><obj6>…</o>

Referring

Q: Please detect man with a green hat…

A: <g>man with a green hat</g><o><obj4>…</o>

Grounded Conversation

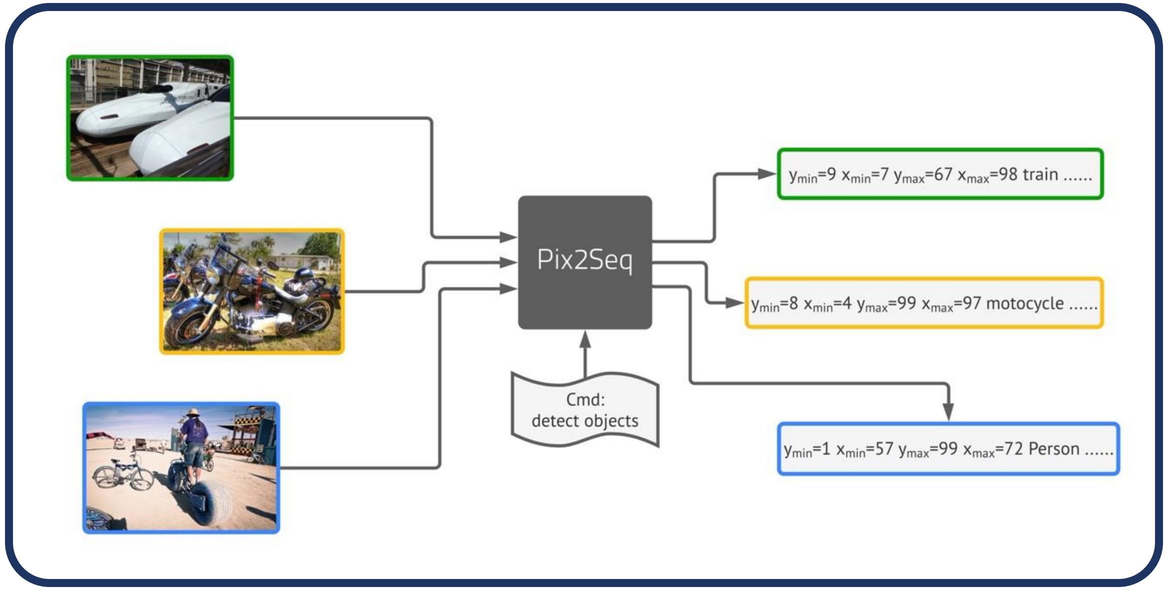
Q: Answer the question in grounded format. What is unusual about this image?

A: The unusual aspect of this image is that it features <g>dog</g><o><obj3>..<o>s and a <g>cupcake</g><o><obj0>…</o>s, which is not a common sight. Typically, images of dogs are not associated with food items like cupcakes. The combination of a dog and a cupcake in the same image is unexpected and creates a unique and intriguing visual experience.

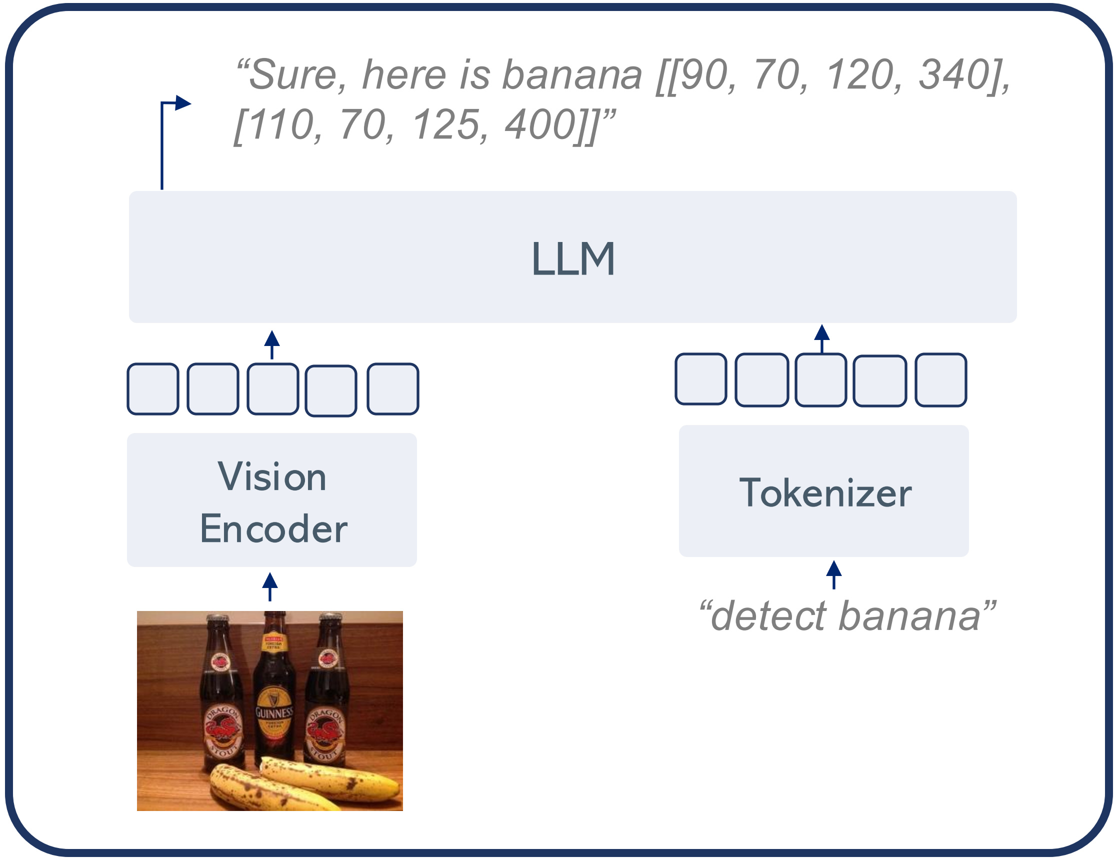
Jiang Q, Luo G, Yang Y, et al. Chatrex: Taming multimodal llm for joint perception and understanding[J]. arXiv preprint arXiv:2411.18363, 2024.

动机: 多模态大语言模型如何做检测?

将坐标当作文本来直接预测[1].


$y_{min}=9$ $x_{min}=7$ $y_{max}=67$ $x_{max}=98$ train ......
Pix2Seq
$y_{min}=8$ $x_{min}=4$ $y_{max}=99$ $x_{max}=97$ motocycle ......
Cmd:
detect objects
$y_{min}=1$ $x_{min}=57$ $y_{max}=99$ $x_{max}=72$ Person ......

Pix2Seq[1]



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

LLM

Vision Encoder

Tokenizer

“detect banana”

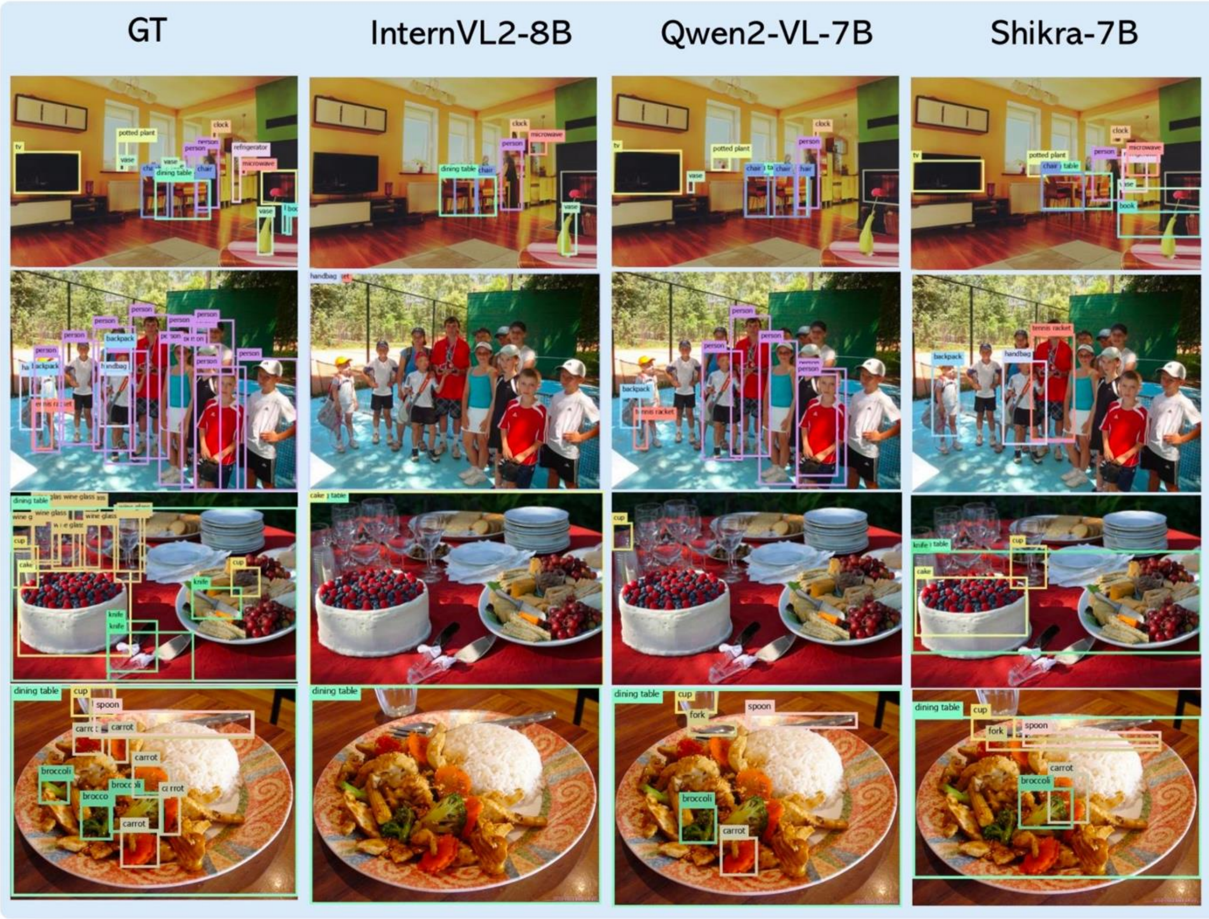
Modern MLLMs

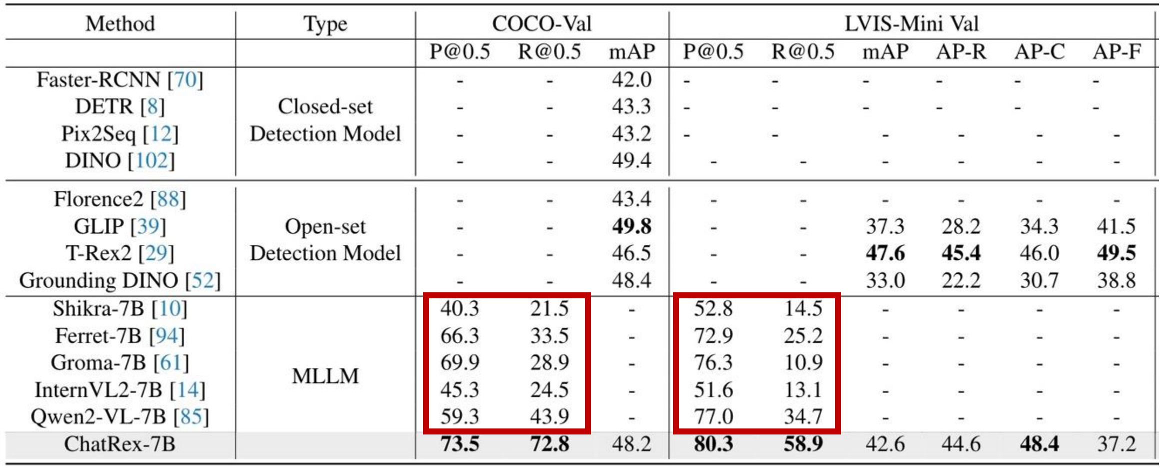
---

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

动机: 多模态大语言模型如何做检测?

但是多模态大语言模型的检测性能很差



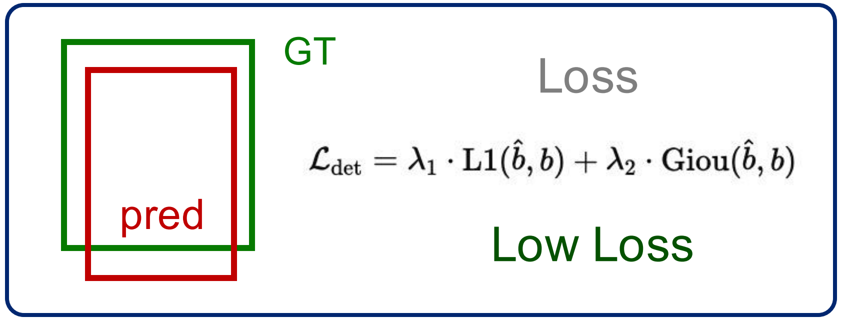


Method	Type	COCO-Val			LVIS-Mini Val					
		P@0.5	R@0.5	mAP	P@0.5	R@0.5	mAP	AP-R	AP-C	AP-F
Faster-RCNN [70]	Closed-set Detection Model	-	-	42.0	-	-	-	-	-	-
DETR [8]		-	-	43.3	-	-	-	-	-	-
Pix2Seq [12]		-	-	43.2	-	-	-	-	-	-
DINO [102]		-	-	49.4	-	-	-	-	-	-
Florence2 [88]	Open-set Detection Model	-	-	43.4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

动机: 挑战在哪?

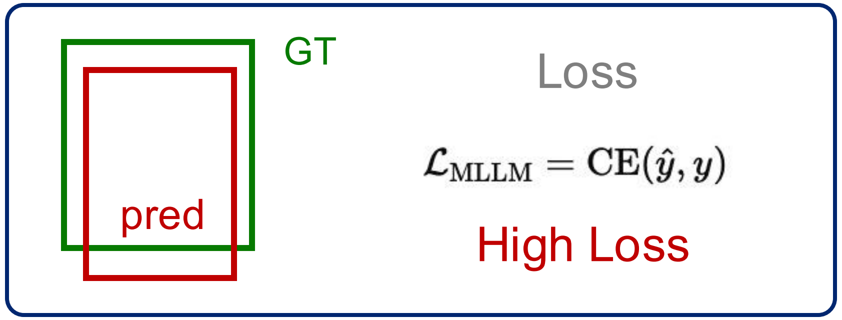
1. **Directly predict the coordinates is a hard task:** Regression V.S. Classification



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

Low Loss

detection model training



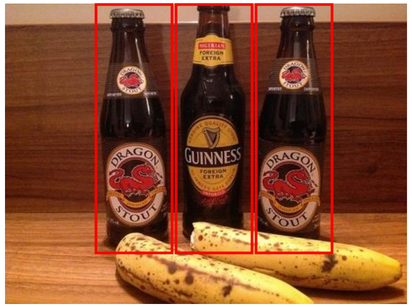
$$\mathcal{L}_{\text{MLLM}} = \text{CE}(\hat{y}, y)$$

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.

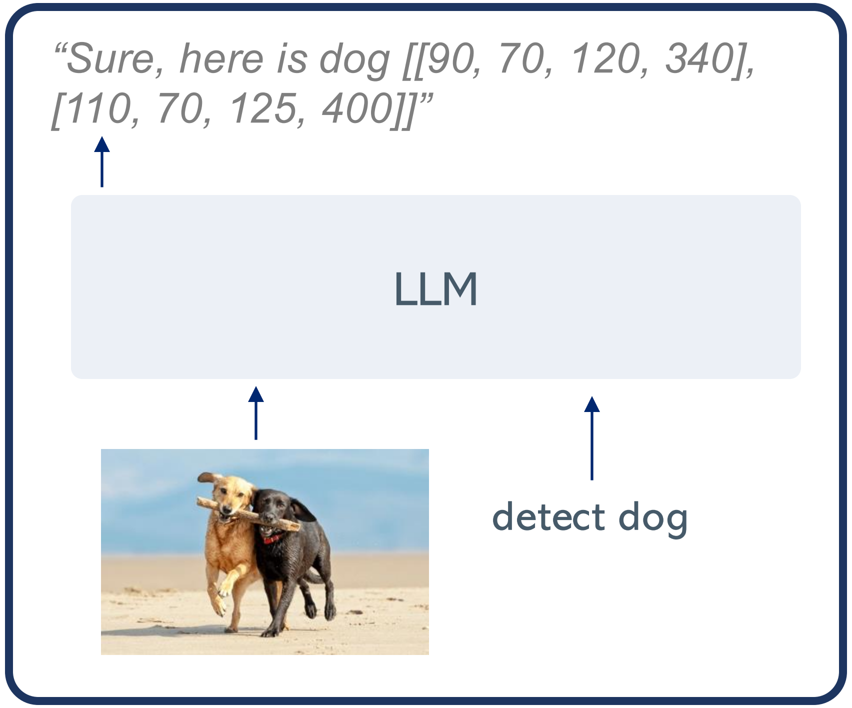


“bottle1, bottle2, bottle3” *“bottle3, bottle2, bottle1”* *“bottle2, bottle1, bottle3”*

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

解决方案:基于检索的感知模型

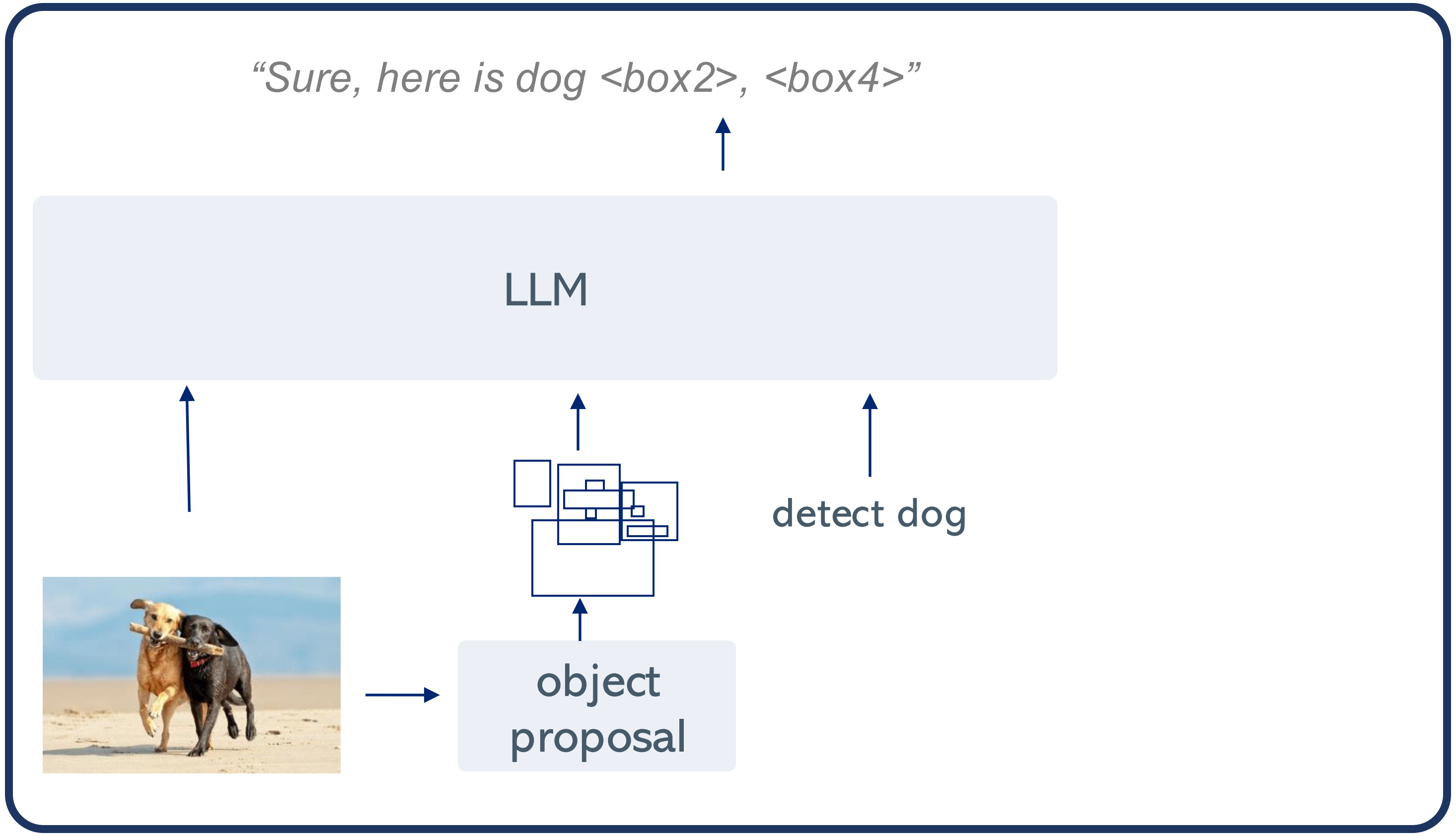
Core idea: LLM has strong understanding capability, while detection model has strong perception capability



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

LLM

detect dog



“Sure, here is dog <box2>, <box4>”

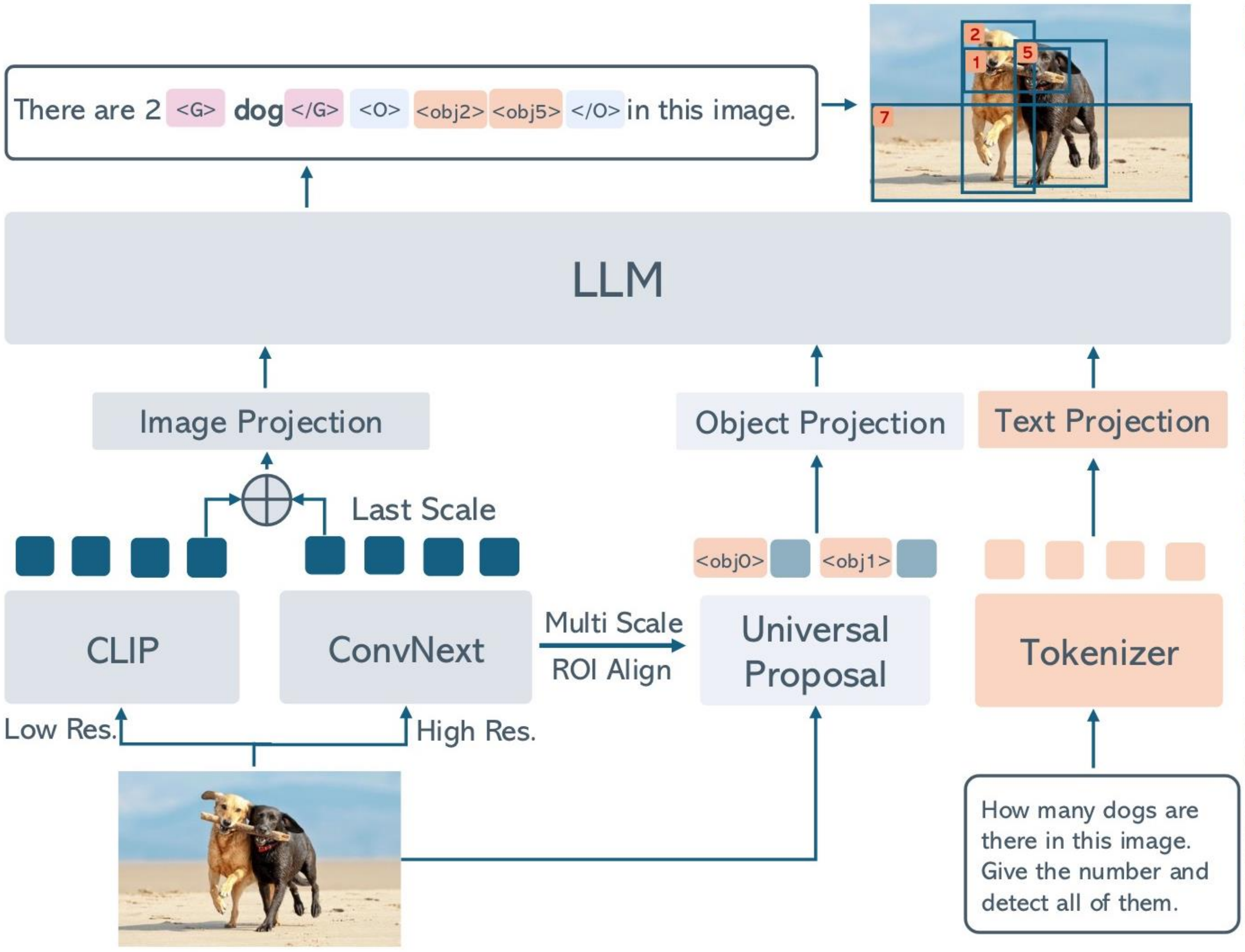
LLM

detect dog

object proposal

Retrieval based method

ChatRex: Detection-Oriented MLLM



There are 2 <G> dog </G> <O> <obj2> <obj5> </O> in this image.

How many dogs are there in this image. Give the number and detect all of them.

■ Vision Token ■ Object Token

■ Text Token <obj1> Object Index Token

<G> Grounding Start Token </G> Grounding End Token

<O> Object Start Token </O> Object End Token

Task1: Detection and Grounding

Q: Please detect the yellow dog in this image.

A: <G> the yellow dog </G> <O> <obj2> </O>

Task2: Region Understanding

Q: Give me a brief description of <obj5>

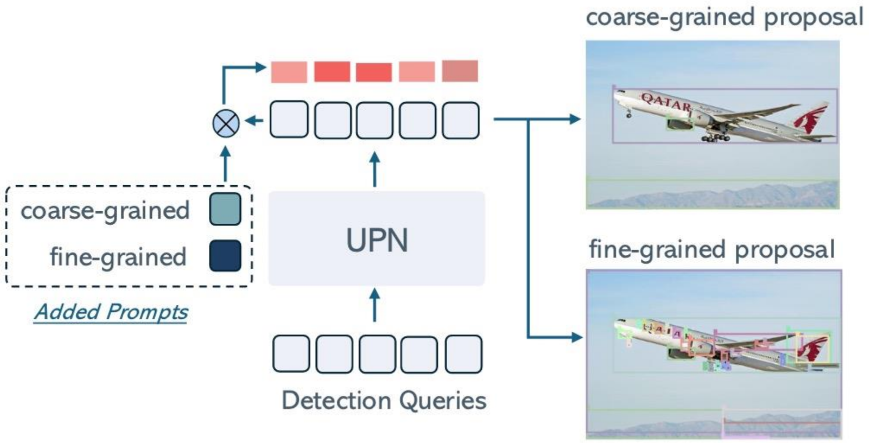
A: <G> Black dog with a stick </G> <O> <obj5> </O>

Task3: Grounded Image Conversation

Q: What are the dogs doing?

A: The <G> dogs </G> <O> <obj2> <obj5> </O> are running on a <G> beach </G> <O> <obj7> </O> while carrying <G> a stick </G> <O> <obj1> </O> together

ChatRex: Detection-Oriented MLLM



coarse-grained proposal

coarse-grained 🟦

fine-grained ■

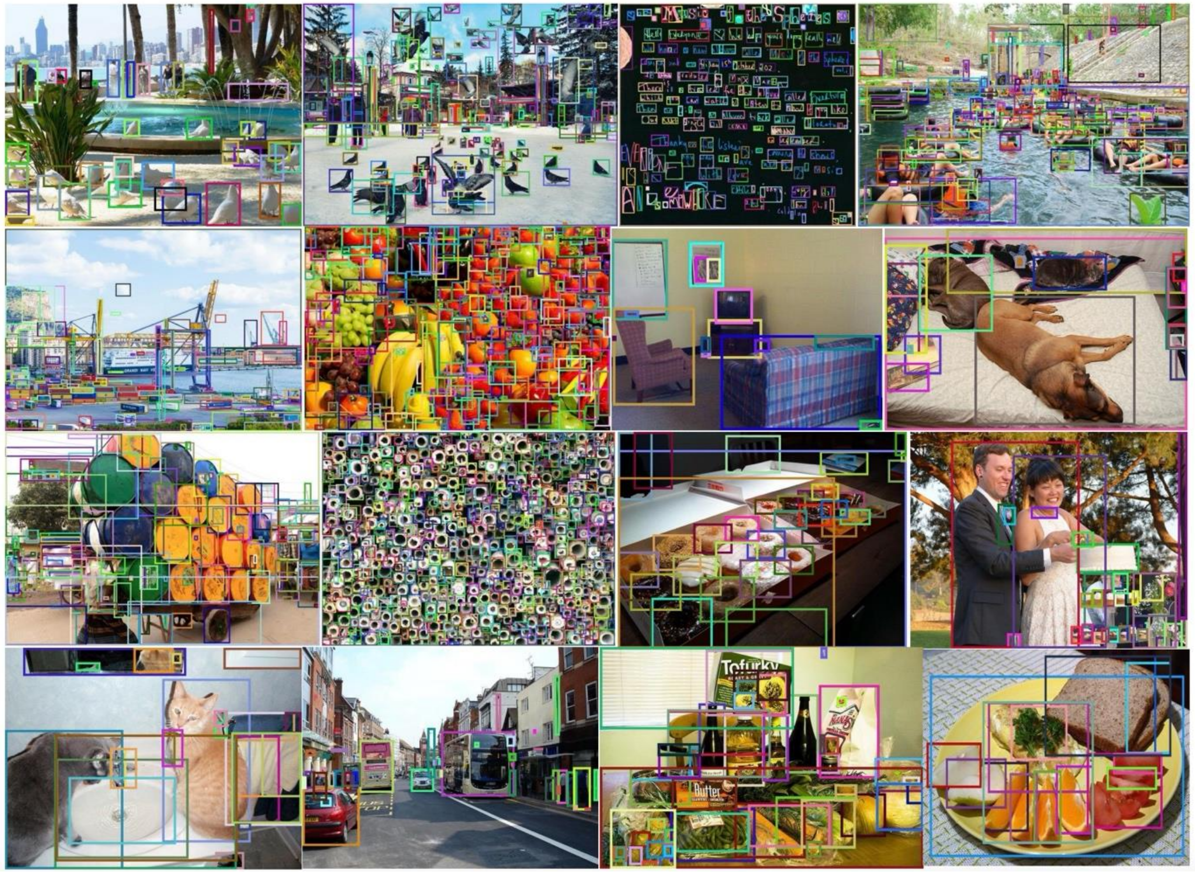
Added Prompts

UPN

fine-grained proposal

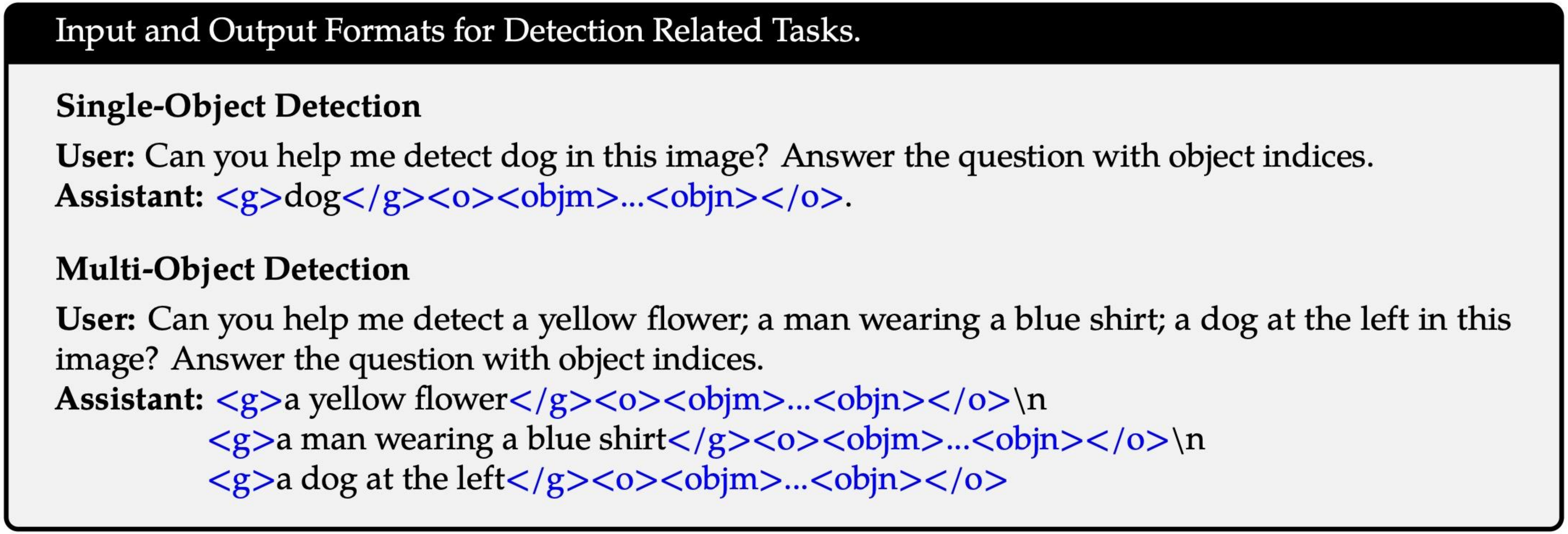
Detection Queries

Universal Proposal Architecture



ChatRex: Detection-Oriented MLLM

Use the input box indices as the output of detection



Input and Output Formats for Detection Related Tasks.

Single-Object Detection

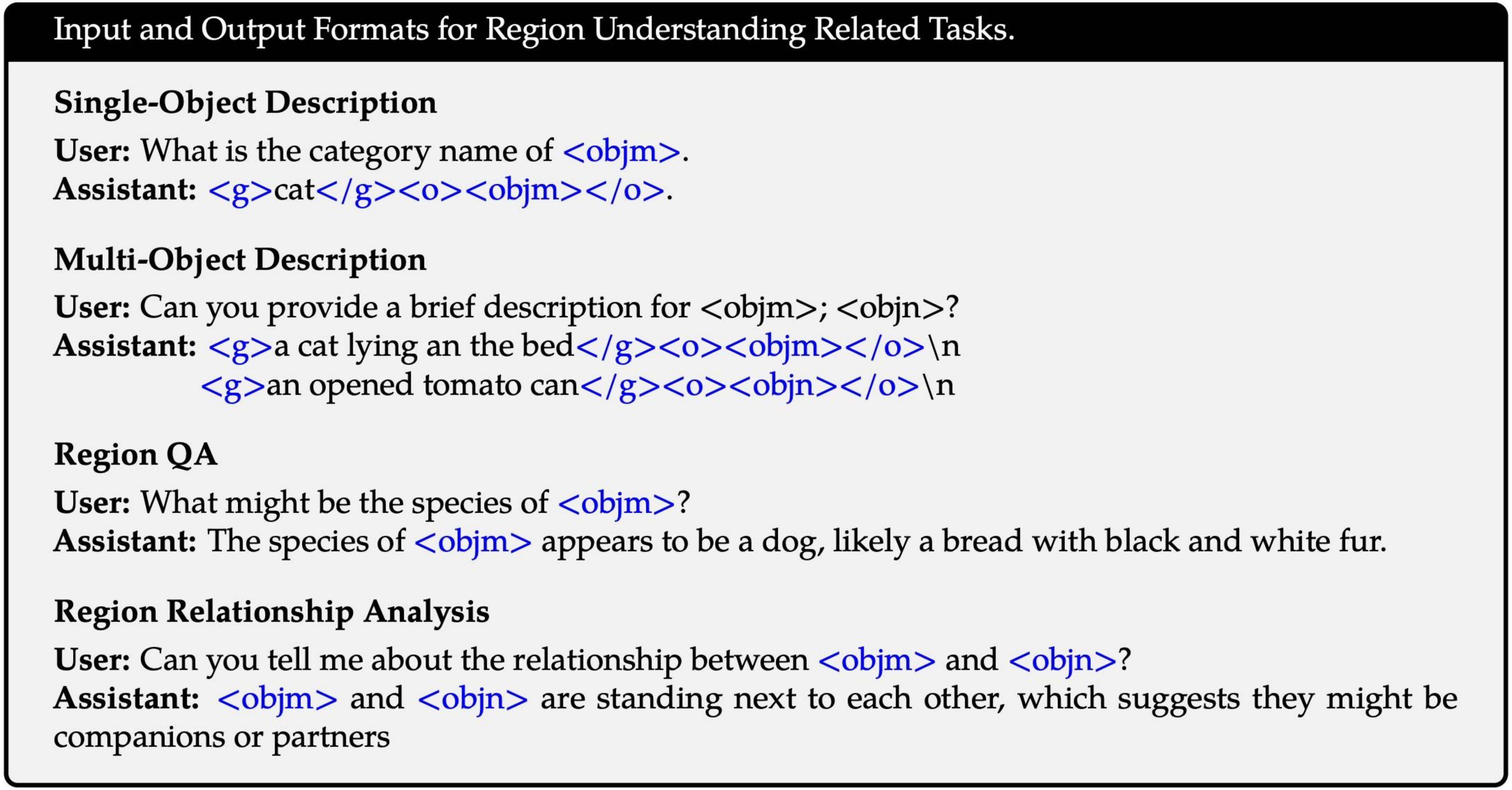
User: Can you help me detect dog in this image? Answer the question with object indices.
Assistant: <g>dog</g><o><objm>...<objn></o>.

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

应用: 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>

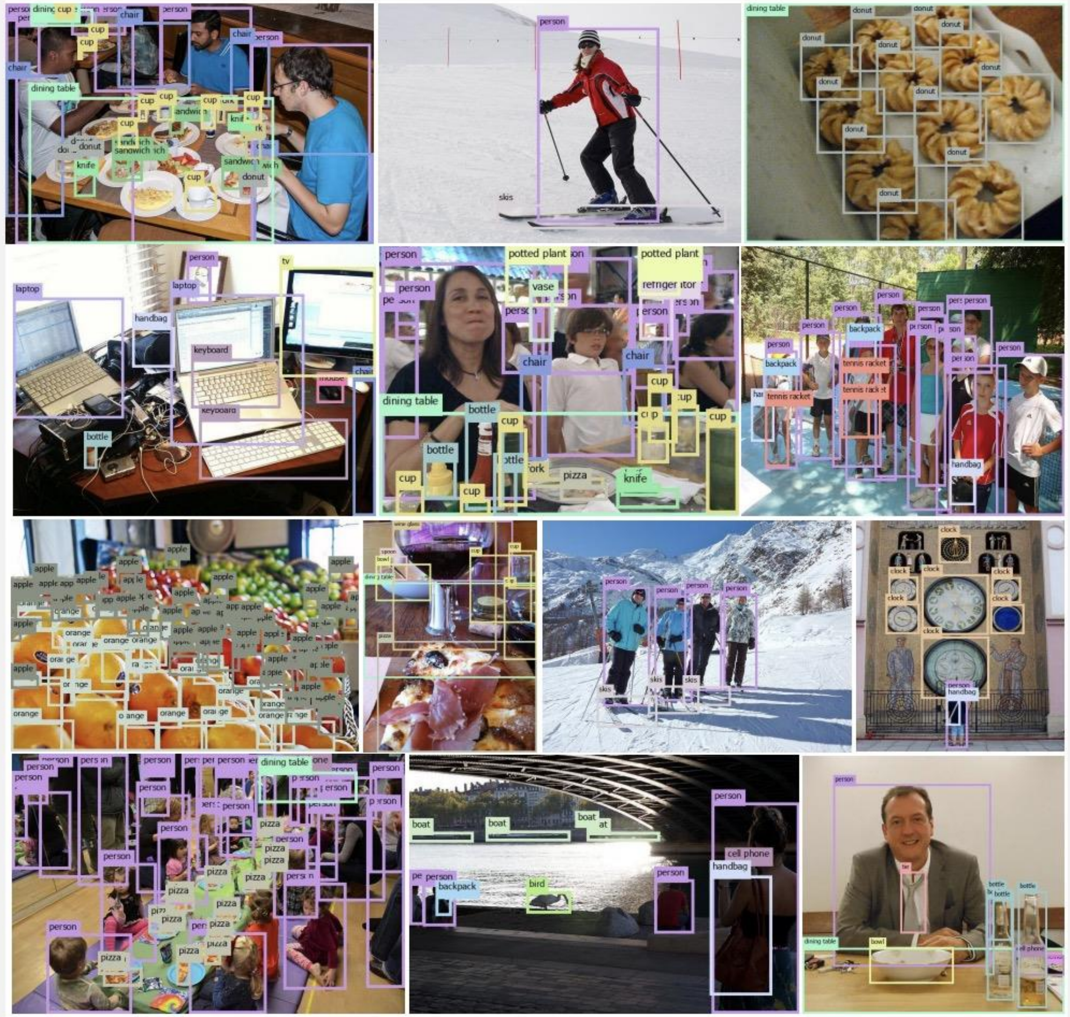


Figure 7: Visualization on Common Object Detection Task.

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:

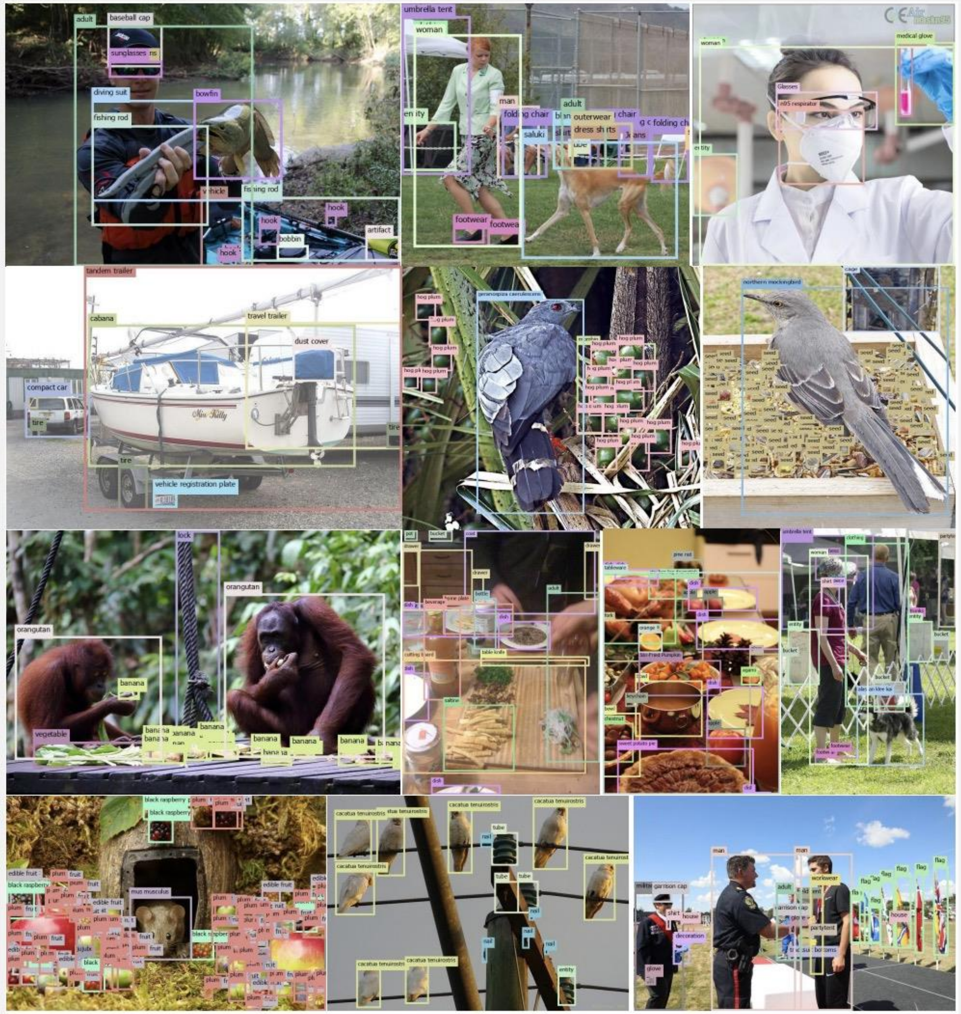
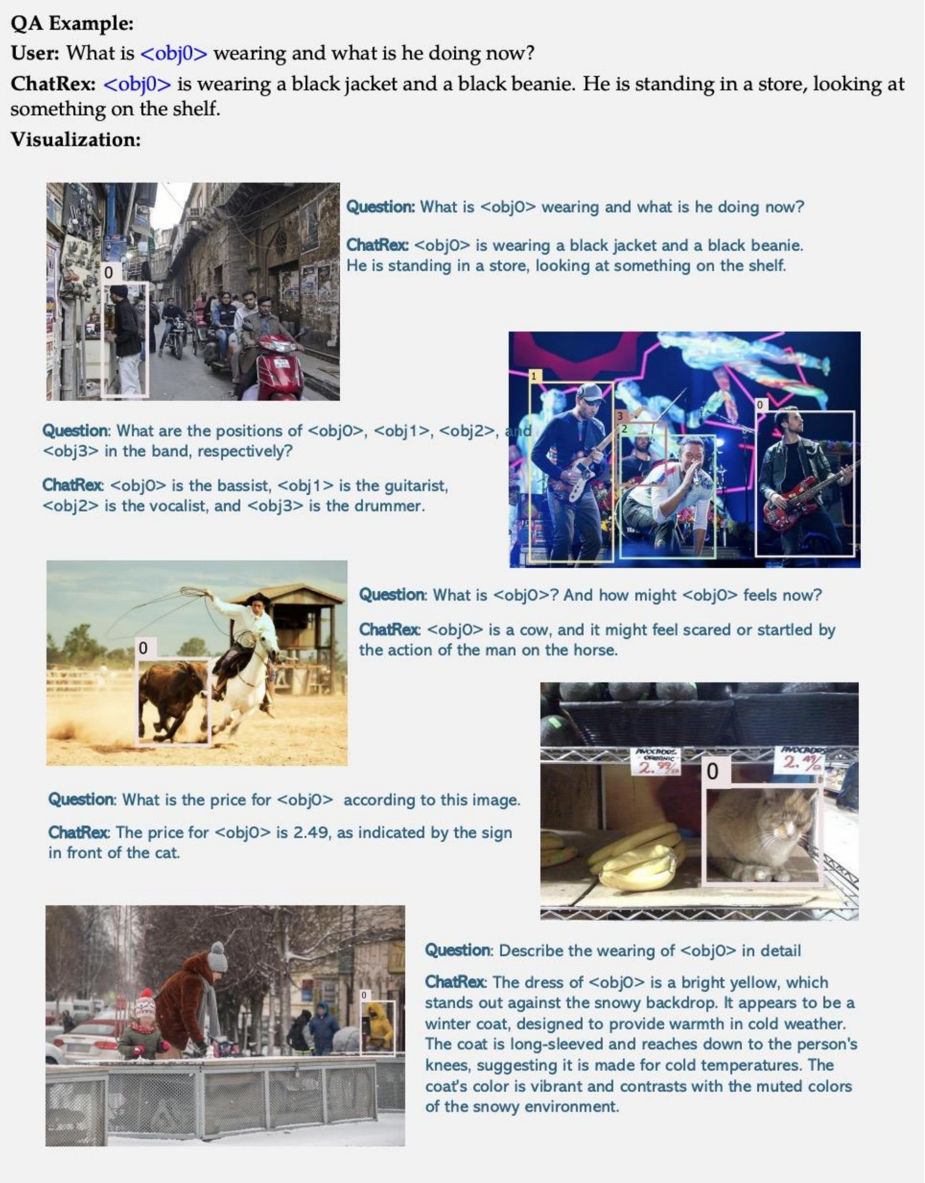


Figure 8: Visualization on Long-tailed Object Detection Task.

应用: 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> feels now?

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

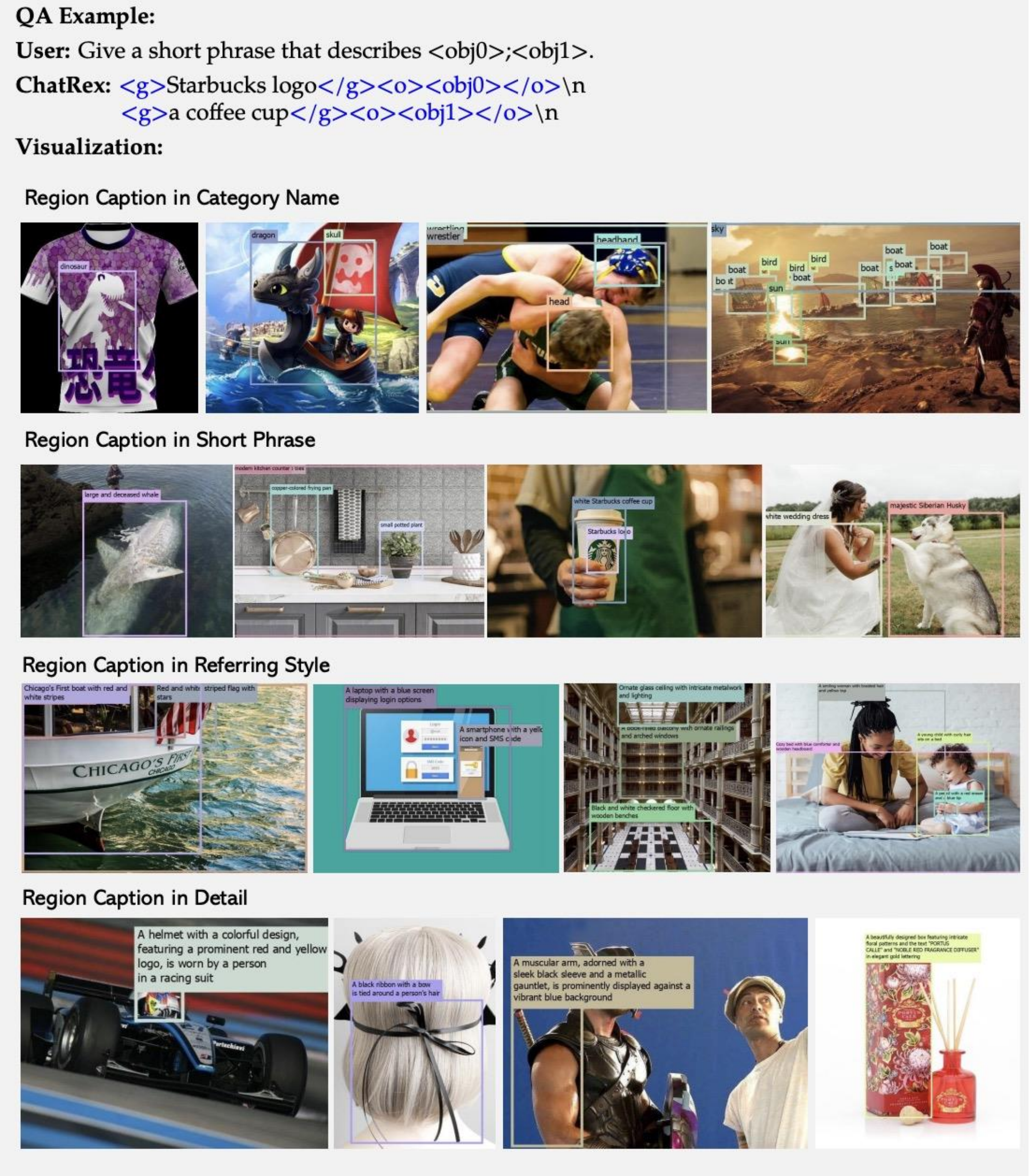
Question: What is the price for <obj0> according to this image.

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

dinosaur
dragon
skull
wrestling wrestler
headband
head
sky
boat
bird
sun

Region Caption in Short Phrase

large and deceased whale
copper-colored frying pan
small potted plant
white Starbucks coffee cup
Starbucks logo
white wedding dress
majestic Siberian Husky

Region Caption in Referring Style

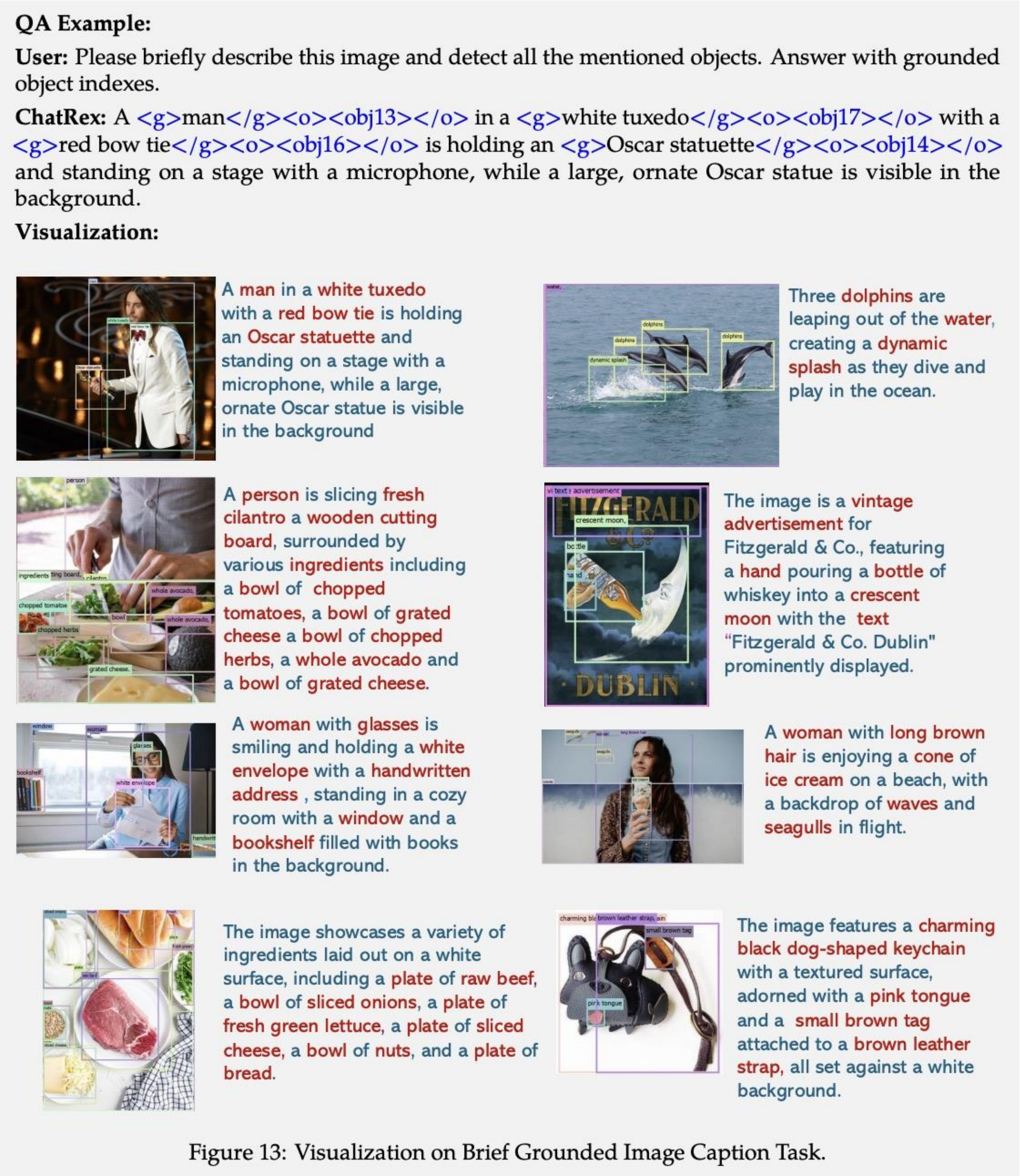
Chicago's First boat with red and white stripes
Red and white striped flag with stars
A laptop with a blue screen displaying login options
A smartphone with a yellow icon and SMS code
Ornate glass ceiling with intricate metalwork and lighting
A book-lined balcony with ornate railings and arched windows
Black and white checkered floor with wooden benches
Cozy bed with blue comforter and wooden headboard
A young child with curly hair sits on a bed

Region Caption in Detail

A helmet with a colorful design, featuring a prominent red and yellow logo, is worn by a person in a racing suit
A black ribbon with a bow is tied around a person's hair
A muscular arm, adorned with a sleek black sleeve and a metallic gauntlet, is prominently displayed against a vibrant blue background
A beautifully designed box featuring intricate floral patterns and the text "PORTUS CALLE" and "NOBLE RED FRAGRANCE DIFFUSER" in elegant gold lettering

Figure 11: Visualization on Detailed Grounded Image Caption Task.

应用: 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.

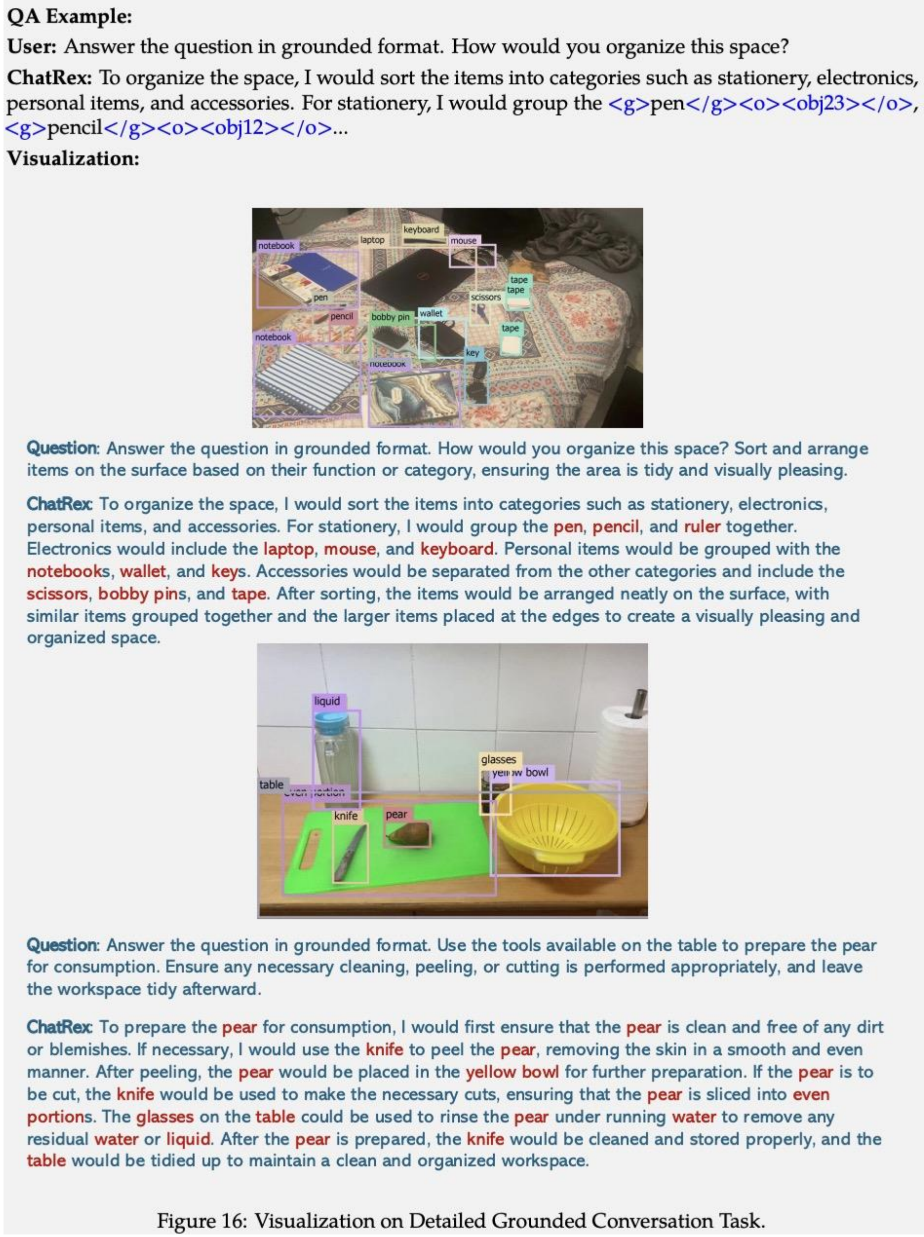
A woman with glasses is smiling and holding a white envelope with a handwritten address , standing in a cozy room with a window and a bookshelf filled with books in the background.

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.

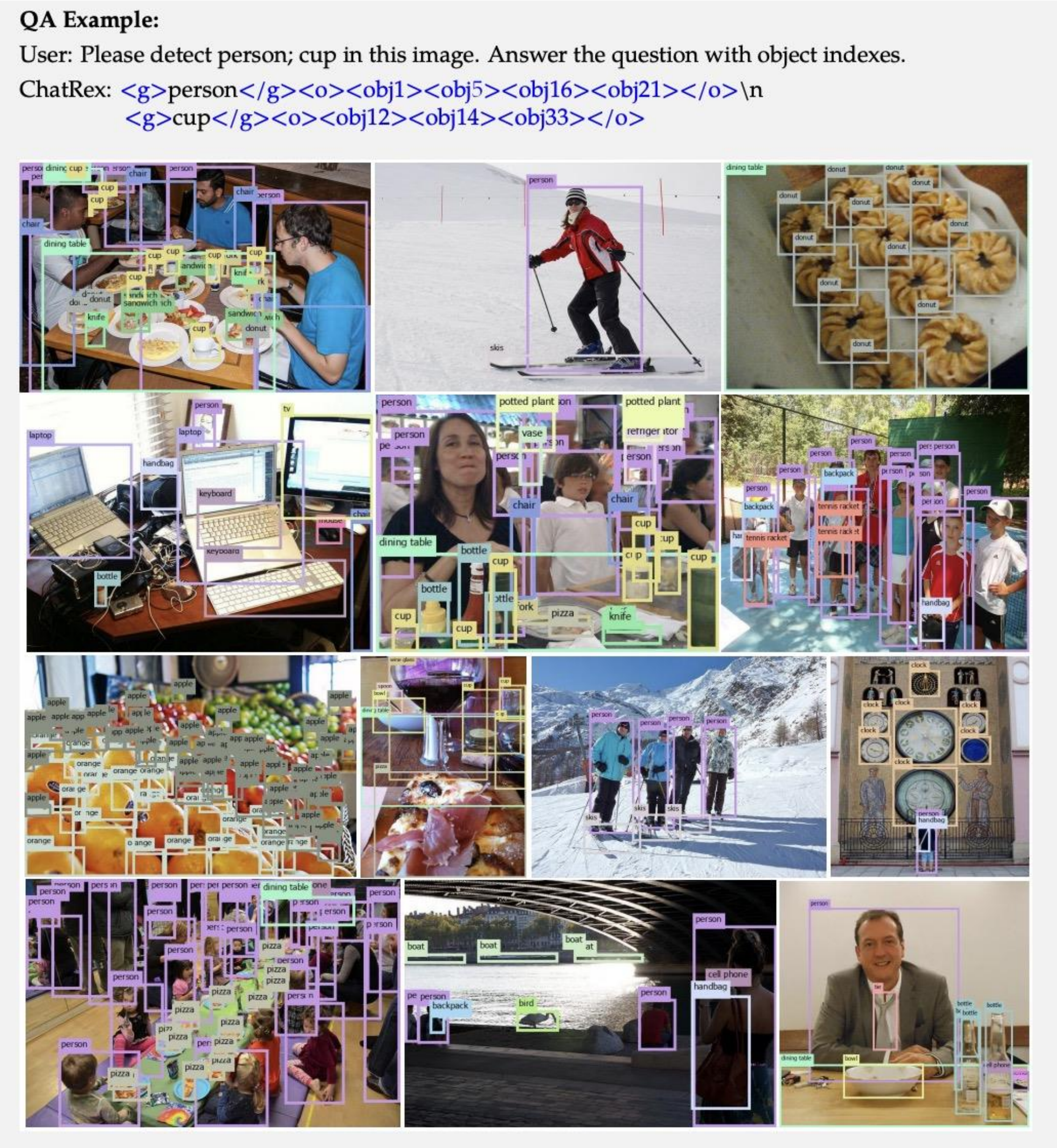
ChatRex: To organize the space, I would sort the items into categories such as stationery, electronics, personal items, and accessories. For stationery, I would group the pen, pencil, and ruler together. Electronics would include the laptop, mouse, and keyboard. Personal items would be grouped with the 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.

应用: Common/Long-tailed Object Detection

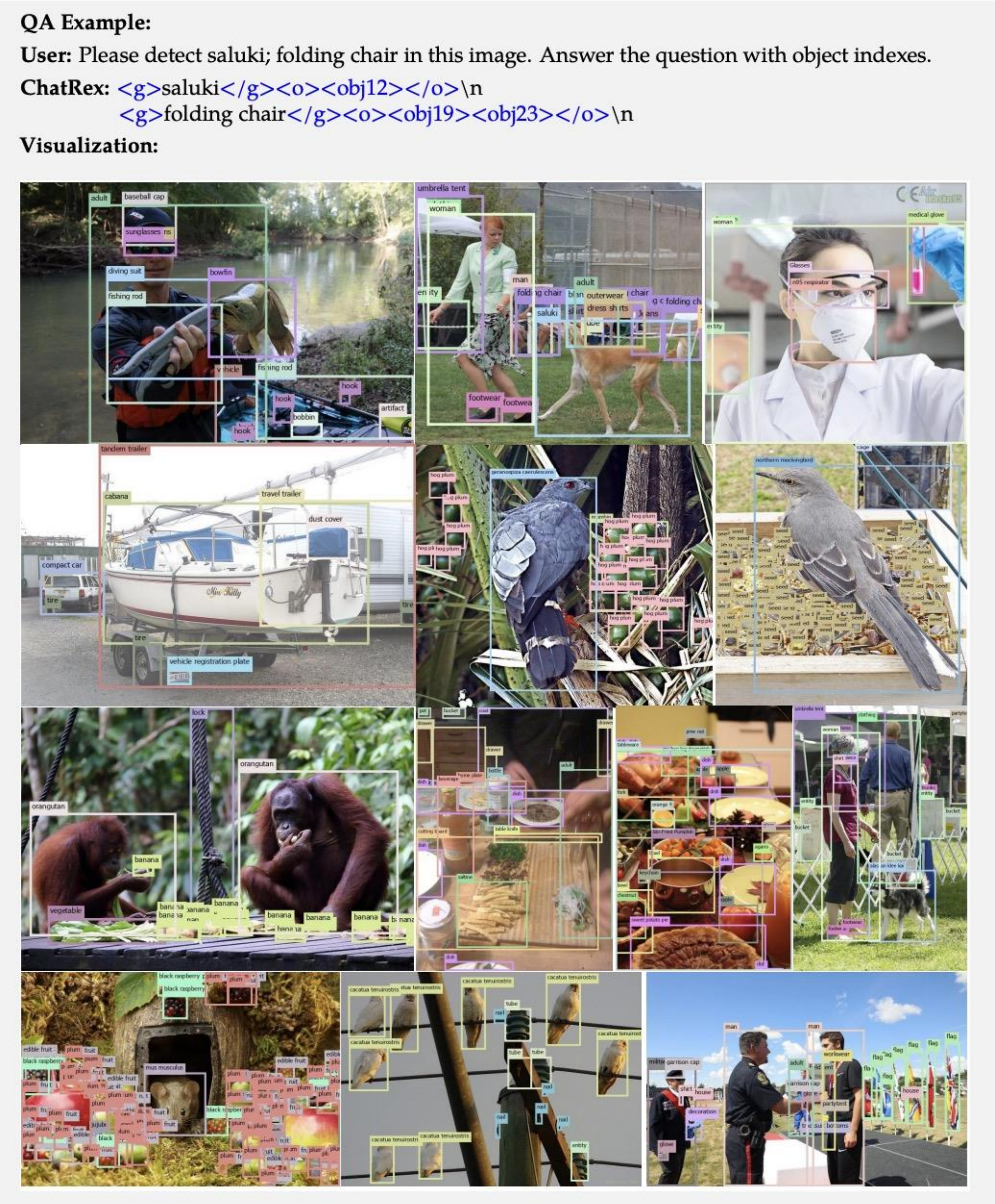


QA Example:

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:

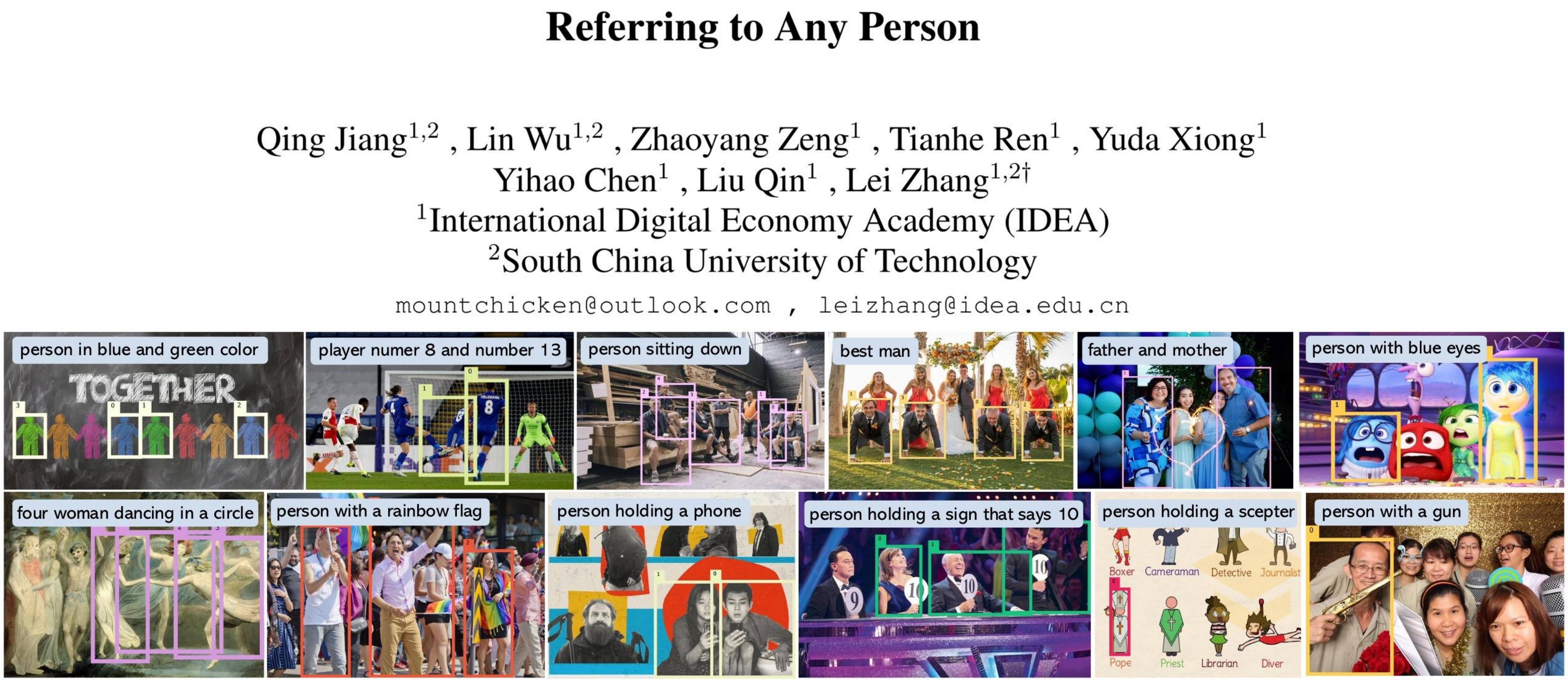
User: Please detect saluki; folding chair in this image. Answer the question with object indexes.

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.

Referring to Any Person

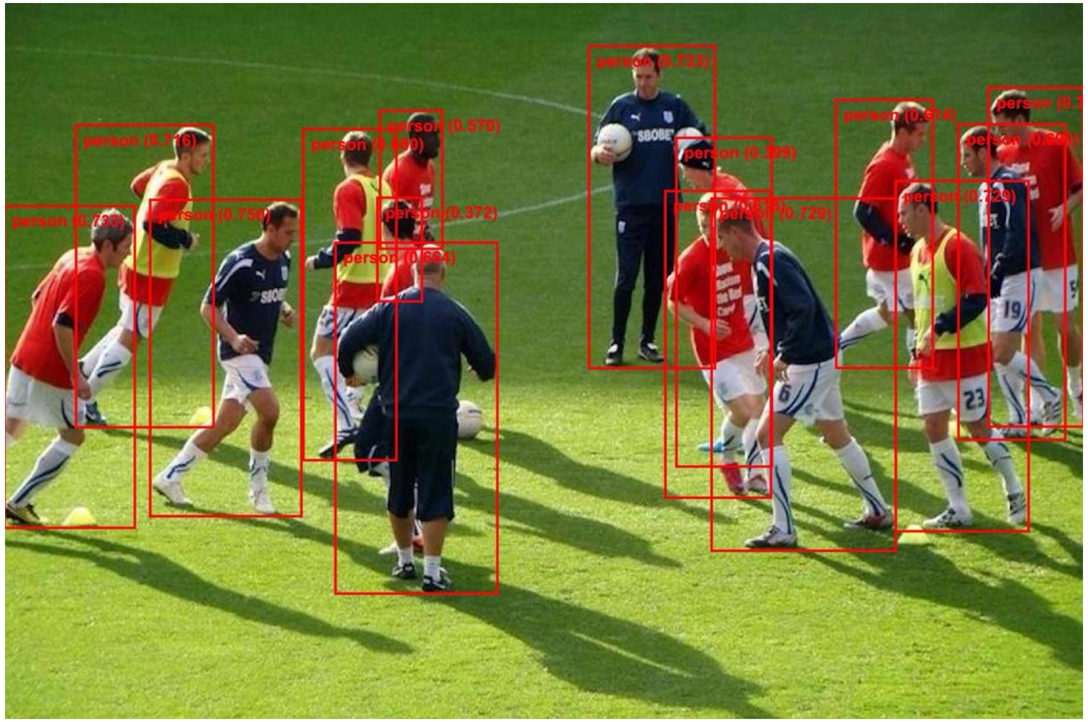


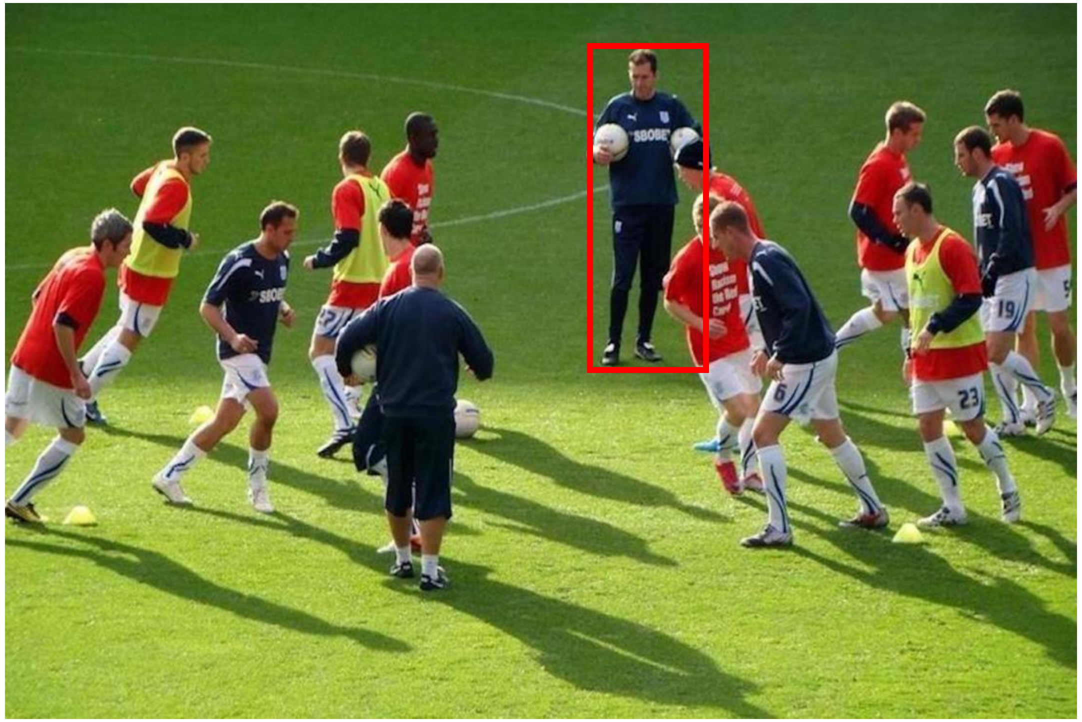
Referring to Any Person

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

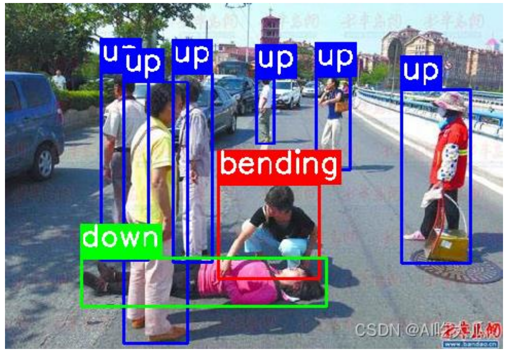




Detection: “person”

Referring: “person who is holding two footballs”

Most Detection Tasks Can be formulated as Referring



摔倒检测

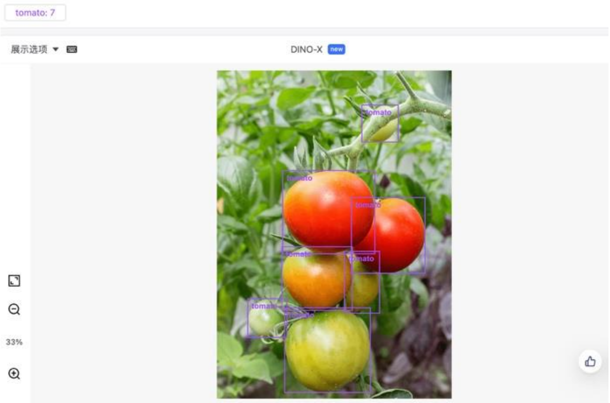
“person fallen”

佩戴安全帽检测

“person that are not wearing helmet”

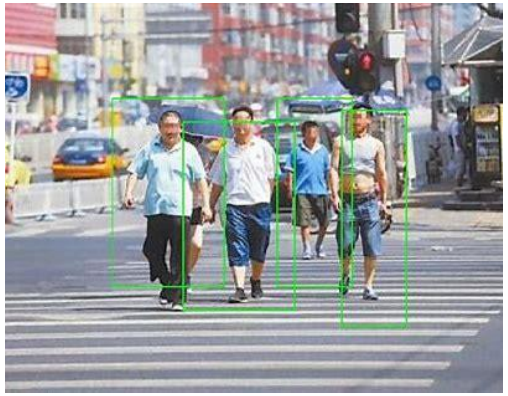
工位睡觉检测

“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

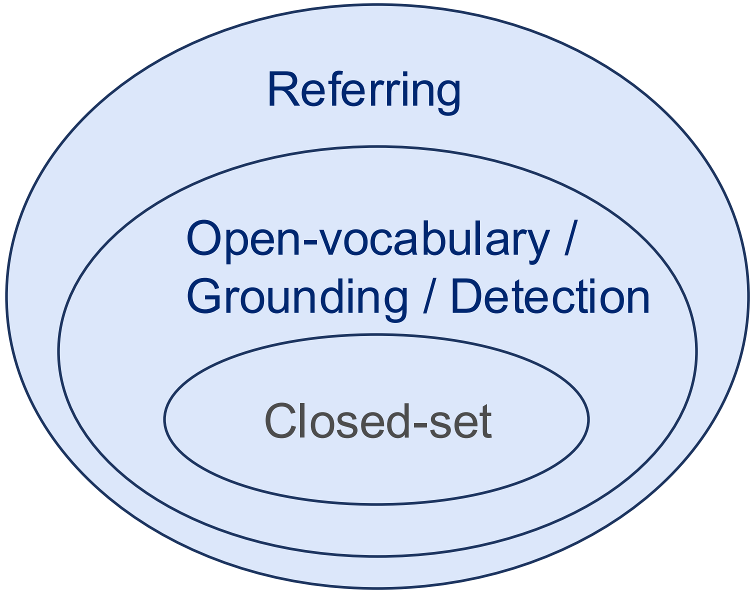
Detection: Category name e.g. man

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

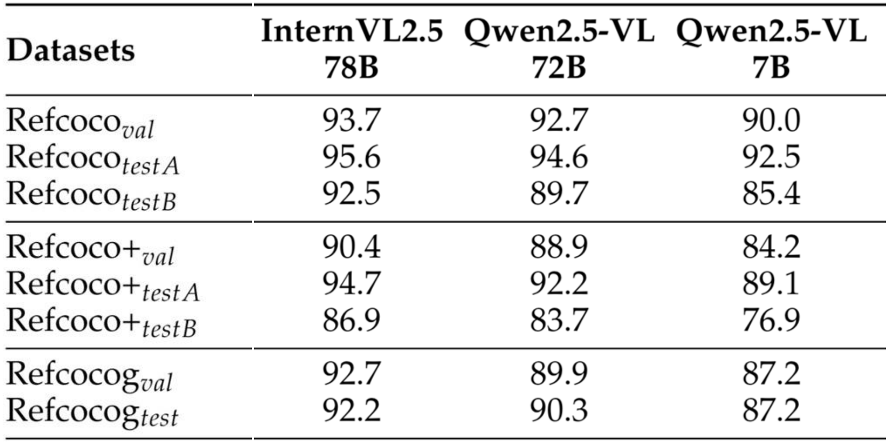


Referring

Open-vocabulary /
Grounding / Detection

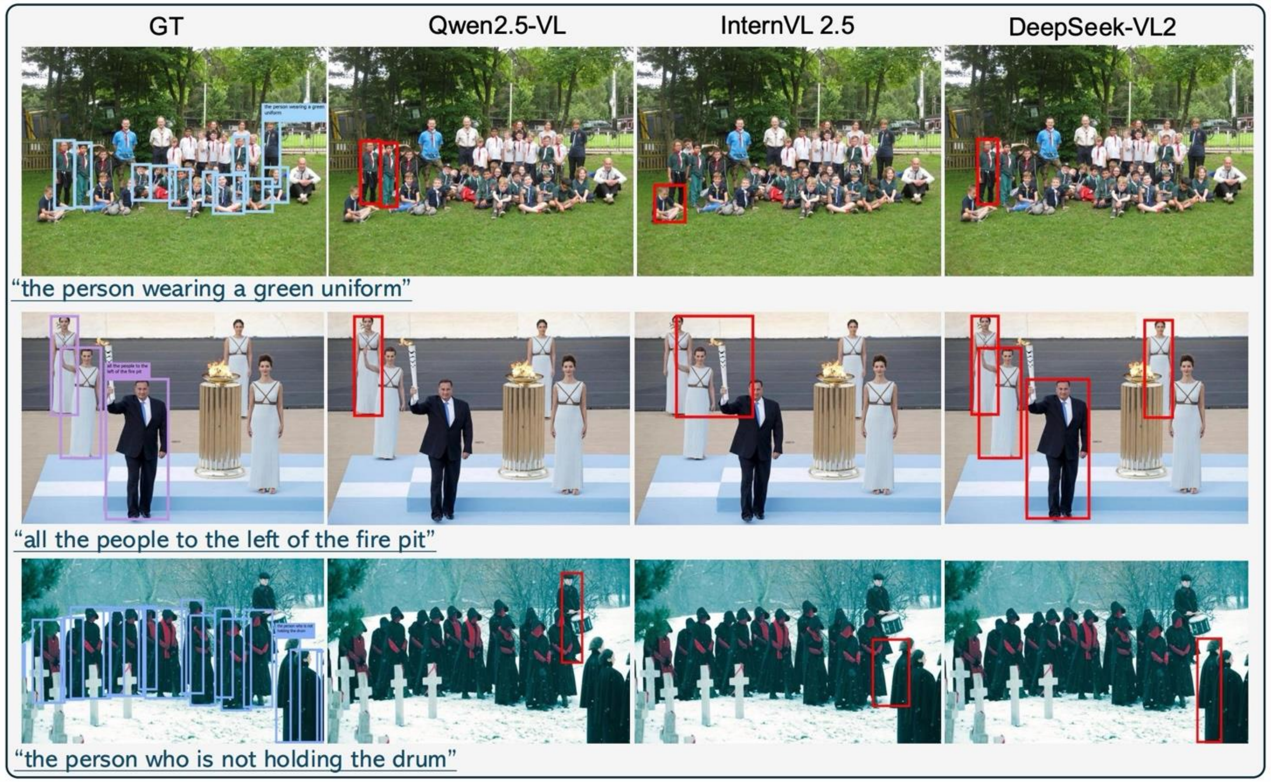
Closed-set

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**

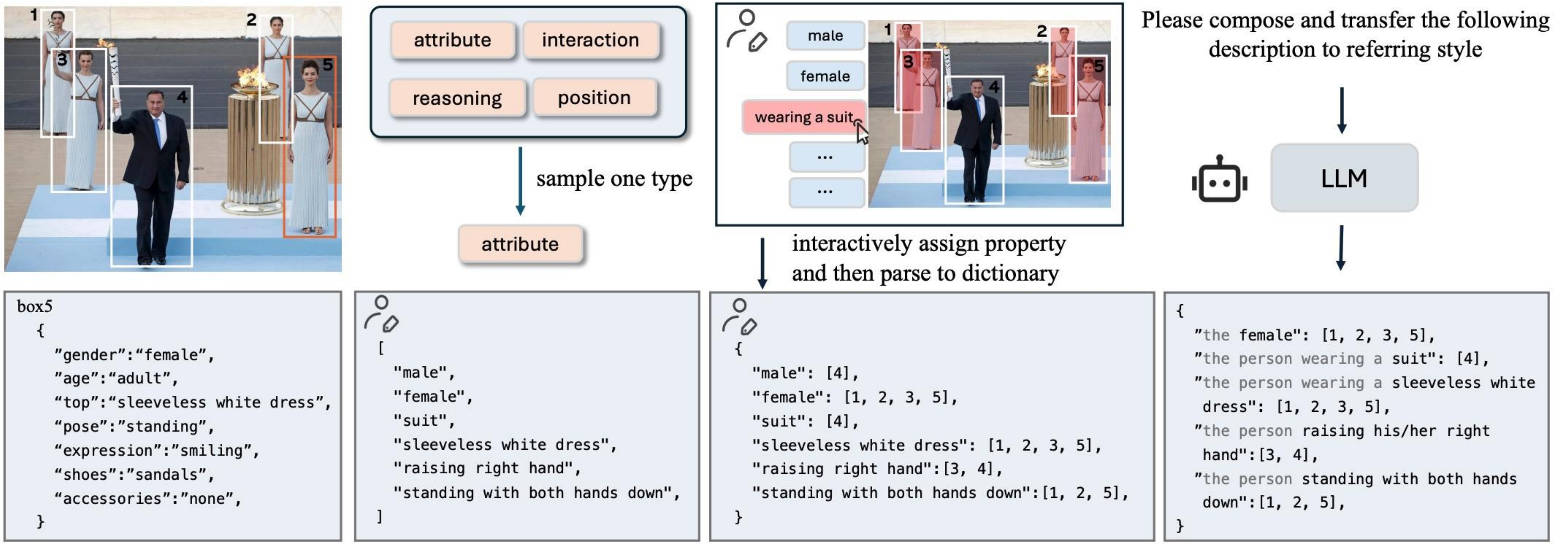

GT
Qwen2.5-VL
InternVL 2.5
DeepSeek-VL2
"the person wearing a green uniform"
"all the people to the left of the fire pit"
"the person who is not holding the drum"

Low Performance in real-world scenarios

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

Solutions: Data + Model

HumanRef Dataset


1
2
3
4
5
attribute
interaction
reasoning
position
sample one type
attribute
male
female
wearing a suit
...
...
interactively assign property
and then parse to dictionary
Please compose and transfer the following
description to referring style
LLM

```
box5
  {
    ”gender”:“female”,
    ”age”:“adult”,
    “top”:“sleeveless white dress”,
    “pose”:”standing”,
    “expression”:”smiling”,
    “shoes”:”sandals”,
    “accessories”:”none”,
  }
```

```
[
  "male",
  "female",
  "suit",
  "sleeveless white dress",
  "raising right hand",
  "standing with both hands down",
]
```

```
{
  "male": [4],
  "female": [1, 2, 3, 5],
  "suit": [4],
  "sleeveless white dress": [1, 2, 3, 5],
  "raising right hand":[3, 4],
  "standing with both hands down":[1, 2, 5],
}
```

```
{
  ”the female": [1, 2, 3, 5],
  ”the person wearing a suit": [4],
  ”the person wearing a sleeveless white
   dress": [1, 2, 3, 5],
  ”the person raising his/her right
   hand":[3, 4],
  ”the person standing with both hands
   down":[1, 2, 5],
}
```

a) pseudo labeling

b) write property list

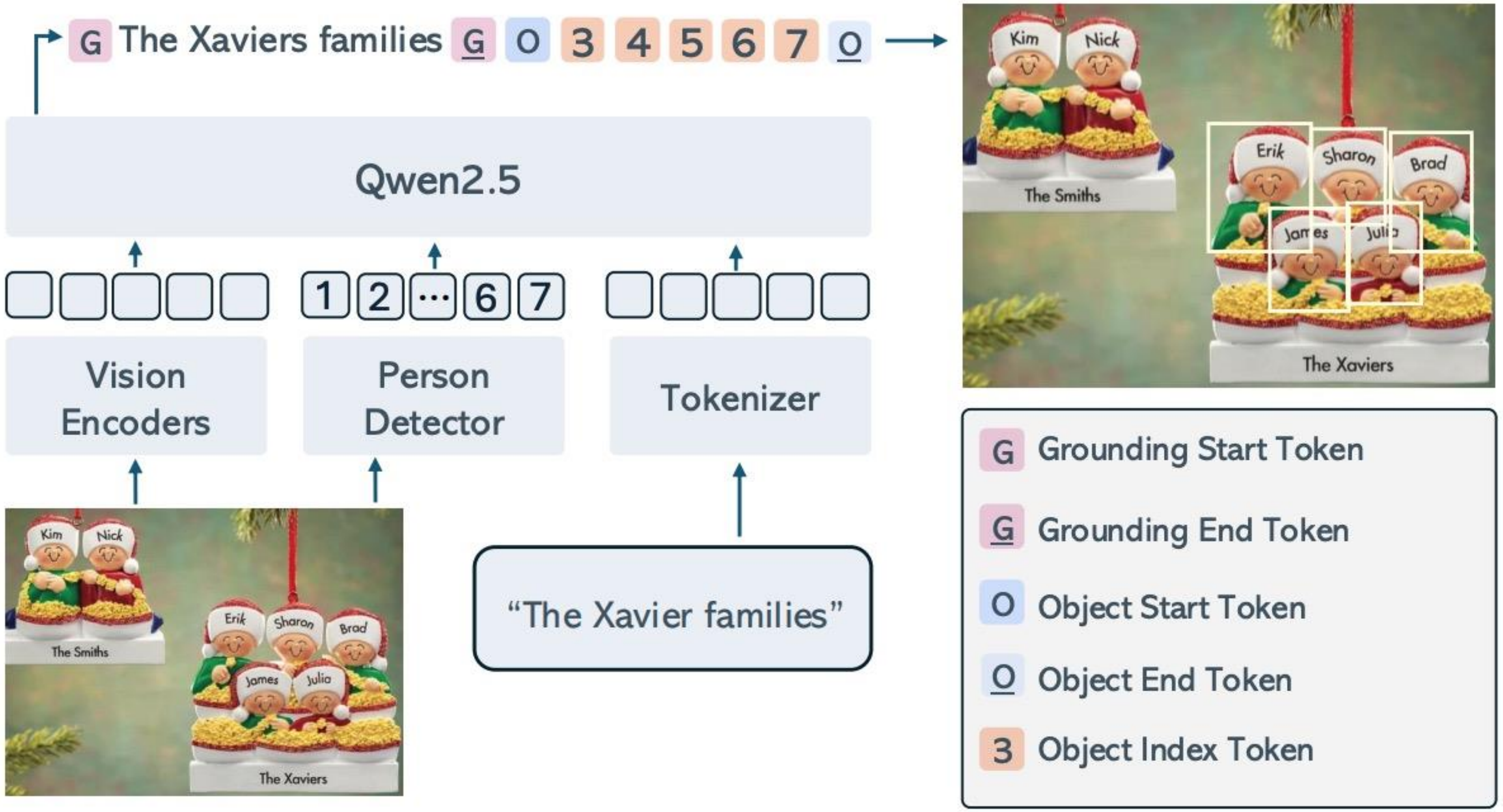
c) assign property to each person

d) transfer to referring style with LLM

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

Solutions: Model

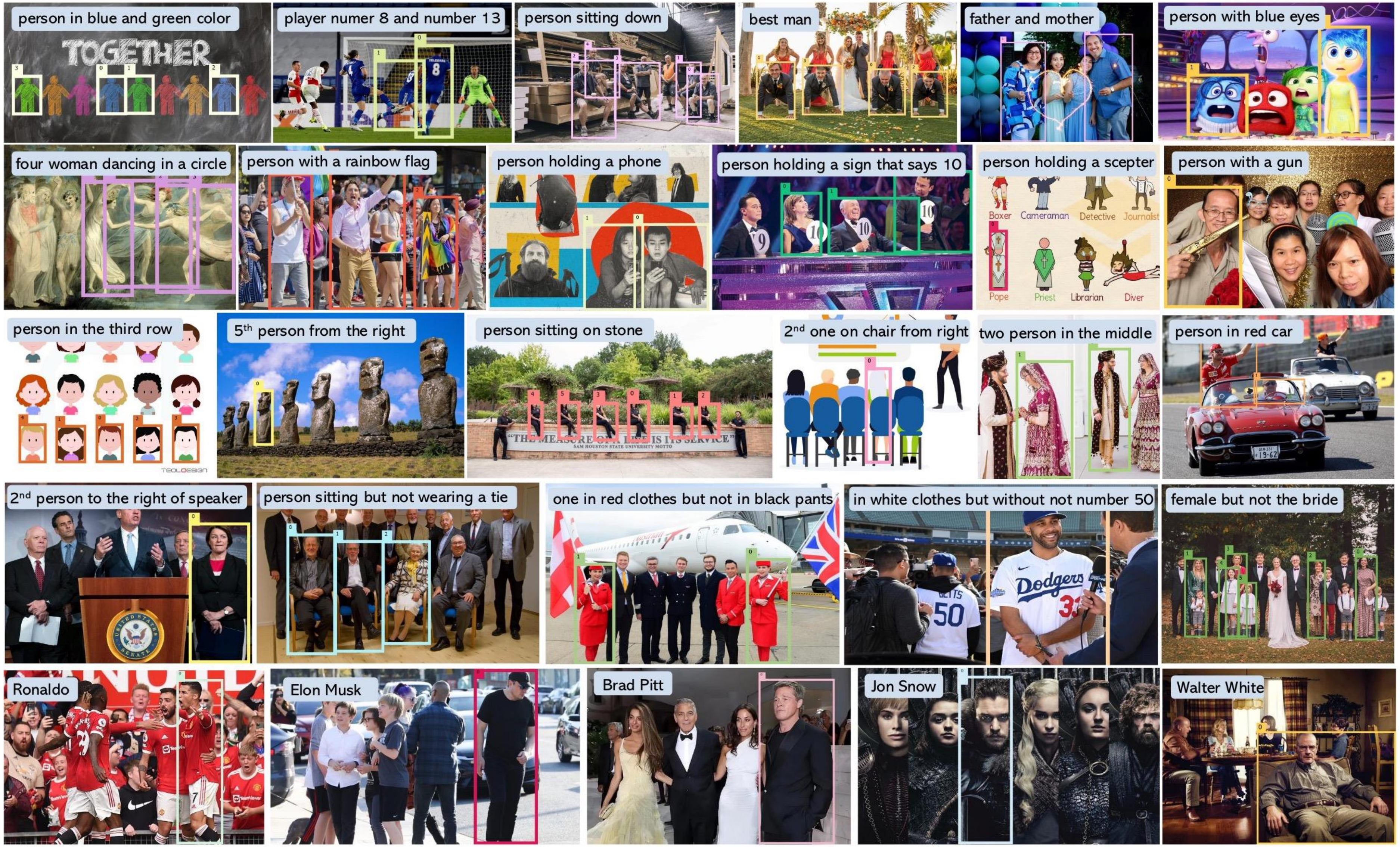
RexSeek



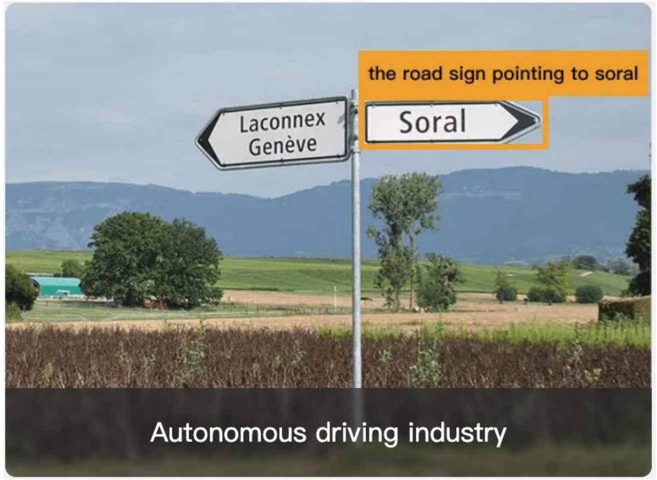
G	Grounding Start Token
G	Grounding End Token
O	Object Start Token
O	Object End Token
3	Object Index Token

1. Strong perception capability
2. Strong language comprehension

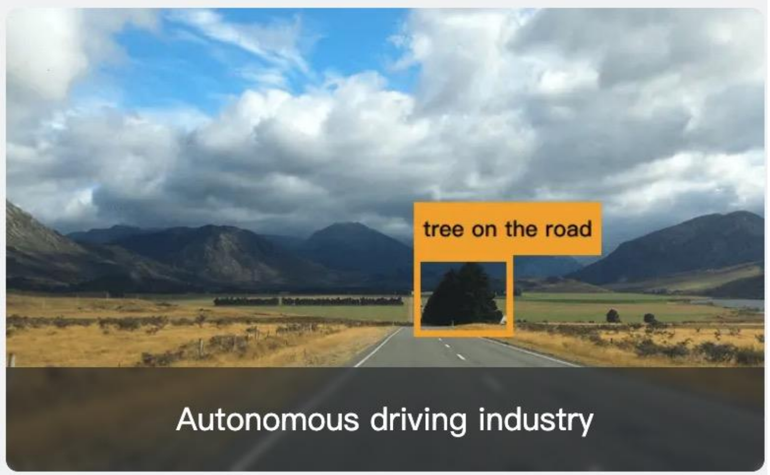
Applications: Referring Any Person


person in blue and green color
TOGETHER
player numer 8 and number 13
person sitting down
best man
father and mother
person with blue eyes
four woman dancing in a circle
person with a rainbow flag
person holding a phone
person holding a sign that says 10
person holding a scepter
Boxer
Cameraman
Detective
Journalist
Pope
Priest
Librarian
Diver
person with a gun
person in the third row
5th person from the right
person sitting on stone
2nd one on chair from right
two person in the middle
person in red car
2nd person to the right of speaker
person sitting but not wearing a tie
one in red clothes but not in black pants
in white clothes but without not number 50
female but not the bride
Ronaldo
Elon Musk
Brad Pitt
Jon Snow
Walter White

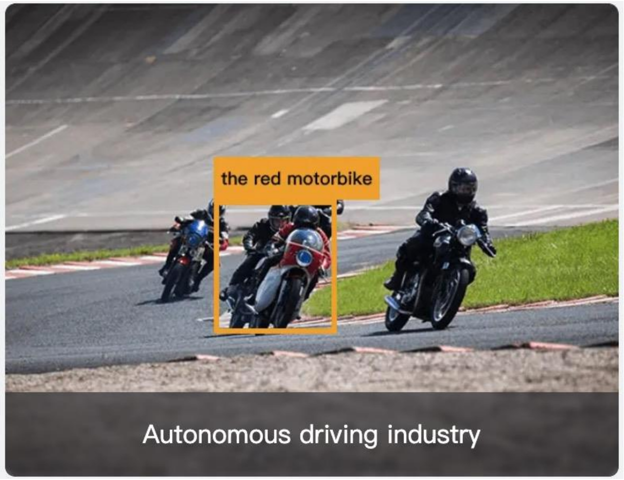
Applications: Any Object Referring



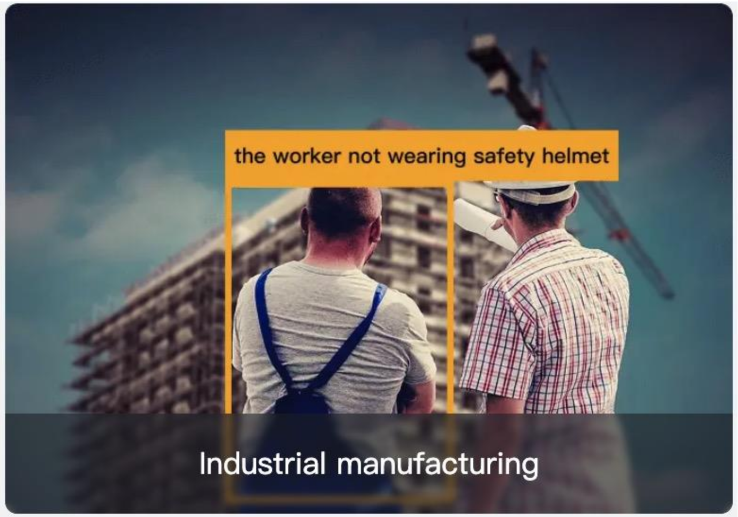
Autonomous driving industry



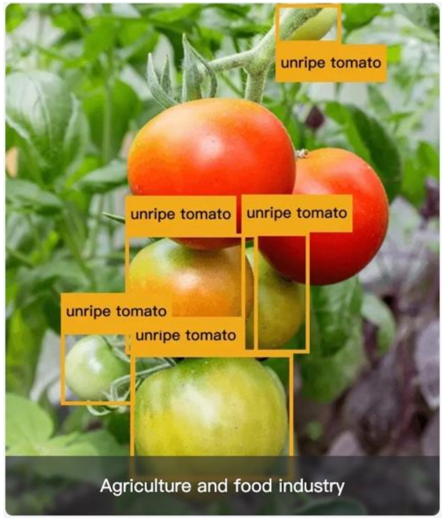
Autonomous driving industry



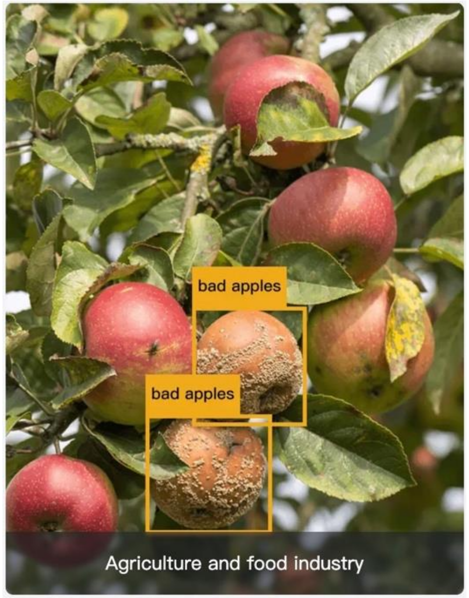
Autonomous driving industry



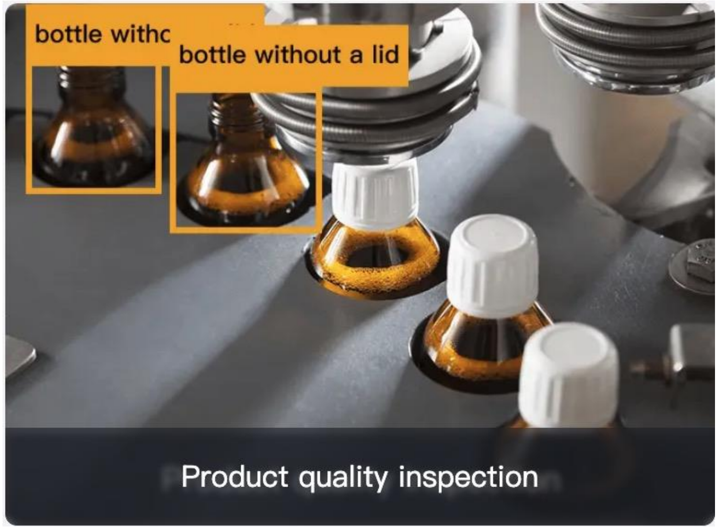
Industrial manufacturing



Agriculture and food industry

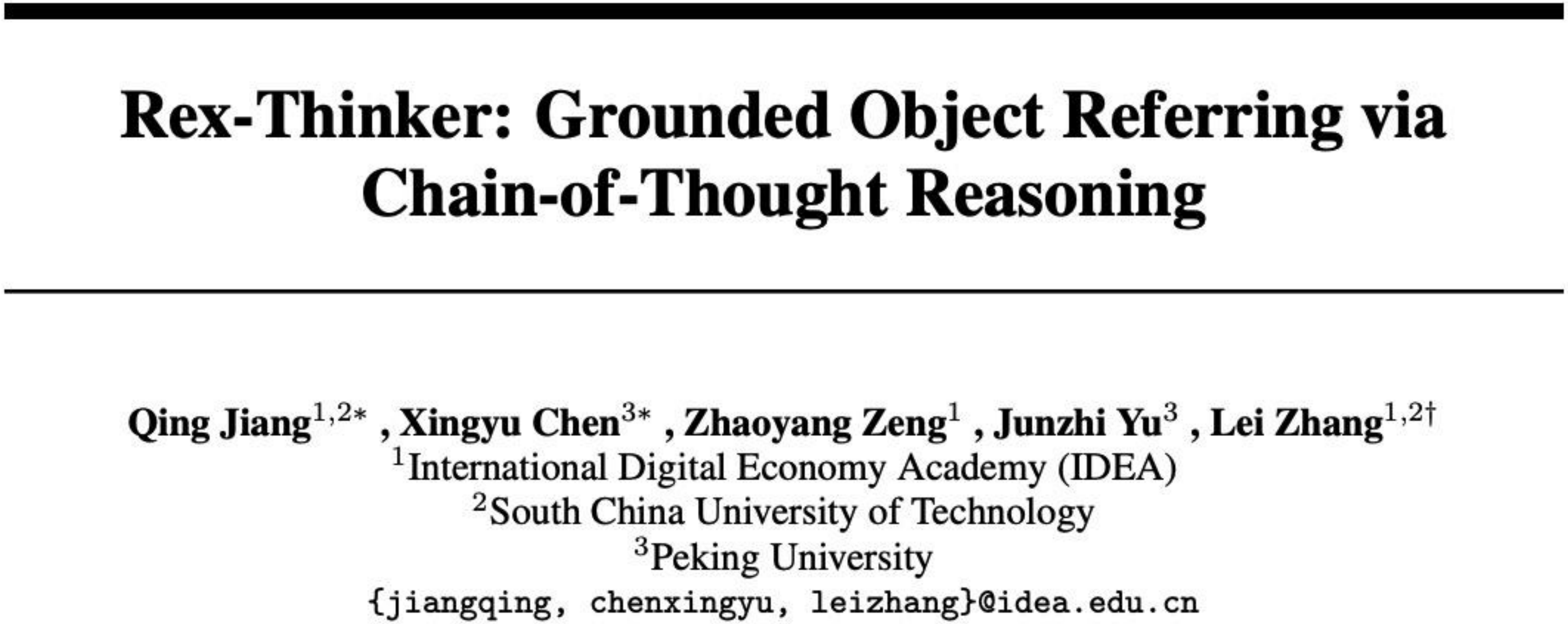


Agriculture and food industry



Product quality inspection

Rex-Thinker: Referring with CoT Reasoning



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

Qing Jiang[1,2*] , **Xingyu Chen**[3*] , **Zhaoyang Zeng**[1] , **Junzhi Yu**[3] , **Lei Zhang**[1,2†]
[1] International Digital Economy Academy (IDEA)
[2] South China University of Technology
[3] Peking University
{jiangqing, chenxingyu, leizhang}@idea.edu.cn

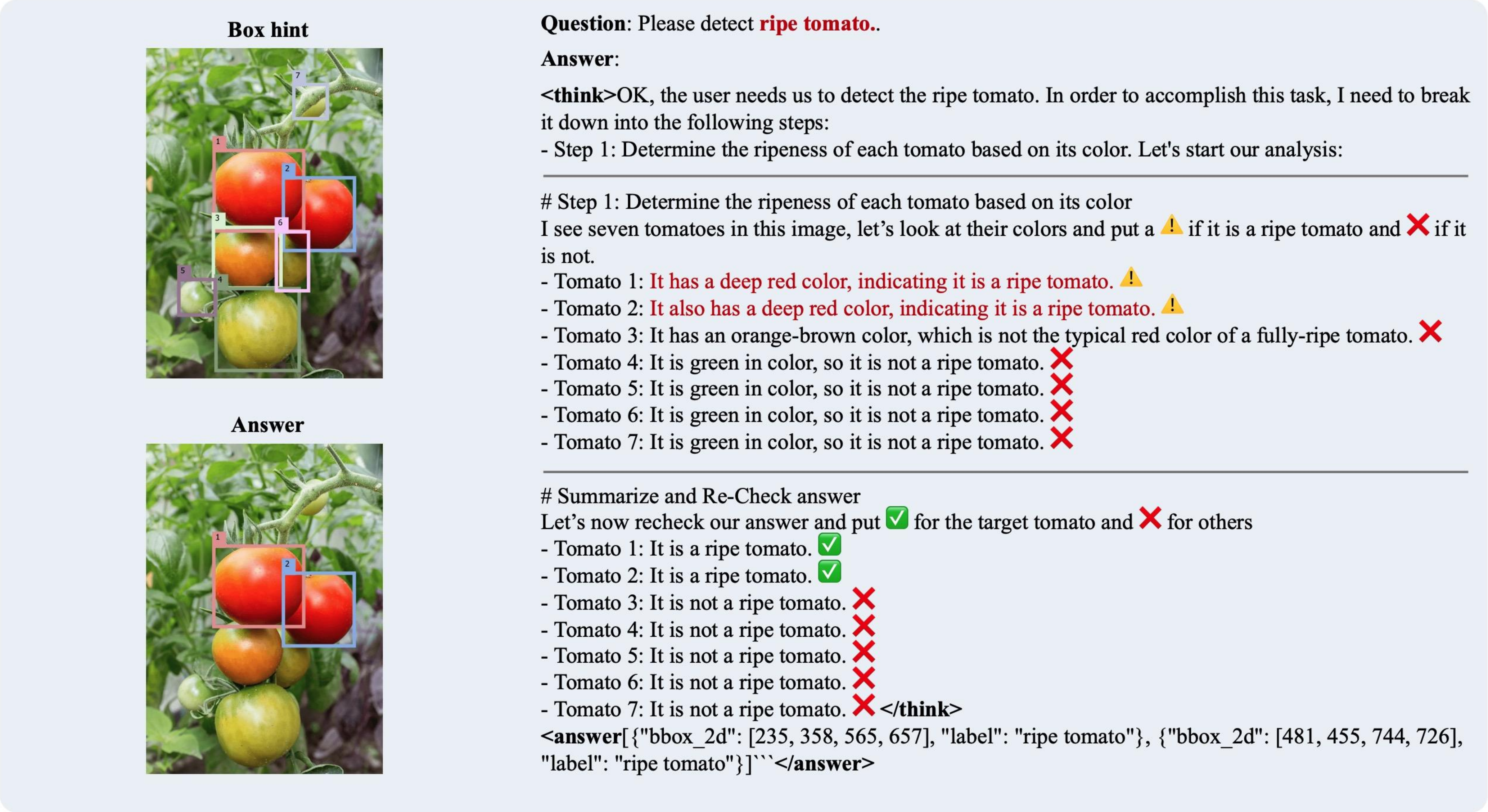
Rex-Thinker: Referring with CoT Reasoning

Live Reasoning Demo

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

Human:

Rex-Thinker: Referring with CoT Reasoning



Box hint

Answer

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

Answer:

Planning

<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:

Action

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. ✗

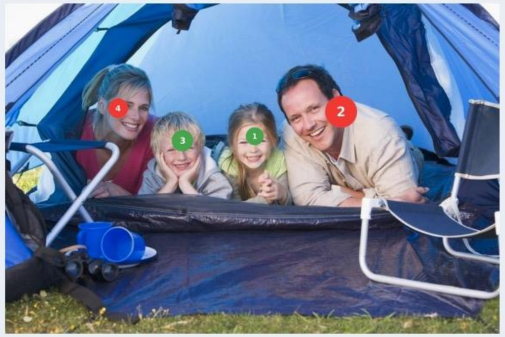
Summarization

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>**

1. CoT Referring Dataset

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.

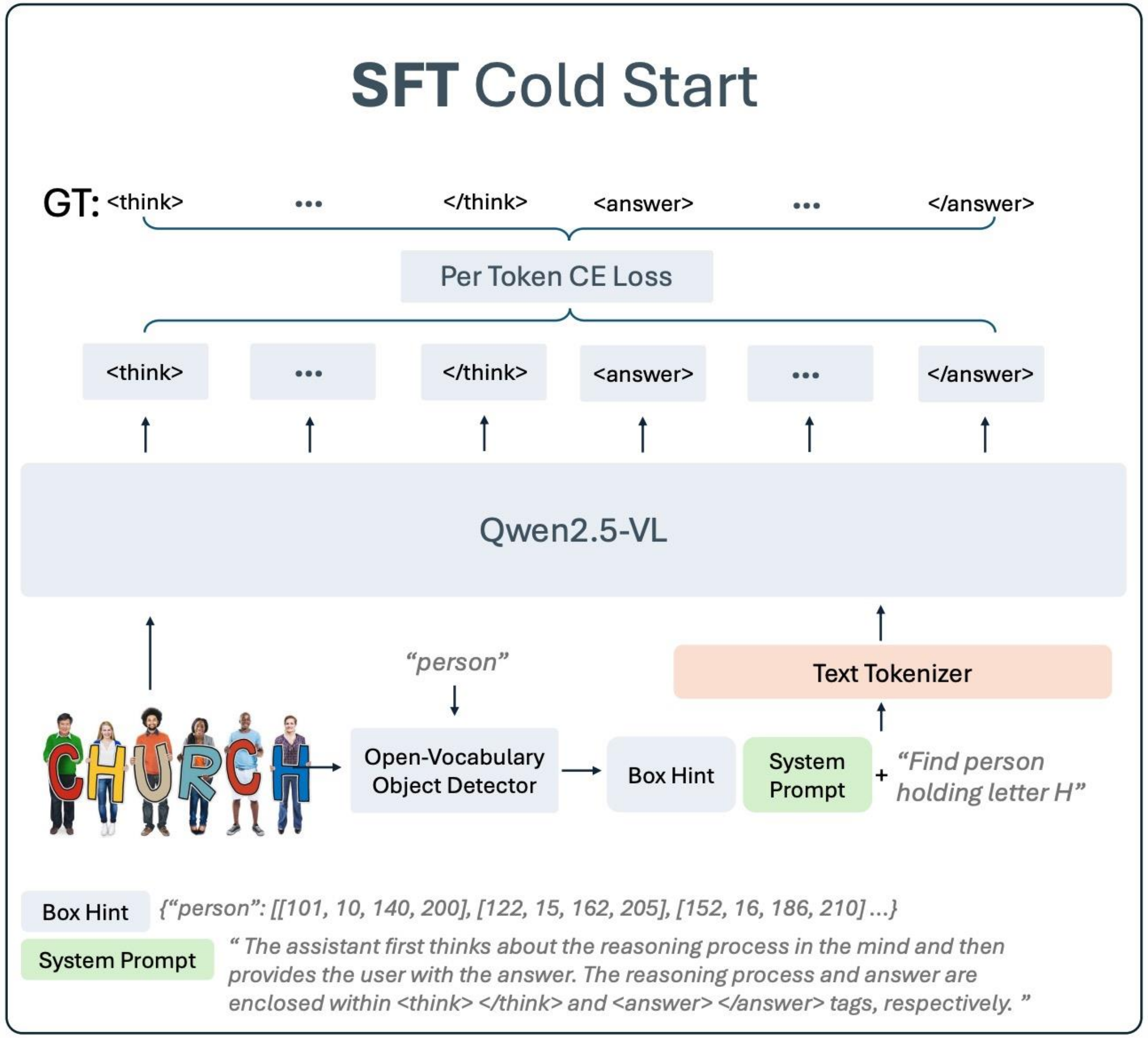
3. Summarization:

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. → ✗

2. SFT + GRPO Training



SFT Cold Start

GT: <think> ••• </think> <answer> ••• </answer>

Per Token CE Loss

<think> ••• </think> <answer> ••• </answer>

Qwen2.5-VL

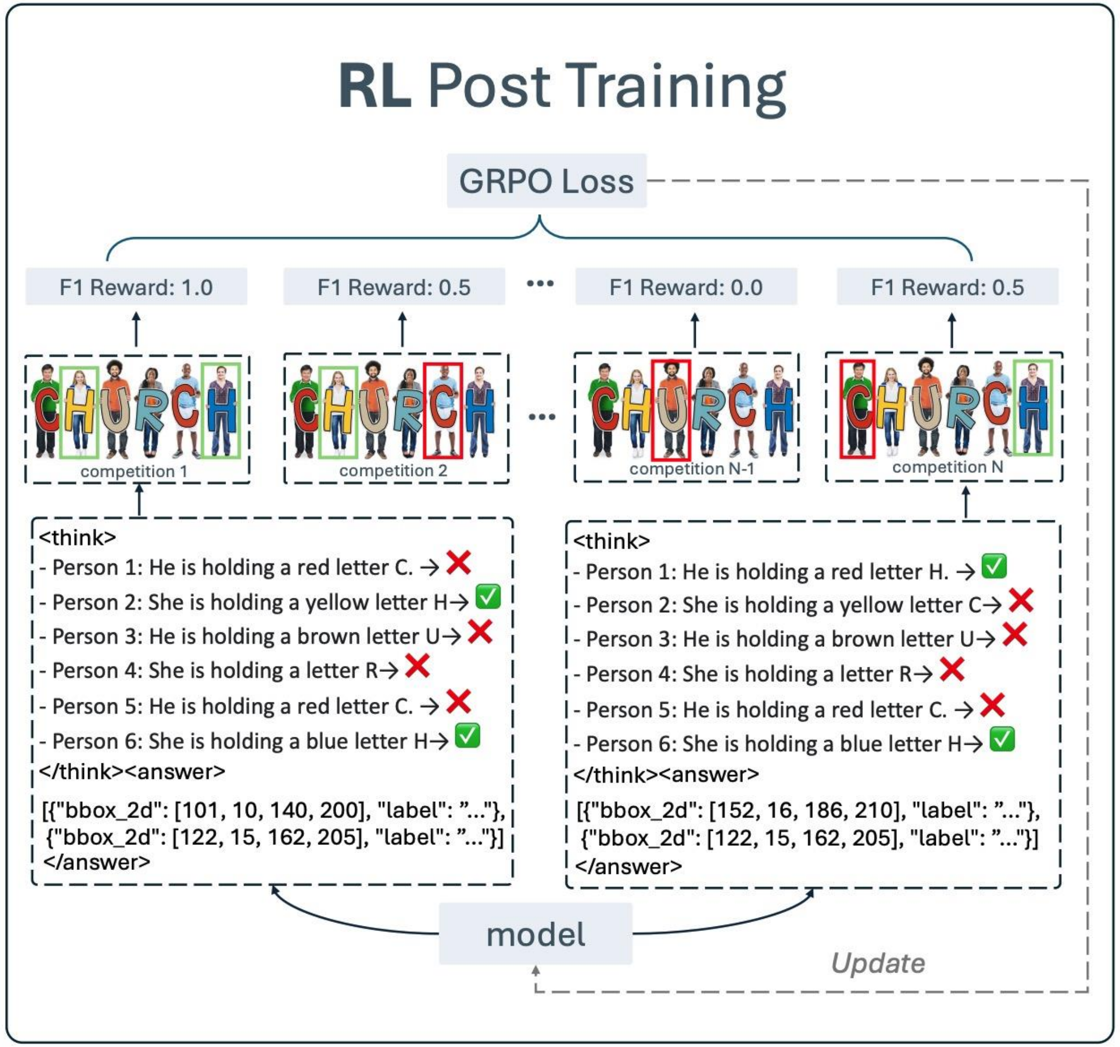
"person"

Open-Vocabulary Object Detector → Box Hint System Prompt + *"Find person holding letter H"*

Text Tokenizer

Box Hint *{"person": [[101, 10, 140, 200], [122, 15, 162, 205], [152, 16, 186, 210] ...}*

System Prompt *" The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively. "*



RL Post Training

GRPO Loss

F1 Reward: 1.0 F1 Reward: 0.5 ••• F1 Reward: 0.0 F1 Reward: 0.5

competition 1 competition 2 ••• competition N-1 competition N

<think>
- Person 1: He is holding a red letter C. → ❌
- Person 2: She is holding a yellow letter H→ ✅
- Person 3: He is holding a brown letter U→ ❌
- Person 4: She is holding a letter R→ ❌
- Person 5: He is holding a red letter C. → ❌
- Person 6: She is holding a blue letter H→ ✅
</think><answer>
[{"bbox_2d": [101, 10, 140, 200], "label": "..."},
{"bbox_2d": [122, 15, 162, 205], "label": "..."}]
</answer>

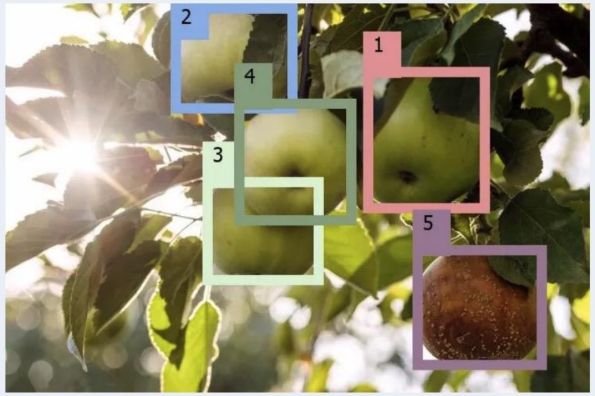
<think>
- Person 1: He is holding a red letter H. → ✅
- Person 2: She is holding a yellow letter C→ ❌
- Person 3: He is holding a brown letter U→ ❌
- Person 4: She is holding a letter R→ ❌
- Person 5: He is holding a red letter C. → ❌
- Person 6: She is holding a blue letter H→ ✅
</think><answer>
[{"bbox_2d": [152, 16, 186, 210], "label": "..."},
{"bbox_2d": [122, 15, 162, 205], "label": "..."}]
</answer>

model

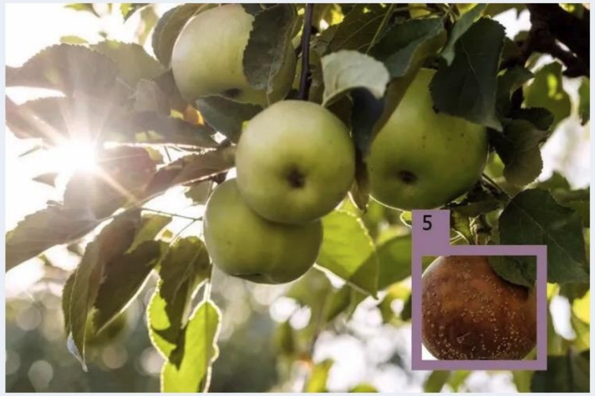
Update

Visualization Results

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. ⚠

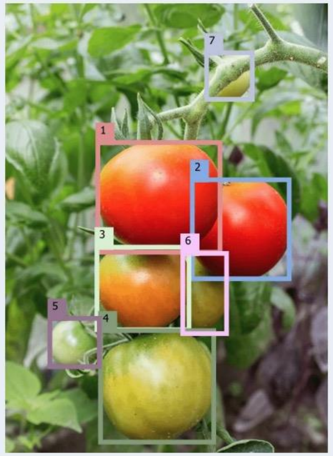
Step 2: Identify the type of the disease
The brownish spot on Apple 5 suggests that it may have a fungal or bacterial infection, which are common types of diseases for apples.

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**>

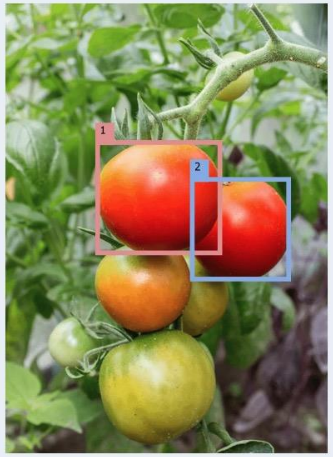
<**answer**[{"bbox_2d": [429, 261, 570, 395], "label": "apple with disease"}]```</**answer**>

Visualization Results

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:

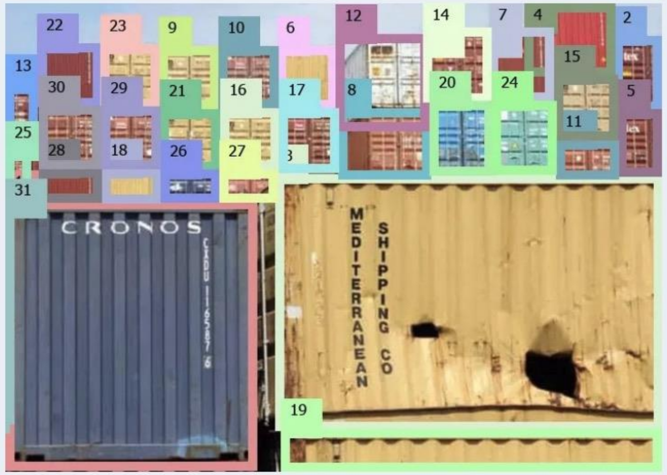
# Step 1: Determine the ripeness of each tomato based on its color
I see seven tomatoes in this image, let's look at their colors and put a ⚠ if it is a ripe tomato and ❌ if it is not.
- Tomato 1: It has a deep red color, indicating it is a ripe tomato. ⚠
- Tomato 2: It also has a deep red color, indicating 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>**

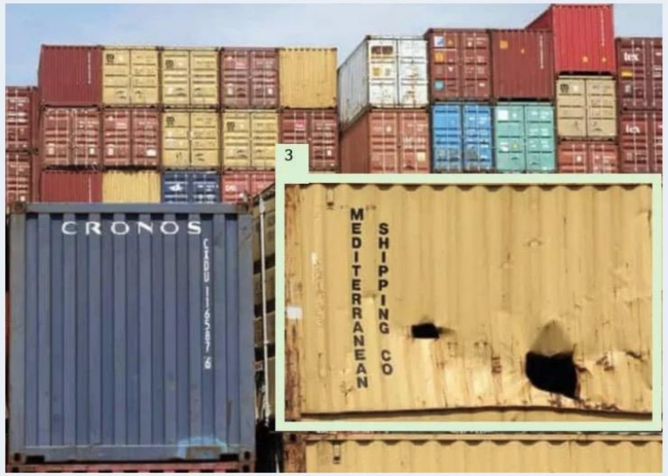
<answer[{"bbox_2d": [235, 358, 565, 657], "label": "ripe tomato"}, {"bbox_2d": [481, 455, 744, 726], "label": "ripe tomato"}]```**</answer>**

Visualization Results

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. ✗

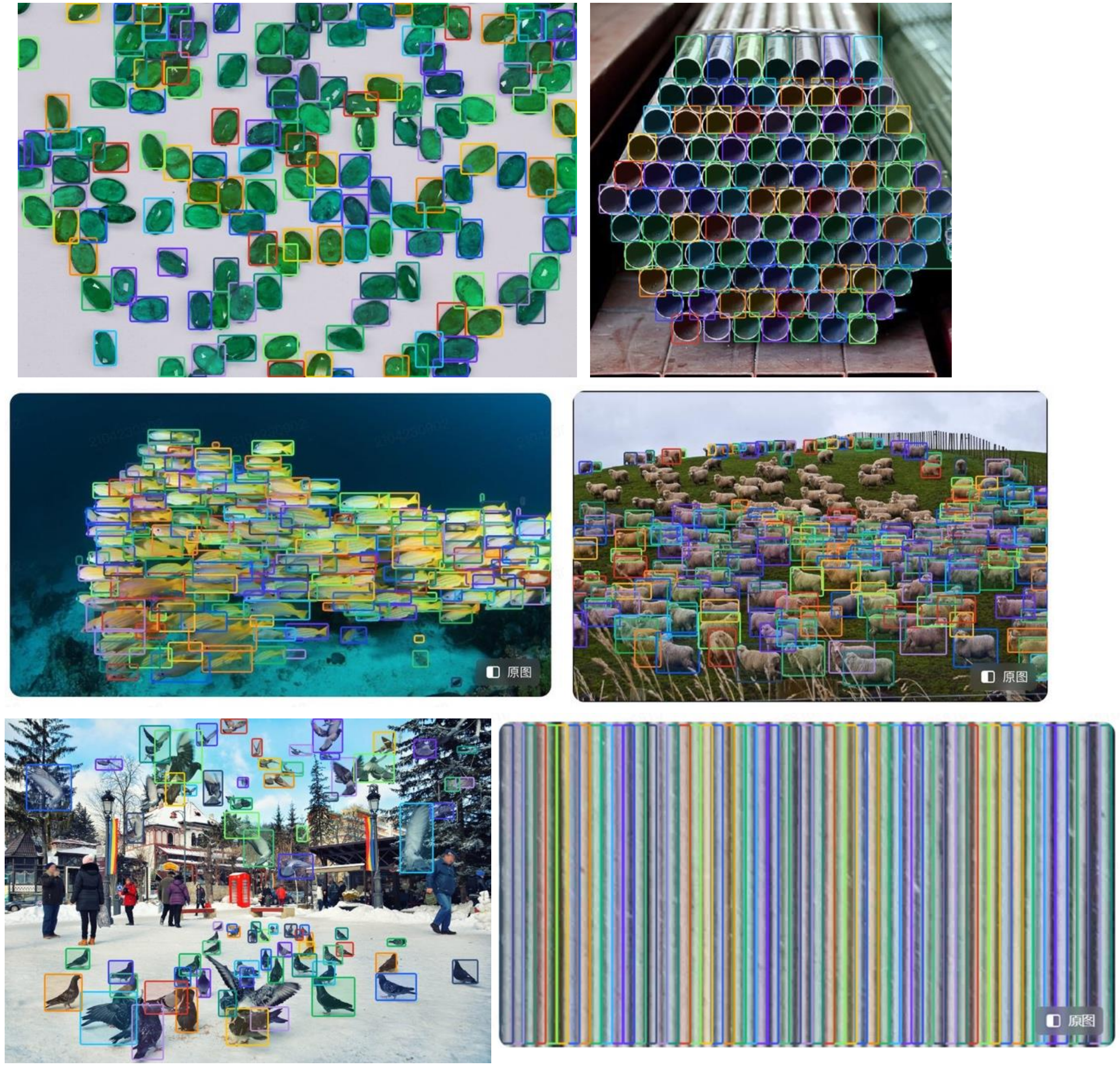
Summarize and Re-Check answer
Let's now recheck our answer and put ✅ for the target container and ✗ for others
- Container 1: No visible signs of damage → ✗
- Container 2: No visible signs of damage → ✗
- Container 3: Has a large hole, is damaged → ✅
- Container 4: No visible signs of damage → ✗

…
- 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"}]```**</answer>**

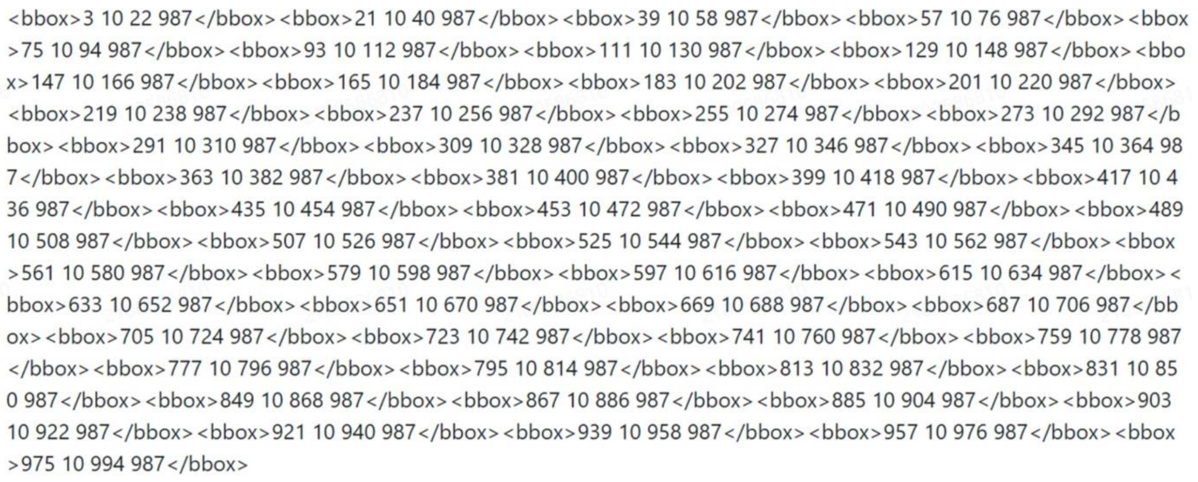
SEED-VL-1.5

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

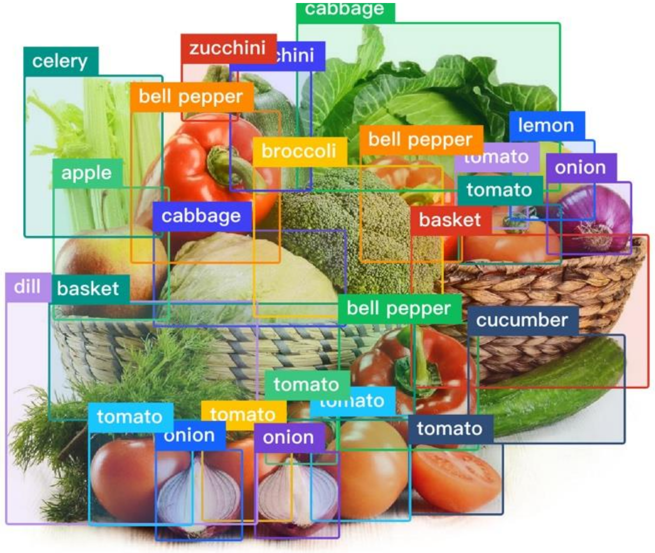
Abswer:

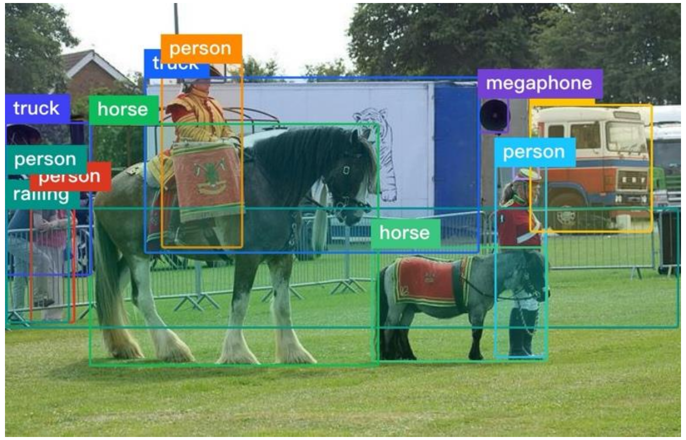


```
<bbox>3 10 22 987</bbox><bbox>21 10 40 987</bbox><bbox>39 10 58 987</bbox><bbox>57 10 76 987</bbox><bbox
>75 10 94 987</bbox><bbox>93 10 112 987</bbox><bbox>111 10 130 987</bbox><bbox>129 10 148 987</bbox><bbo
x>147 10 166 987</bbox><bbox>165 10 184 987</bbox><bbox>183 10 202 987</bbox><bbox>201 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</b
box><bbox>291 10 310 987</bbox><bbox>309 10 328 987</bbox><bbox>327 10 346 987</bbox><bbox>345 10 364 98
7</bbox><bbox>363 10 382 987</bbox><bbox>381 10 400 987</bbox><bbox>399 10 418 987</bbox><bbox>417 10 4
36 987</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</bb
ox><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 85
0 987</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

SEED-VL-1.5 (Universal Proposal)





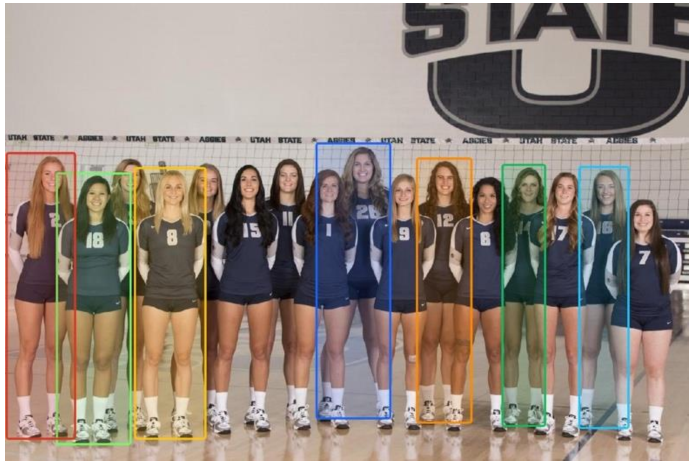
Question:请帮我识别一下图片，然后告诉我图中有哪些物体,并框选这些物体，物体类别名请用英文表示，对于每个物体，请提供其类别和边界框，格式为：[{"category": category, "bbox": "<bbox>x1 y1 x2 y2</bbox>"}, {"category": category, "bbox": "<bbox>x1 y1 x2 y2</bbox>"}]。记住每个坐标都要用 <bbox> </bbox> 进行包裹

Answer:

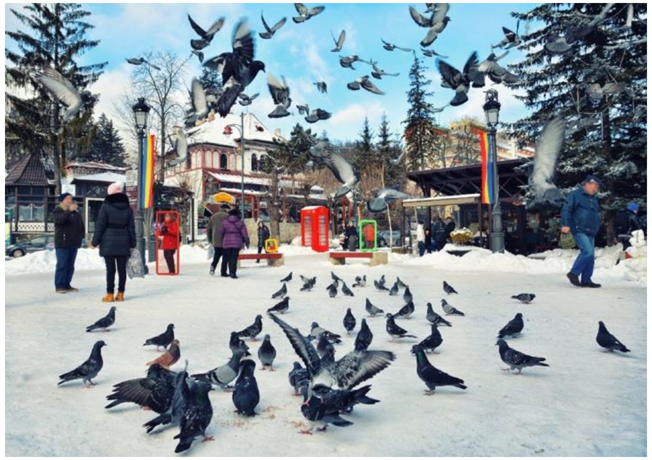
[{"category": "person", "bbox": "<bbox>3 392 95 736</bbox>"}, {"category": "person", "bbox": "<bbox>39 433 103 736</bbox>"}, {"category": "truck", "bbox": "<bbox>0 276 128 627</bbox>"}, {"category": "horse", "bbox": "<bbox>125 278 553 835</bbox>"}, {"category": "person", "bbox": "<bbox>232 138 350 564</bbox>"}, {"category": "truck", "bbox": "<bbox>207 171 700 576</bbox>"}, {"category": "megaphone", "bbox": "<bbox>701 217 745 297</bbox>"}, {"category": "truck", "bbox": "<bbox>778 234 957 527</bbox>"}, {"category": "horse", "bbox": "<bbox>543 564 803 825</bbox>"}, {"category": "person", "bbox": "<bbox>725 377 800 817</bbox>"}, {"category": "hat", "bbox": "<bbox>262 138 326 184</bbox>"}, {"category": "hat", "bbox": "<bbox>753 377 790 406</bbox>"}, {"category": "railing", "bbox": "<bbox>0 474 996 748</bbox>"}]

SEED-VL-1.5

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

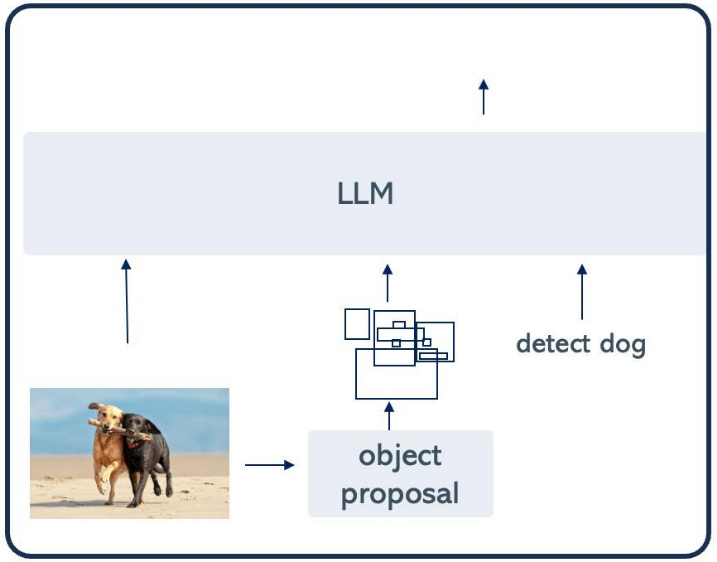


帮我框选图中穿红色衣服的人，以<bbox>x1 y1 x2 y2</bbox>的形式表示

What’s Next?

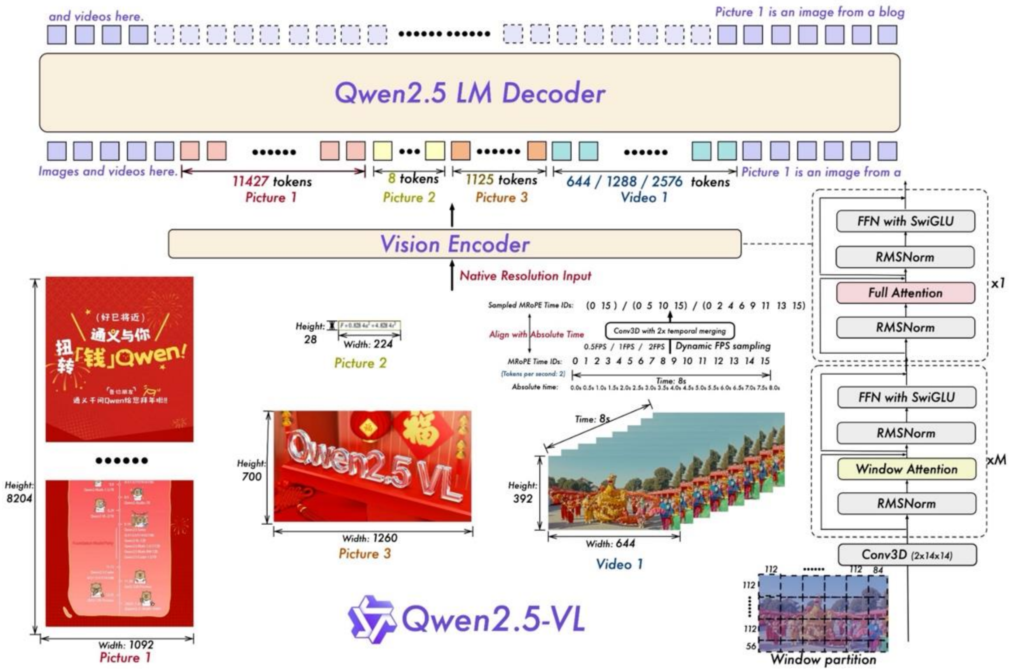
ChatRex **RexSeek** **Rex-Thinker**

- Proposal can not be provided in advance for many scenarios



Retrieval based method

- VLM has powerful detection capabilities of its own
- VLM has strong comprehension capabilities
- The proposal boxes can be inputted or not inputted at the same time.
- Support streaming or video input



从开集检测迈向通用视觉感知

感谢！