idea INTERNATIONAL DIGITAL ECONOMY ACADEMY
粤港澳大湾区数字经济研究院

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

Qing Jiang

2024 6.14

Team Introduction —— Computer Vision & Robotics, IDEA

idea

Dr. Lei Zhang

Chief Scientist

DETR Series Research in CVR

DAB-DETR ICLR 2022 → DN-DETR CVPR 2022 → DINO ICLR 2023 → Mask DINO CVPR 2023

DAB-DETR ICLR 2022
- Introduce Anchor Box to DETR
- Offer a better understanding of DETR Queries

DN-DETR CVPR 2022
- Introduce Denoising to DETR
- Reduce DETR training to 12 epochs

DINO ICLR 2023
- Further improve DAB and DN
- Achieve SOTA on COCO Detection

Mask DINO CVPR 2023
- Unify Object Detection and Segmentation

Open-Set Object Detection Research in CVR

Grounding DINO → T-Rex → T-Rex2 → Grounding DINO 1.5

Grounding DINO
- Text prompt based open-set object detection model

T-Rex
- Visual prompt based open-set object detection model

T-Rex2
- Combine both text prompt and visual prompt in one model

Grounding DINO 1.5
- SOTA text prompt based model with both Pro and Edge version models.

What is Object Detection?

idea

What is Object Detection?

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

Detection results from Grounding DINO

Role of Object Detection in 2024

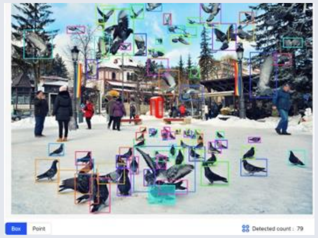
Object-Centric understanding is the perceptual basis for machine-physical world interaction.

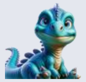
Solution for Hallucinations

: How many pigeons are there in this image

: In the image, there are approximately 59 pigeons.

User provide visual prompt

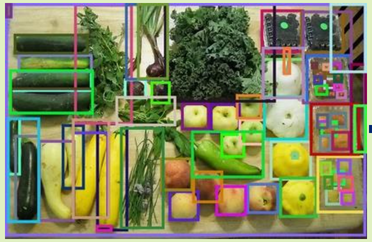




: 79

T-Rex2

Fine-grained Perception



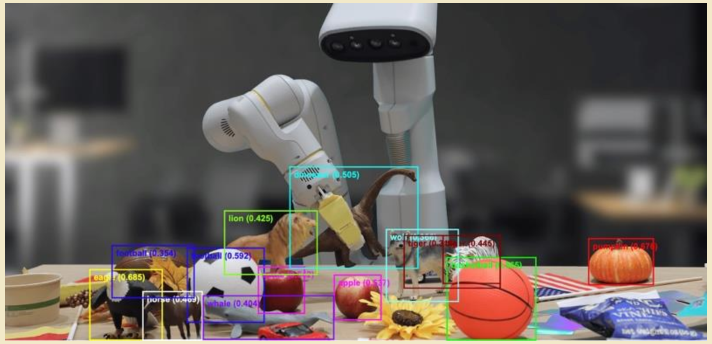
Box

Box

VLM

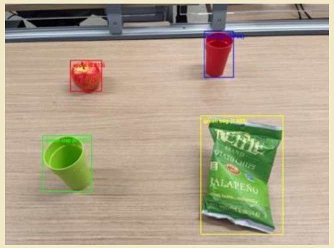
This image depicted … (Fine-grained long caption)

Eyes for Embodied AI



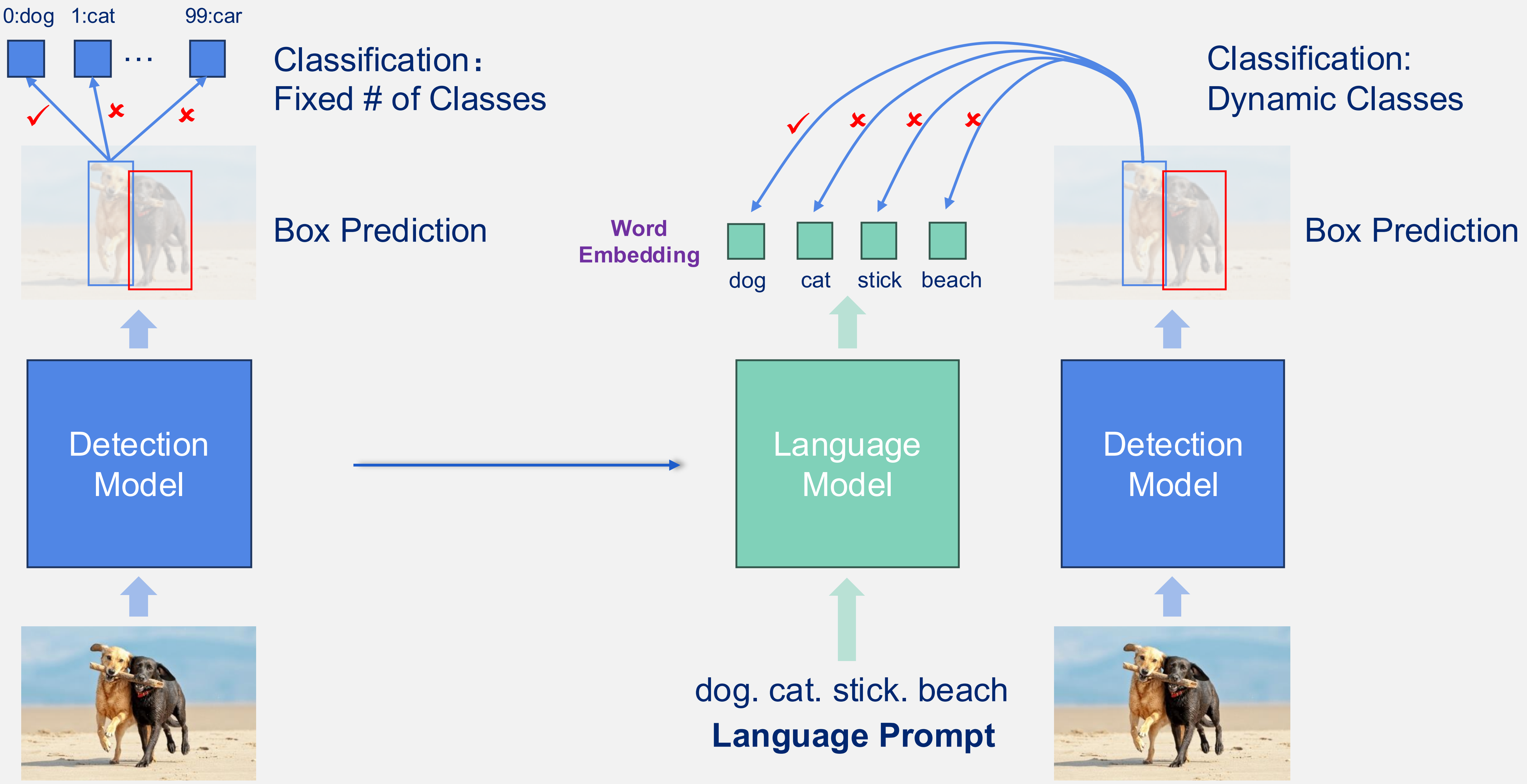
eagle . horse . lion. wolf .dinosaur . tiger . whale . apple . football . basketball . pumpkin

RT-2



apple . red cup . green cup . green bag

Paradigm Shift in Object Detection (Close-Set to Open-Set)



Goal of Open-Set Object Detection

- Given an image and arbitrary prompts
- We expect a model predicts all objects mentioned in prompts without finetuning.

Prompt* , *Detect

visual* *Text

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

Two Paths to Open-Set Object Detection (text prompt based)

idea

Path 1: **Referring** (CLIP-based)

ViLD
RCNN+CLIP

RegionCLIP
RCNN+finetune CLIP

OV-DETR
DETR+CLIP

X-Decoder
DETR+CLIP+caption

DetCLIP
RCNN+CLIP+large data

OpenSeeD

SEEM

2020 | 2021 | 2022 | 2023

MDETR
Grounding+DETR

GLIP
Grounding+ATSS
Extend to Detection

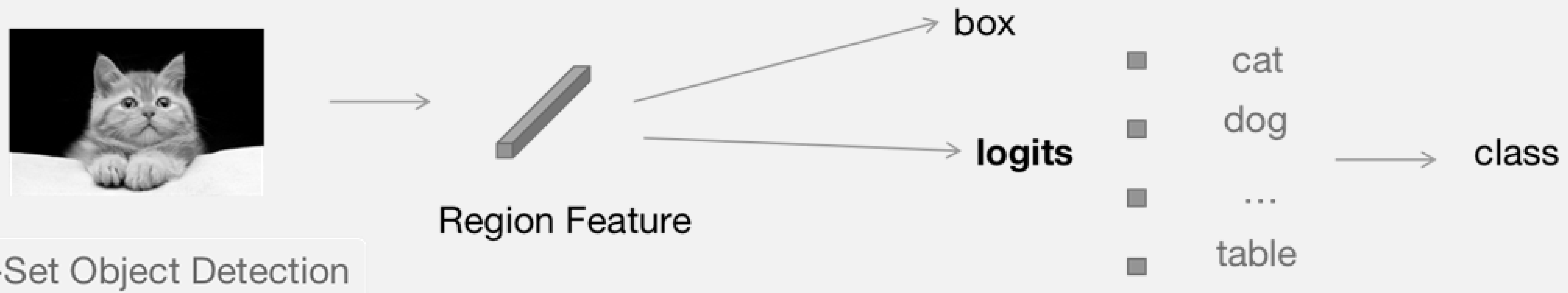
GLIPv2

Grounding DINO
Grounding + DINO

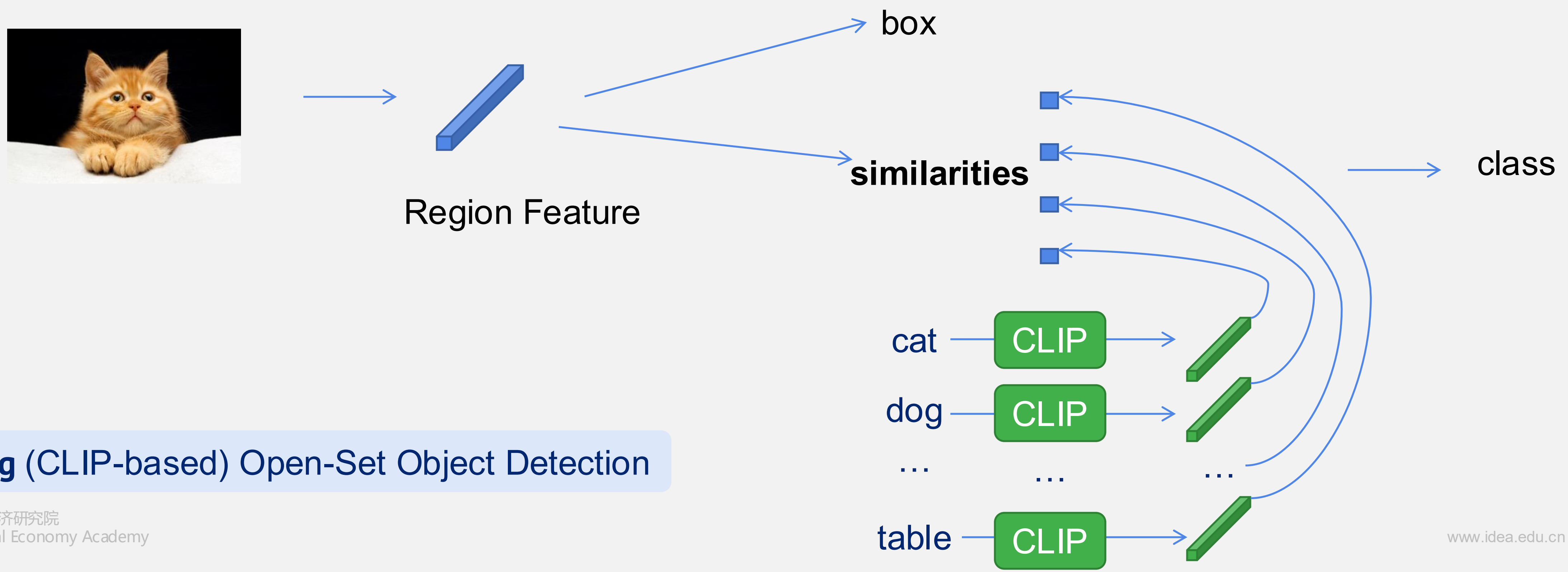
Path 2: **Grounding**

Path1: Referring (CLIP-based) Open-Set Object Detection

idea



Closed-Set Object Detection



Referring (CLIP-based) Open-Set Object Detection

Path2: Grounding based Open-Set Object Detection

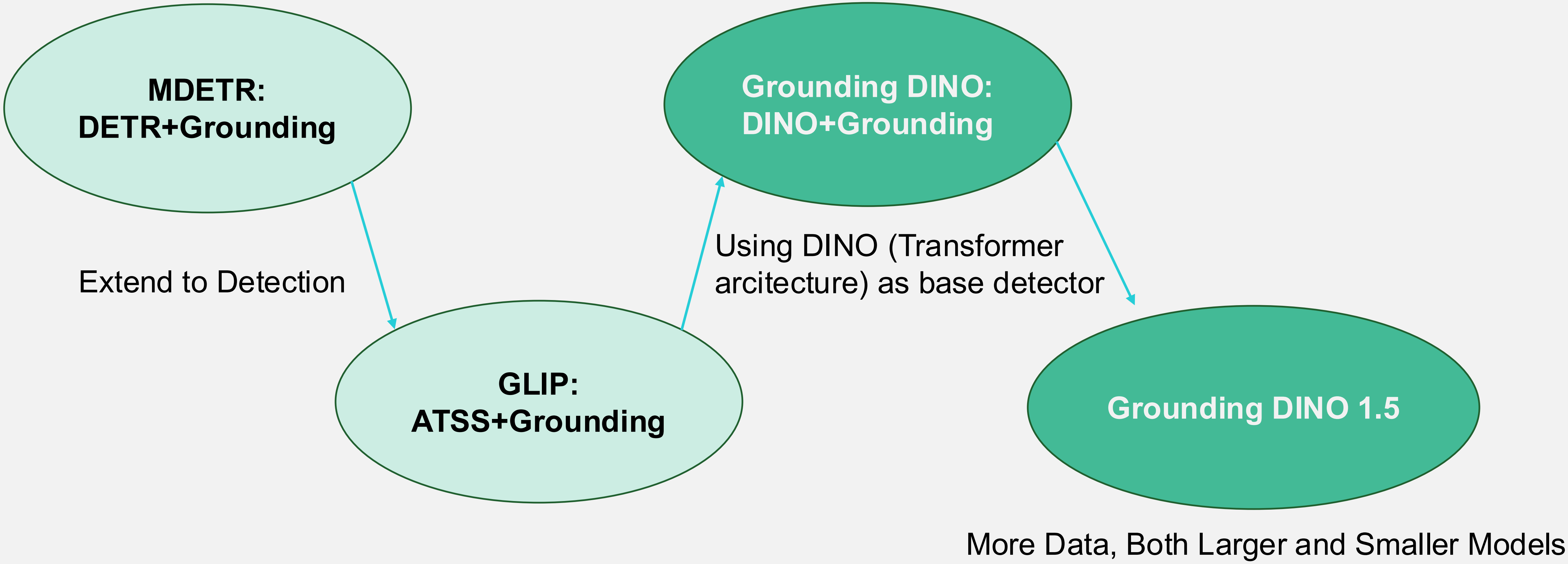
What is Grounding?

The image shows an individual lying prone on grassy ground, aiming a bolt-action rifle with their right eye close to the sight. This person is wearing a military-style uniform with a steel helmet with netting, suggesting a military or historical reenactment context. On the individual's back, you can see a backpack and what appears to be a canteen, both typical of military field equipment.

→ identifying the bounding boxes in an image that correspond to the noun phrases in a given sentence.

Grounding DINO -> Grounding DINO 1.5

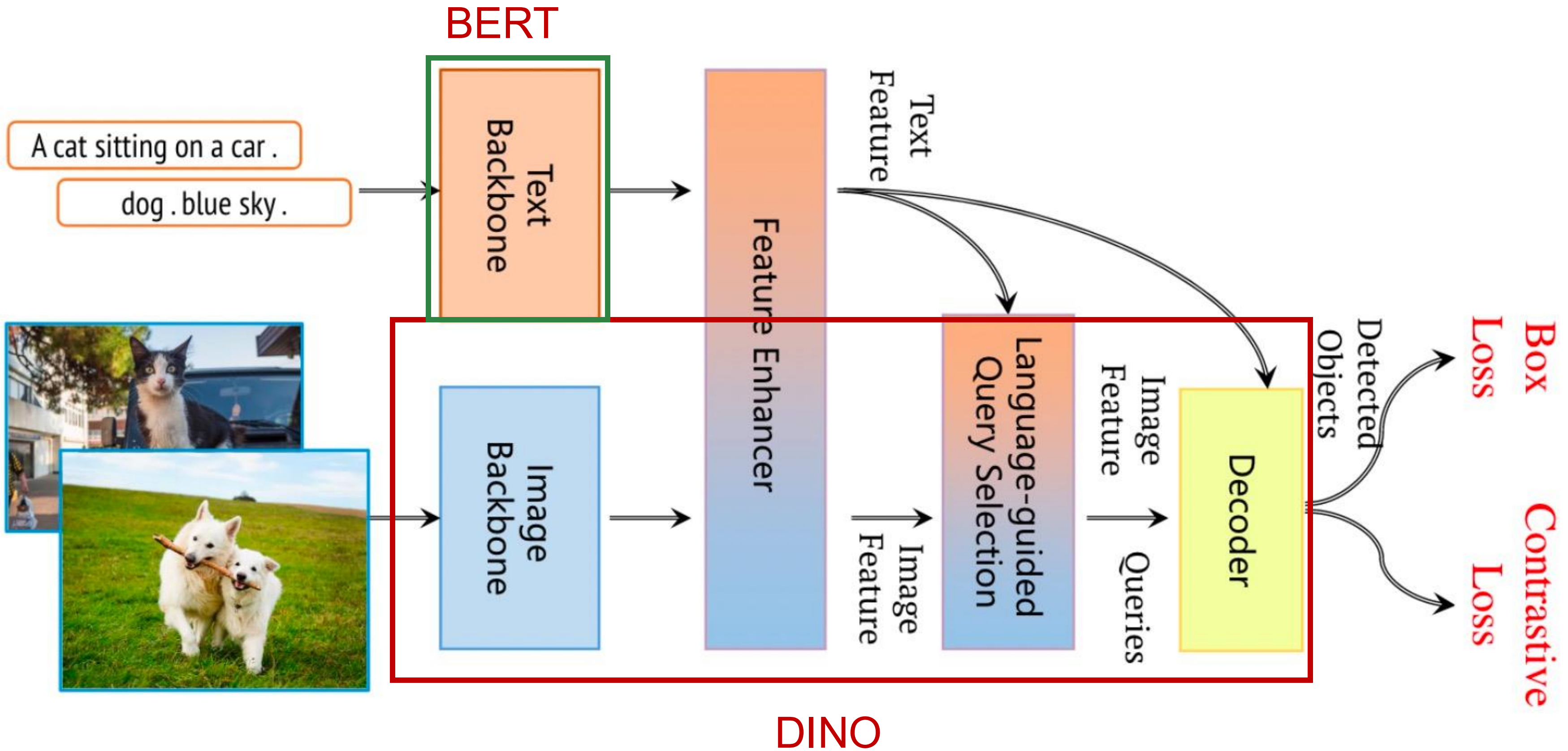
idea



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

We use “edge” for its dual meaning both as in pushing the boundaries and as in running on edge devices.

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

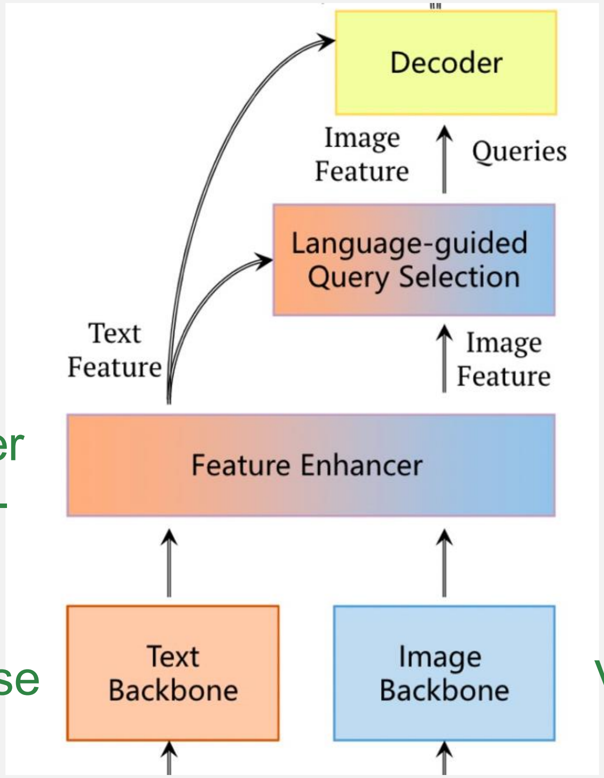


(a) Model Framework

Pro V.S. Edge: Overall Architecture

6 x Transformer Encoder, Multi-Scale

BERT-Base



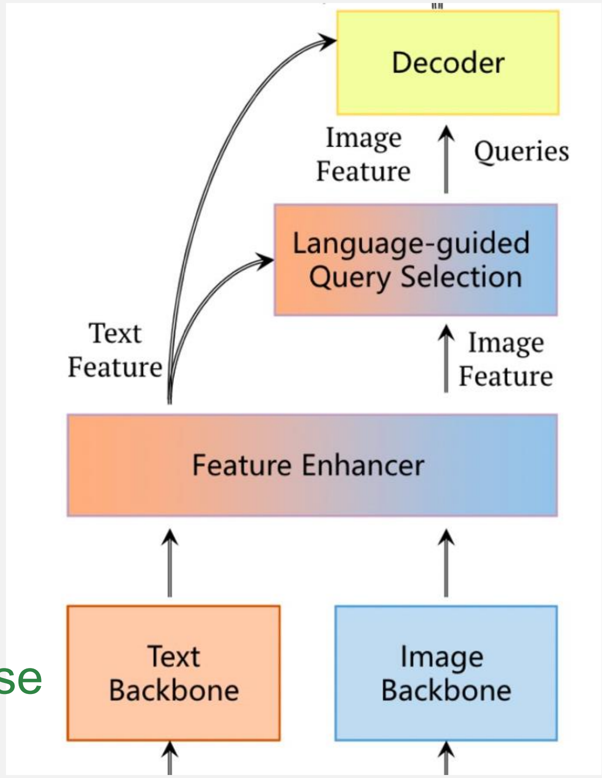
6 x Transformer Decoder

ViT-L

1.5 Pro

1 x Transformer Encoder, Multi Scale

BERT-Base



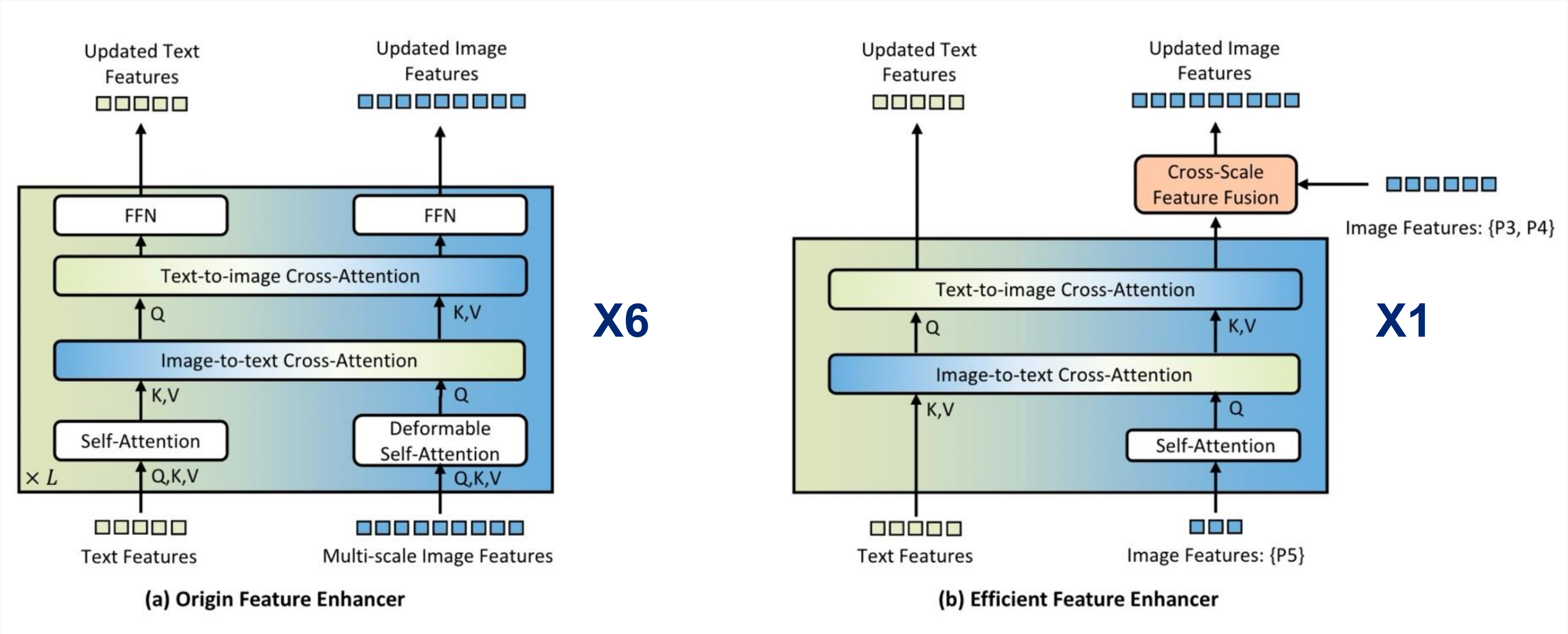
6 x Transformer Decoder

EfficientViT-L1

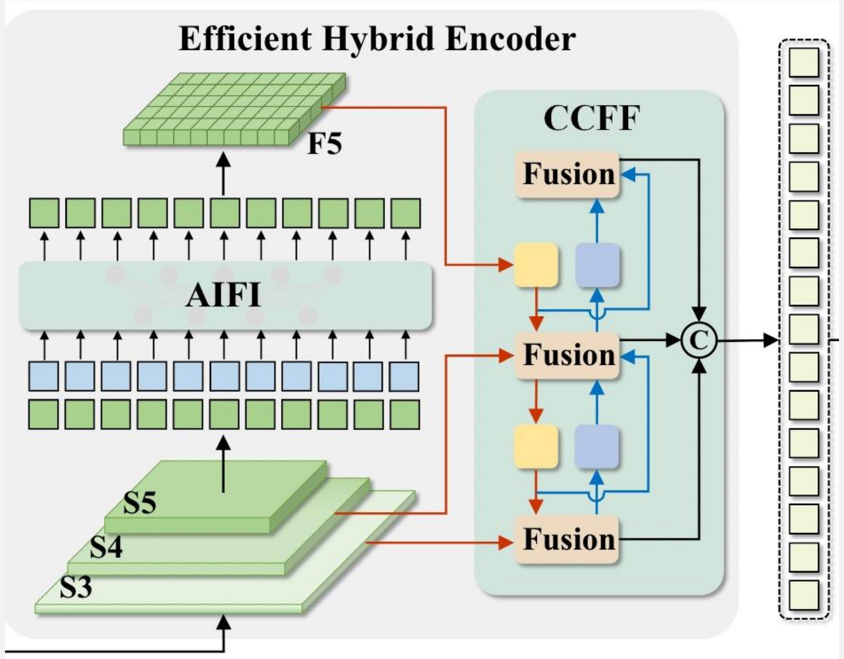
1.5 Edge

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

Pro V.S. Edge: Encoder



(a) Origin Feature Enhancer (b) Efficient Feature Enhancer



RT-DETR

Yian, Zhao, et al. “DETRs Beat YOLOs on Real-time Object Detection”. CVPR 2023.

Pro V.S. Edge: Running Time for Each Module (Pytorch Time)

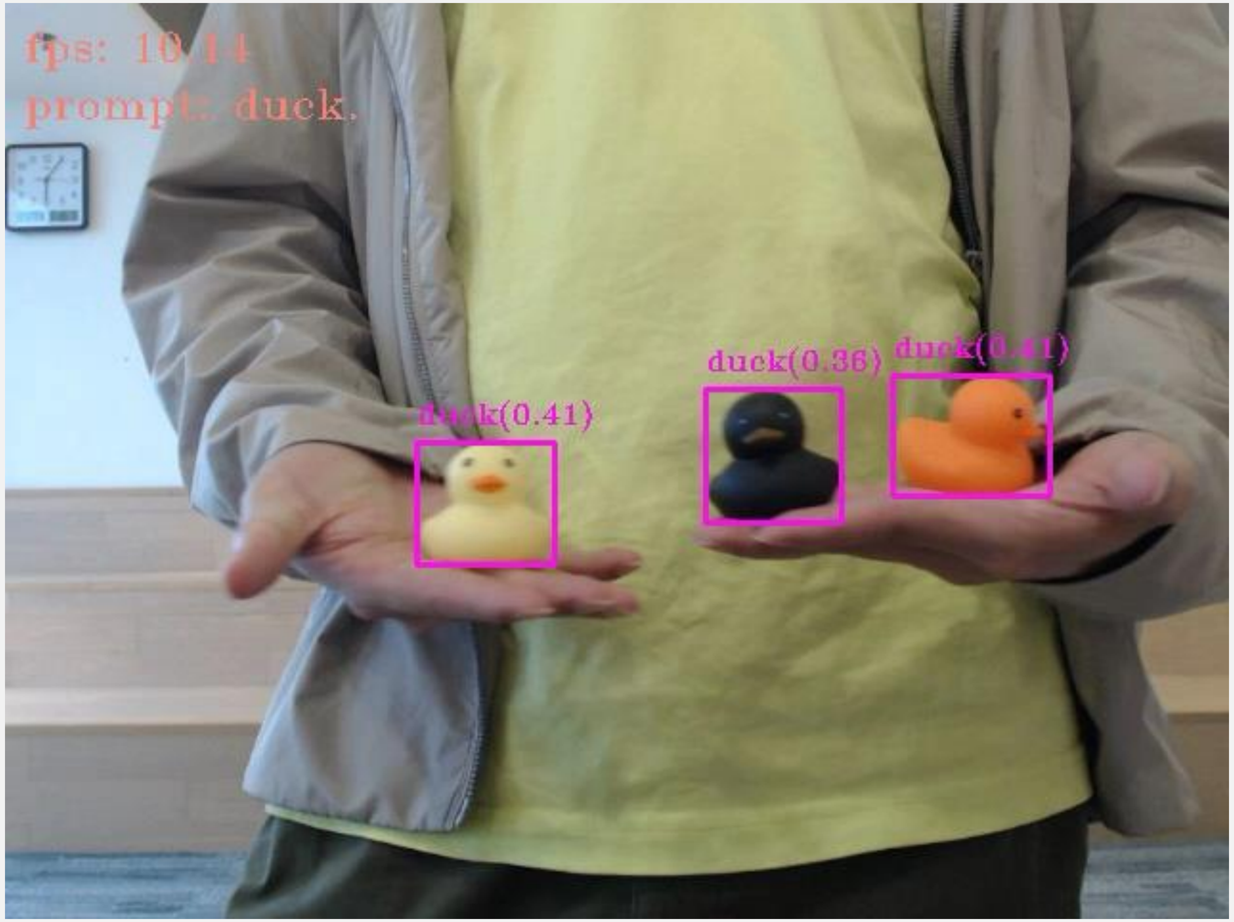
ViT-L

Model	BERT	Backbone	Encoder	Decoder	FPS
Pro	0.008 (1.7%)	0.367 （79.2%）	0.073 (15.7%)	0.015 (3.34%)	2.16
Edge	0.009 (15.3%)	0.012 (18.7%)	0.021 (32.75%)	0.021 (33.3%)	15.9

EfficientViT-L2

- Measured on a RTX 3090
- backbone takes the most time
- how to further optimize encoder and decoder time consumption is the next step

Deploy Edge on Edge Device (NVIDIA Orin NX)



Deploy Edge on NVIDIA Orin NX

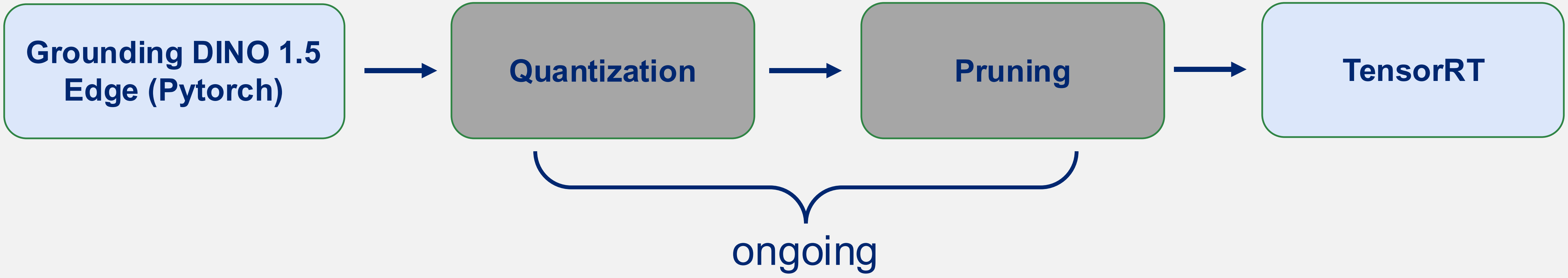
	Jetson AGX Orin series				Jetson Orin NX series		Jetson Orin Nano series		
	Jetson AGX Orin Developer Kit	Jetson AGX Orin 64GB	Jetson AGX Orin Industrial	Jetson AGX Orin 32GB	Jetson Orin NX 16GB	Jetson Orin NX 8GB	Jetson Orin Nano Developer Kit	Jetson Orin Nano 8GB	Jetson Orin Nano 4GB
AI Performance	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

Only the TOPS of Orin NX is close to that of the 3090, which means the model should be quantized to INT8 for optimal performance.

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

Deploy Edge on NVIDIA Orin NX



Challenges:

- Remove deformable attention in decoder

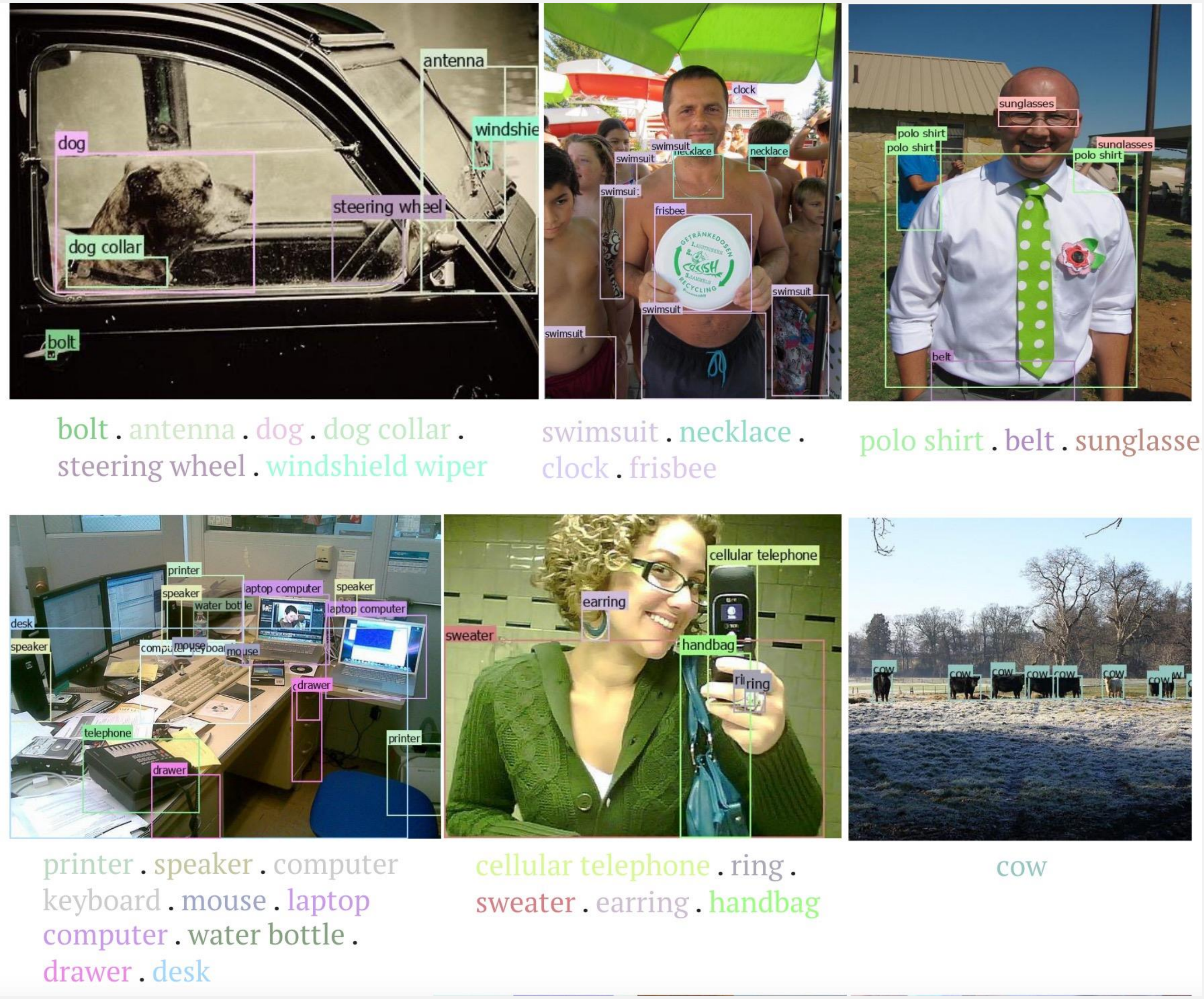
Results: Grounding DINO 1.5 Pro

Method	Backbone	Pre-training data	COCO	LVIS$^{\text{minival}}$				LVIS$^{\text{val}}$				ODinW35	ODinW13
			AP_{all}	AP_{all}	AP_r	AP_c	AP_f	AP_{all}	AP_r	AP_c	AP_f	AP_{avg}	AP_{avg}
Supervised Models (Pre-training data includes COCO, LVIS, etc.)													
GLIPv2 [35]	Swin-H [32]	FourODs,COCO,GoldG,CC15M,SBU	60.6	50.1	-	-	-	-	-	-	-	-	55.5
Grounding DINO [18]	Swin-L [19]	O365,OID,GoldG,Cap4M,COCO,RefC	60.7	33.9	22.2	30.7	38.8	-	-	-	-	-	-
APE (B) [24]	ViT-L	COCO,LVIS,O365,OID,VG	57.7	62.5	-	-	-	57.0	-	-	-	29.4	**59.8**
APE (D) [24]	ViT-L [6]	COCO,LVIS,O365,OID,VG,RefC,SA-1B,GoldG,PhraseCut	58.3	64.7	-	-	-	59.6	-	-	-	28.8	57.9
GLEE-Pro [27]	ViT-L [6]	GLEE-merged-10M (COCO,LVIS,etc)	62.0	-	-	-	-	55.7	49.2	-	-	-	53.4
Zero-shot Transfer Models													
OWL-ViT [22]	ViT-L [5]	O365,OID,VG,LiT	42.2	-	-	-	-	34.6	31.2	-	-	-	-
MDETR [11]	ResNet101 [8]	COCO,GoldG	-	22.5	7.4	22.7	25.0	-	-	-	-	-	-
GLIP [16]	Swin-L	FourODs,GoldG,Cap24M	49.8	37.3	28.2	34.3	41.5	26.9	17.1	23.3	35.4	-	52.1
Grounding DINO [18]	Swin-T	O365,GoldG,Cap4M	48.4	27.4	18.1	23.3	32.7	-	-	-	-	22.3	49.8
Grounding DINO [18]	Swin-L	O365,OID,GoldG	52.5	-	-	-	-	-	-	-	-	26.1	56.9
OpenSeeD [34]	Swin-L	COCO,O365	-	23.0	-	-	-	-	-	-	-	15.2	-
UniDetector [26]	ResNet50 [8]	COCO,O365,OID	-	-	-	-	-	19.8	18.0	19.2	21.2	-	47.3
OmDet-Turbo-B [36]	ConvNeXt-B [20]	O365,GoldG,PhraseCut,Hake,HOI-A	53.4	34.7	-	-	-	-	-	-	-	30.1	54.7
OWL-ST [21]	CLIP L/14 [23]	WebLI2B	-	40.9	41.5	-	-	35.2	36.2	-	-	24.4	53.0
MQ-GLIP [28]	Swin-L	O365	-	43.4	34.5	41.2	46.9	34.7	26.9	32.0	41.3	23.9	54.1
DetCLIP [30]	Swin-L	O365,GoldG,YFCC1M	-	38.6	36.0	38.3	39.3	28.4	25.0	27.0	31.6	-	-
DetCLIPv2 [29]	Swin-L	O365,GoldG,CC15M	-	44.7	43.1	46.3	43.7	36.6	33.3	36.2	38.5	-	-
DetCLIPv3 [31]	Swin-L	O365,V3Det,GoldG,GranuCap50M	-	48.8	49.9	49.7	47.8	41.4	41.4	40.5	42.3	-	-
YOLO-World [3]	YOLOv8-L [10]	O365,GoldG,CC3M	45.1	35.4	27.6	34.1	38.0	-	-	-	-	-	-
DINOv [14]	Swin-L	COCO,SA-1B	-	-	-	-	-	-	-	-	-	15.7	-
T-Rex2 (visual) [9]	Swin-L	O365,OID,HierText,CrowdHuman,SA-1B	46.5	47.6	45.4	46.0	49.5	45.3	43.8	42.0	49.5	27.8	-
T-Rex2 (text) [9]	Swin-L	O365,OID,GoldG,CC3M,SBU,LAION	52.2	54.9	49.2	54.8	**56.1**	45.8	42.7	43.2	**50.2**	22.0	-
Grounding DINO 1.5 Pro (zero-shot)	ViT-L [6]	Grounding-20M	**54.3**	**55.7**	**56.1**	**57.5**	54.1	**47.6**	**44.6**	**47.9**	48.7	**30.2**	58.7

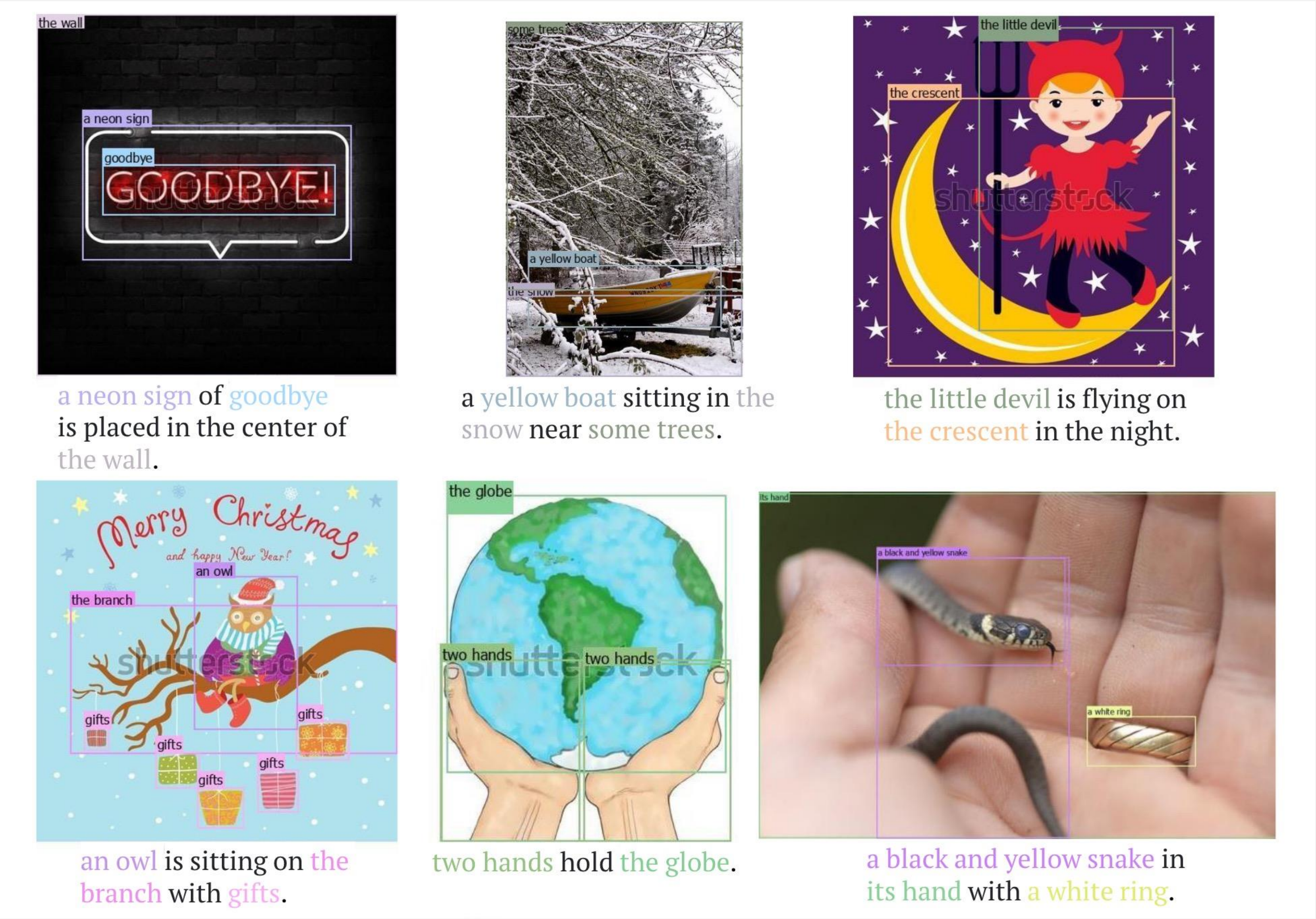
Results: Grounding DINO 1.5 Edge

Method	Backbone	Pre-training data	test size	COCO	$\text{LVIS}^{\text{minival}}$				LVIS^{val}				FPS(A100/TensorRT)	FPS(Orin NX)
					AP_{all}	AP_{r}	AP_{c}	AP_{f}	AP_{all}	AP_{r}	AP_{c}	AP_{f}		
End-to-End Open-Set Object Detection														
GLIP-T	Swin-T	O365,GoldG,Cap4M	800×1333	46.3	26.0	20.8	21.4	31.0	-	-	-	-	-	-
Grounding DINO-T	Swin-T	O365,GoldG,Cap4M	800×1333	48.4	27.4	18.1	23.3	32.7	-	-	-	-	9.4 / 42.6	1.1
Real-time End-to-End Open-Set Object Detection														
YOLO-Worldv2-S†	YOLOv8-S	O365,GoldG	640×640	-	22.7	16.3	20.8	25.5	17.3	11.3	14.9	22.7	47.4 / -	-
YOLO-Worldv2-M†	YOLOv8-M	O365,GoldG	640×640	-	30.0	25.0	27.2	33.4	23.5	17.1	20.0	30.1	42.7 / -	-
YOLO-Worldv2-L†	YOLOv8-L	O365,GoldG	640×640	-	33.0	22.6	32.0	35.8	26.0	18.6	23.0	**32.6**	37.4 / -	-
YOLO-Worldv2-L†	YOLOv8-L	O365,GoldG,CC3M-Lite	640×640	-	32.9	25.3	31.1	35.8	26.1	20.6	22.6	32.3	37.4 / -	-
OmDet-Turbo-T‡	Swin-T	O365,GoldG	640×640	42.5	30.3	-	-	-	-	-	-	-	21.5 / 140.0	-
Grounding DINO 1.5 Edge	EfficientViT-L1	Grounding-20M	640×640	42.9	33.5	28.0	34.3	33.9	27.3	26.3	25.7	29.6	21.7 / 111.6	10.7
Grounding DINO 1.5 Edge	EfficientViT-L1	Grounding-20M	800×1333	**45.0**	**36.2**	**33.2**	**36.6**	**36.3**	**29.3**	**28.1**	**27.6**	31.6	18.5 / 75.2	5.5

Visualization Results (Long-tailed Object Detection)



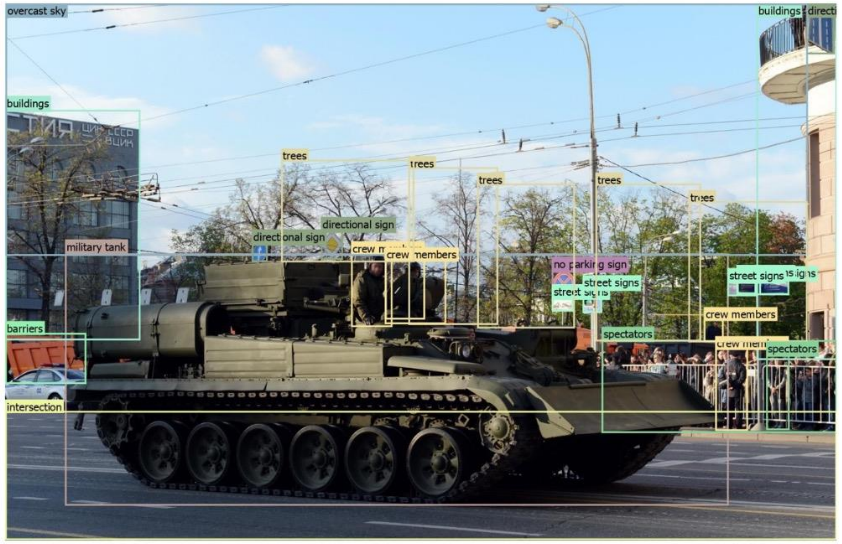
Visualization Results (Short Caption Object Detection)



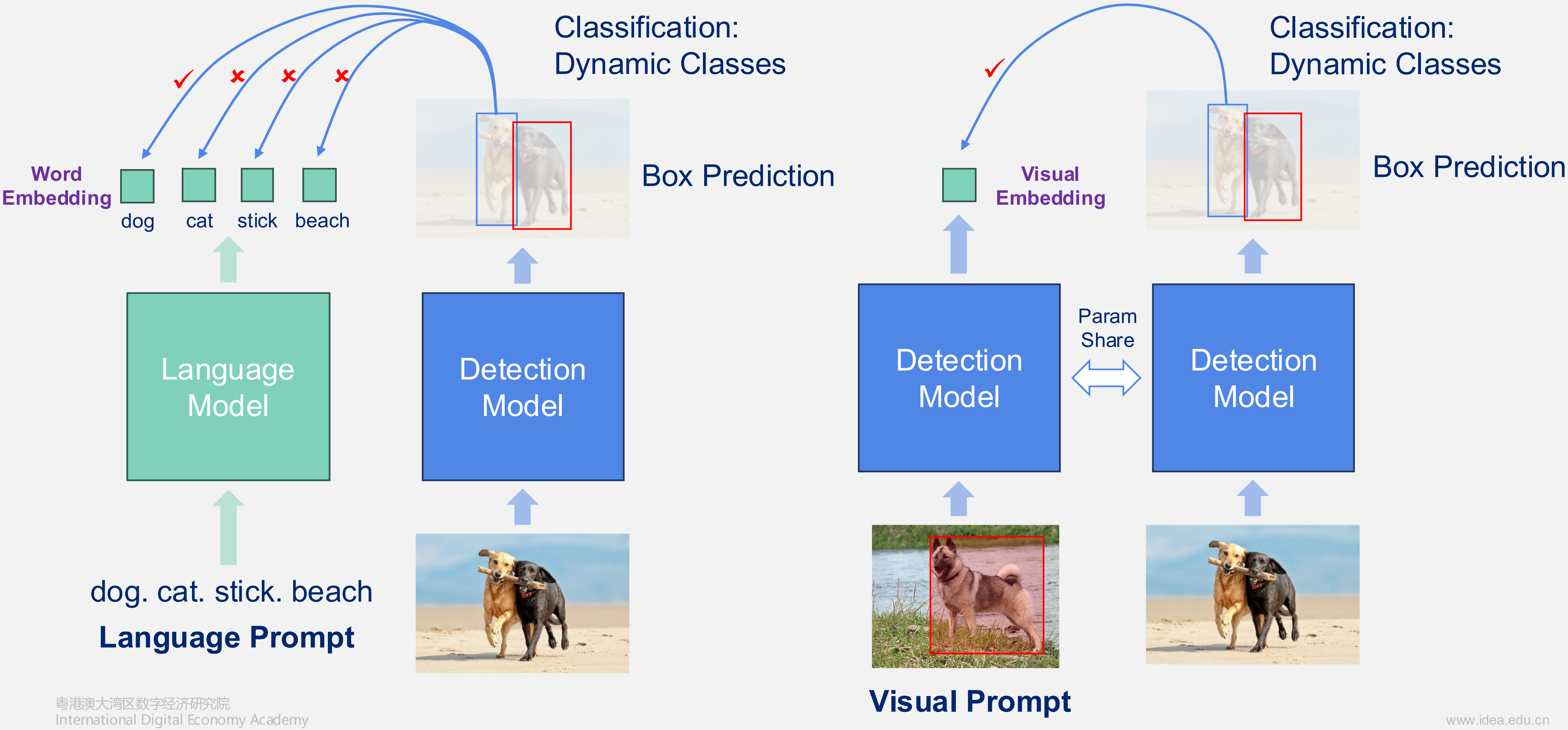
Visualization Results (Long Caption Object Detection)

The image shows a car on display at a motor show, prominently featuring the Fiat logo in red along the side. The car is painted gray with a glossy finish, highlighted by orange and black racing stripes and red accents on the front splitter. Its sporty design includes a large rear spoiler and white multi-spoke wheels, suggesting additional aerodynamic features. In the background. Several people navigate the space, including an individual in a blue jacket with a logo, likely an event staff member.

The image shows a military tank rolling along a city street, with crew members visible on top. Painted in camouflage and equipped with a long barrel, the tank navigates through an intersection, suggesting a setting likely in Europe or Russia. The surroundings, marked by street signs, a no parking sign, and a directional sign pointing right, indicate the tank's participation in a parade or military demonstration. Spectators, including adults and children, stand behind barriers on the sidewalk, observing the tank against a backdrop of buildings and leafless trees, under an overcast sky.



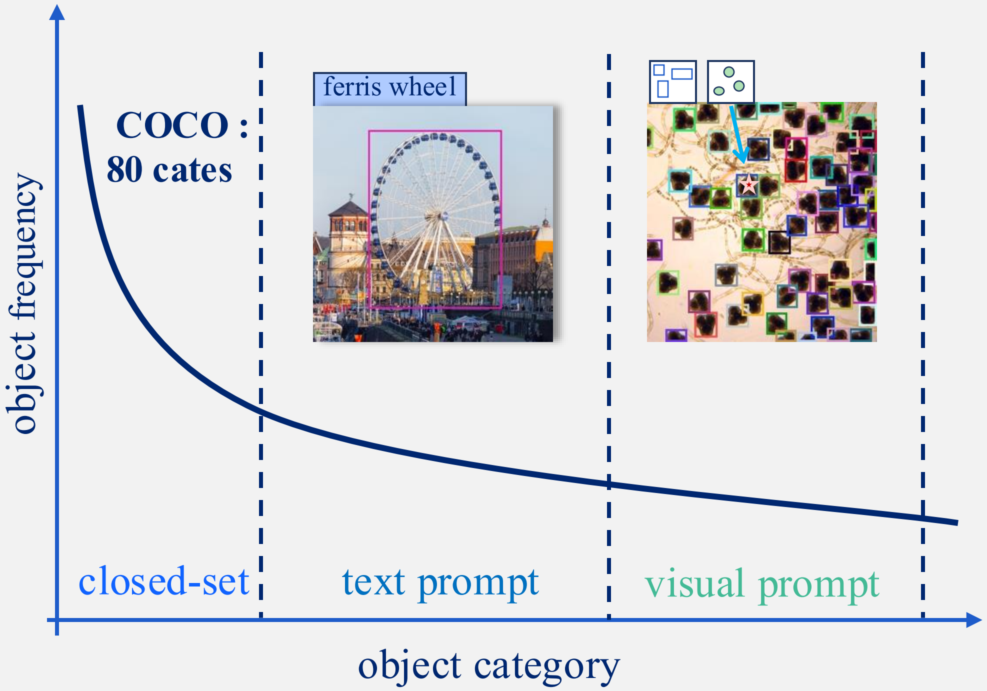
Text Prompt v.s. Visual Prompt

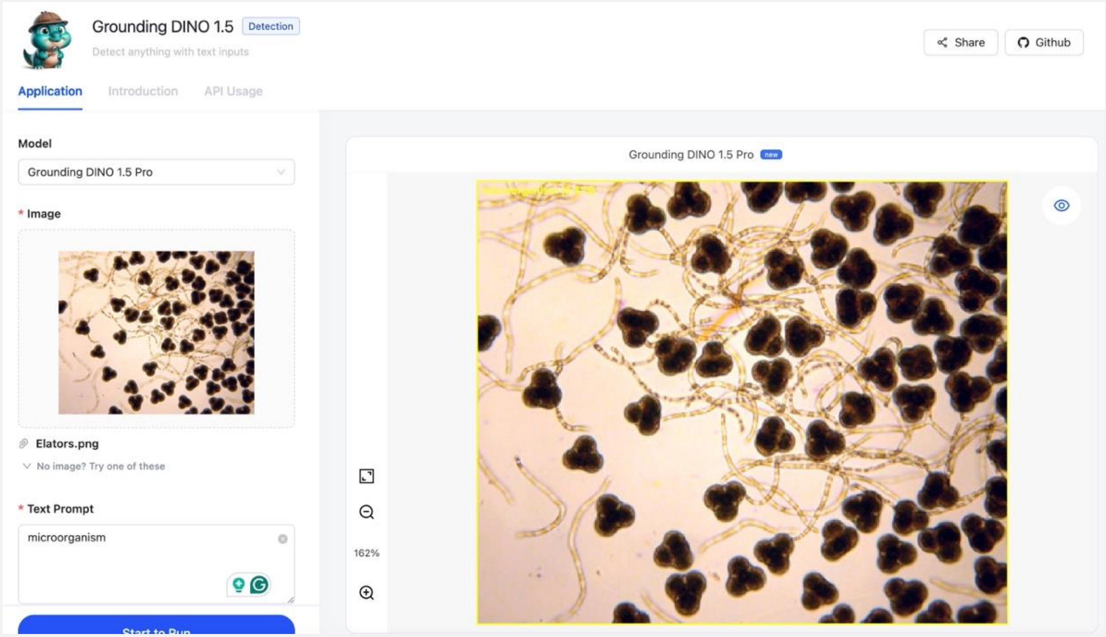


Text Prompt v.s. Visual Prompt

Text Prompt

- describe objects in natural language
- require modality alignment, suffers from long-tailed data shortage
- fall short in describe object that are hard to describe in language

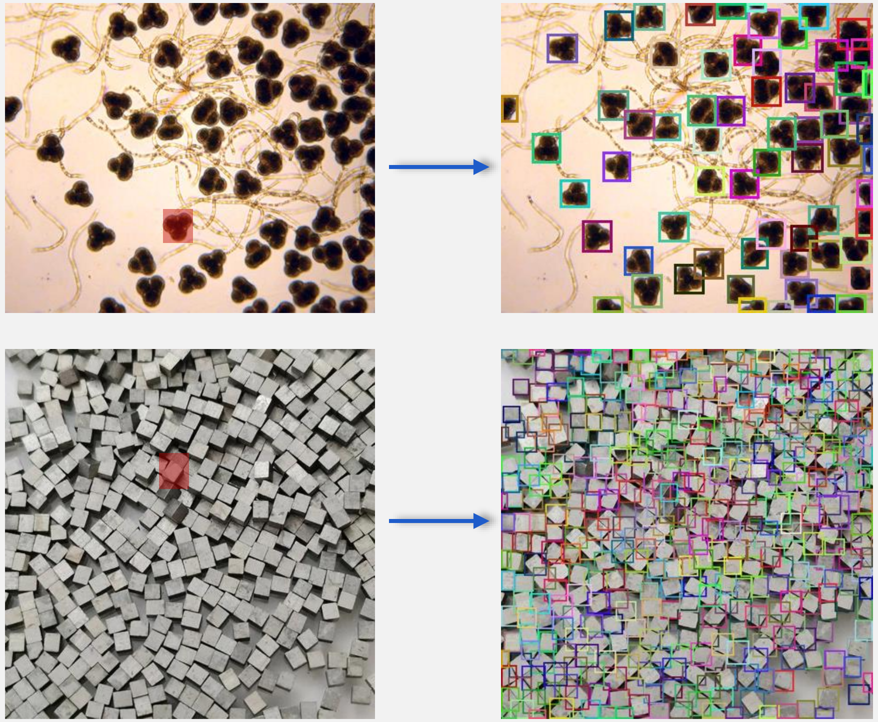


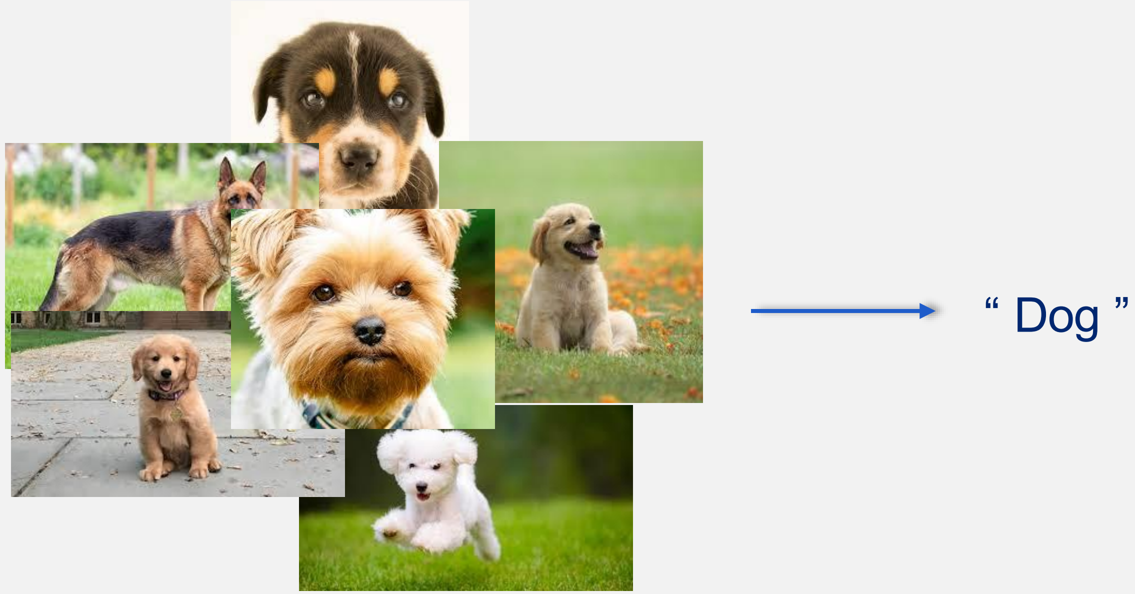


Text Prompt v.s. Visual Prompt

Visual Prompt

- describe objects through visual examples
- less effective at capturing the general concept





require many examples to convey a general concept

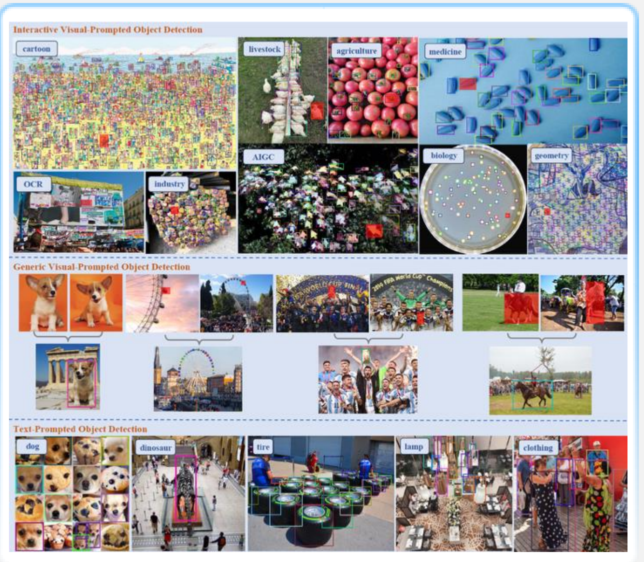
T-Rex2: Combine both Text Prompt and Visual Prompt

idea

T-Rex2: Towards Generic Object Detection via Text-Visual Prompt Synergy

Qing Jiang[1,2] , Feng Li[1,3] , Zhaoyang Zeng[1] , Tianhe Ren[1] , Shilong Liu[1,4] , Lei Zhang[1†]
[1]International Digital Economy Academy (IDEA)
[2]SCUT [3]HKUST [4]Tsinghua
mountchicken@outlook.com , fliay@connect.ust.hk , lius120@mails.tsinghua.edu.cn
{rentianhe, zengzhaoyang, leizhang}@idea.edu.cn
trex-counting.github.io

ECCV 2024



Method	Prompt Type	Backbone	COCO-Val Zero-Shot	LVIS Zero-Shot								ODinW Zero-Shot		Roboflow100 Zero-Shot
			val-80	minival-804				val-1203				35val		100val
			AP	AP	AP_f	AP_c	AP_r	AP	AP_f	AP_c	AP_r	AP_{avg}	AP_{med}	AP_{avg}
GLIP-T [19]	Text	Swin-T	46.7	26.0	31.0	21.4	20.8	17.2	25.5	12.5	10.1	19.6	5.1	-
GLIP-L [19]	Text	Swin-L	49.8	37.3	41.5	34.3	28.2	26.9	35.4	23.3	17.1	23.4	11.0	8.6
Grounding DINO [24]	Text	Swin-T	48.4	27.4	32.7	23.3	18.1	-	-	-	-	22.3	11.9	-
Grounding DINO [24]	Text	Swin-L	**52.5**	33.9	38.8	30.7	22.2	-	-	-	-	26.1	18.4	-
DetCLIPv2 [47]	Text	Swin-T	-	40.4	40.0	41.7	36.0	-	-	-	-	-	-	-
DetCLIPv2 [47]	Text	Swin-L	-	44.7	43.7	46.3	43.1	-	-	-	-	-	-	-
DINOv [17]	Visual-G	Swin-T	-	-	-	-	-	-	-	-	-	14.9	5.4	-
DINOv [17]	Visual-G	Swin-L	-	-	-	-	-	-	-	-	-	15.7	4.8	-
T-Rex2	Text	Swin-T	45.8	42.8	46.5	39.7	37.4	34.8	41.2	31.5	29.0	18.0	4.7	8.2
T-Rex2	Visual-G	Swin-T	38.8	37.4	41.8	33.9	29.9	34.9	41.1	30.3	32.4	23.6	17.5	17.4
T-Rex2	Text	Swin-L	52.2	**54.9**	**56.1**	**54.8**	**49.2**	**45.8**	**50.2**	**43.2**	42.7	22.0	7.3	10.5
T-Rex2	Visual-G	Swin-L	46.5	47.6	49.5	46.0	45.4	45.3	49.5	42.0	**43.8**	**27.8**	**20.5**	**18.5**

Table 1. **One suit of weights** for zero-shot object detection. Red denotes regions where text prompt excels over visual prompt, while green signifies regions favoring visual prompts.

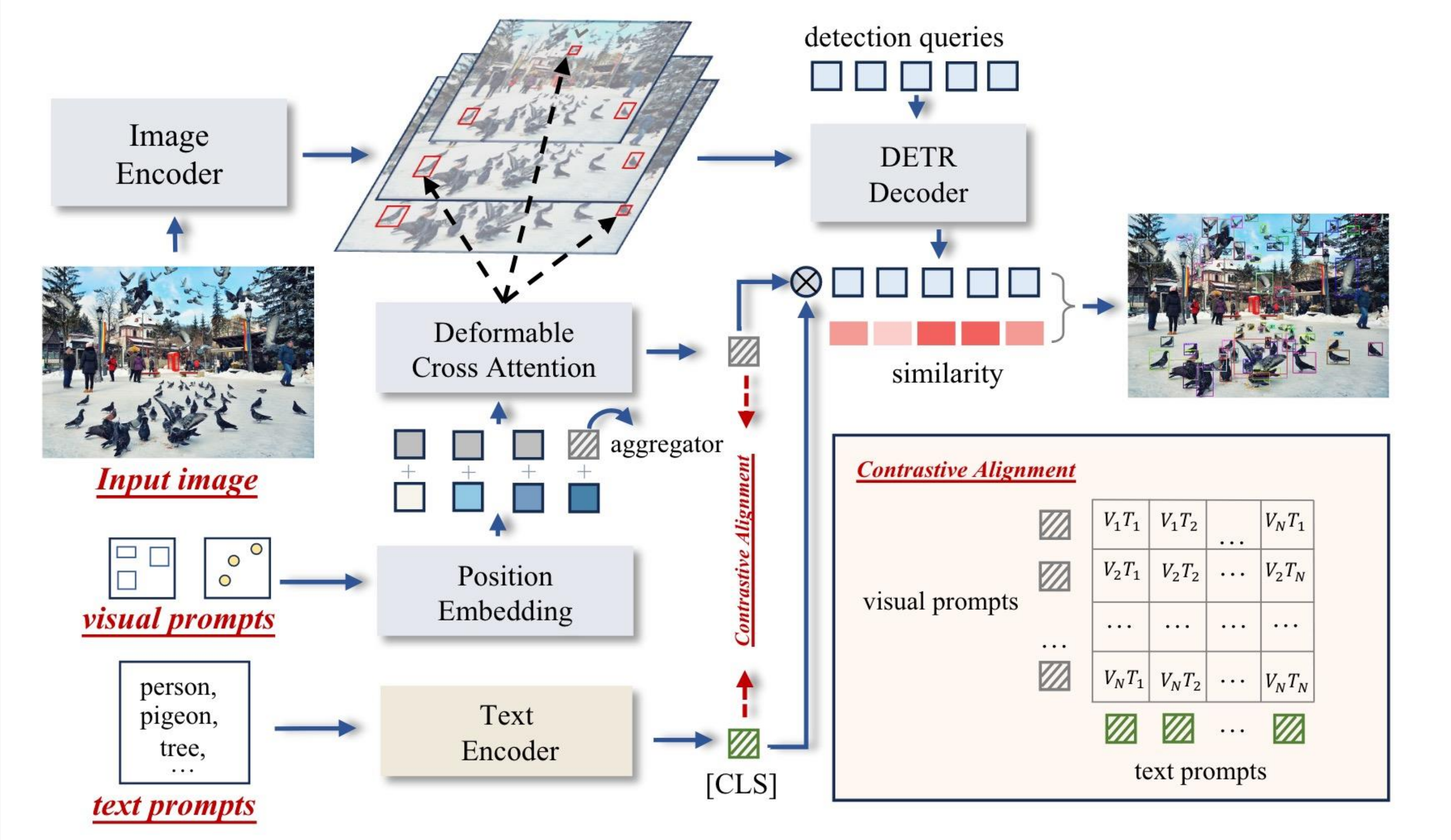
Alessandro Ferrari • 3rd+ + Follow
Founder, CEO @ARGO Vision | Contra...
View my blog
1w •

🦖🦖T-Rex 2: a new SOTA is out!🦖🦖

👉IDEA unveils a novel (VERY STRONG) open-set object detector model. Strong zero-shot capabilities, suitable for various scenarios with only one suit of weights. Demo and Source Code released💙

T-Rex2: Combine both Text Prompt and Visual Prompt

T-Rex2: Combine both Text Prompt and Visual Prompt



DINO-based End-to-End model

Visual Prompt Encoder: Deformable Cross Attention

$$B = \mathrm{Linear}(\mathrm{PE}(b_1, ...b_K); \theta_B) : \mathbb{R}^{K\times 4D} \rightarrow \mathbb{R}^{K\times D}$$
$$P = \mathrm{Linear}(\mathrm{PE}(p_1, ...p_K); \theta_P) : \mathbb{R}^{K\times 2D} \rightarrow \mathbb{R}^{K\times D}$$

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

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

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

Text Prompt Encoder: CLIP

Modality Alignment: Contrastive Learning

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

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, Lei Zhang. **T-Rex2: Towards Generic Object Detection via Text-Visual Prompt Synergy**. Arxiv 2024

Joint prompt leads to generic object detection

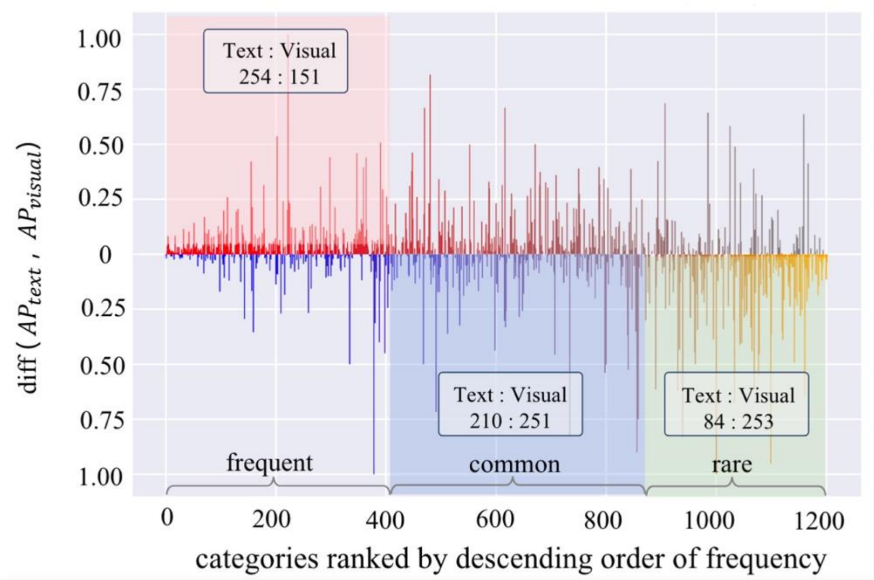
Zero-Shot Generic Object Detection

Method	Prompt Type	Backbone	COCO-Val Zero-Shot	LVIS Zero-Shot								ODinW Zero-Shot		Roboflow100 Zero-Shot
			val-80	minival-804				val-1203				35val		100val
			AP	AP	AP_f	AP_c	AP_r	AP	AP_f	AP_c	AP_r	AP_{avg}	AP_{med}	AP_{avg}
GLIP-T [19]	Text	Swin-T	46.7	26.0	31.0	21.4	20.8	17.2	25.5	12.5	10.1	19.6	5.1	-
GLIP-L [19]	Text	Swin-L	49.8	37.3	41.5	34.3	28.2	26.9	35.4	23.3	17.1	23.4	11.0	8.6
Grounding DINO [24]	Text	Swin-T	48.4	27.4	32.7	23.3	18.1	-	-	-	-	22.3	11.9	-
Grounding DINO [24]	Text	Swin-L	**52.5**	33.9	38.8	30.7	22.2	-	-	-	-	26.1	18.4	-
DetCLIPv2 [47]	Text	Swin-T	-	40.4	40.0	41.7	36.0	-	-	-	-	-	-	-
DetCLIPv2 [47]	Text	Swin-L	-	44.7	43.7	46.3	43.1	-	-	-	-	-	-	-
DINOv [17]	Visual-G	Swin-T	-	-	-	-	-	-	-	-	-	14.9	5.4	-
DINOv [17]	Visual-G	Swin-L	-	-	-	-	-	-	-	-	-	15.7	4.8	-
T-Rex2	Text	Swin-T	45.8	42.8	46.5	39.7	37.4	34.8	41.2	31.5	29.0	18.0	4.7	8.2
T-Rex2	Visual-G	Swin-T	38.8	37.4	41.8	33.9	29.9	<u>34.9</u>	41.1	30.3	<u>32.4</u>	<u>23.6</u>	<u>17.5</u>	<u>17.4</u>
T-Rex2	Text	Swin-L	<u>52.2</u>	**54.9**	**56.1**	**54.8**	**49.2**	**45.8**	**50.2**	**43.2**	42.7	22.0	7.3	10.5
T-Rex2	Visual-G	Swin-L	46.5	47.6	49.5	46.0	45.4	45.3	49.5	42.0	**43.8**	**27.8**	**20.5**	**18.5**

common and frequent case
Text prompt better

rare and novel case
Visual prompt better

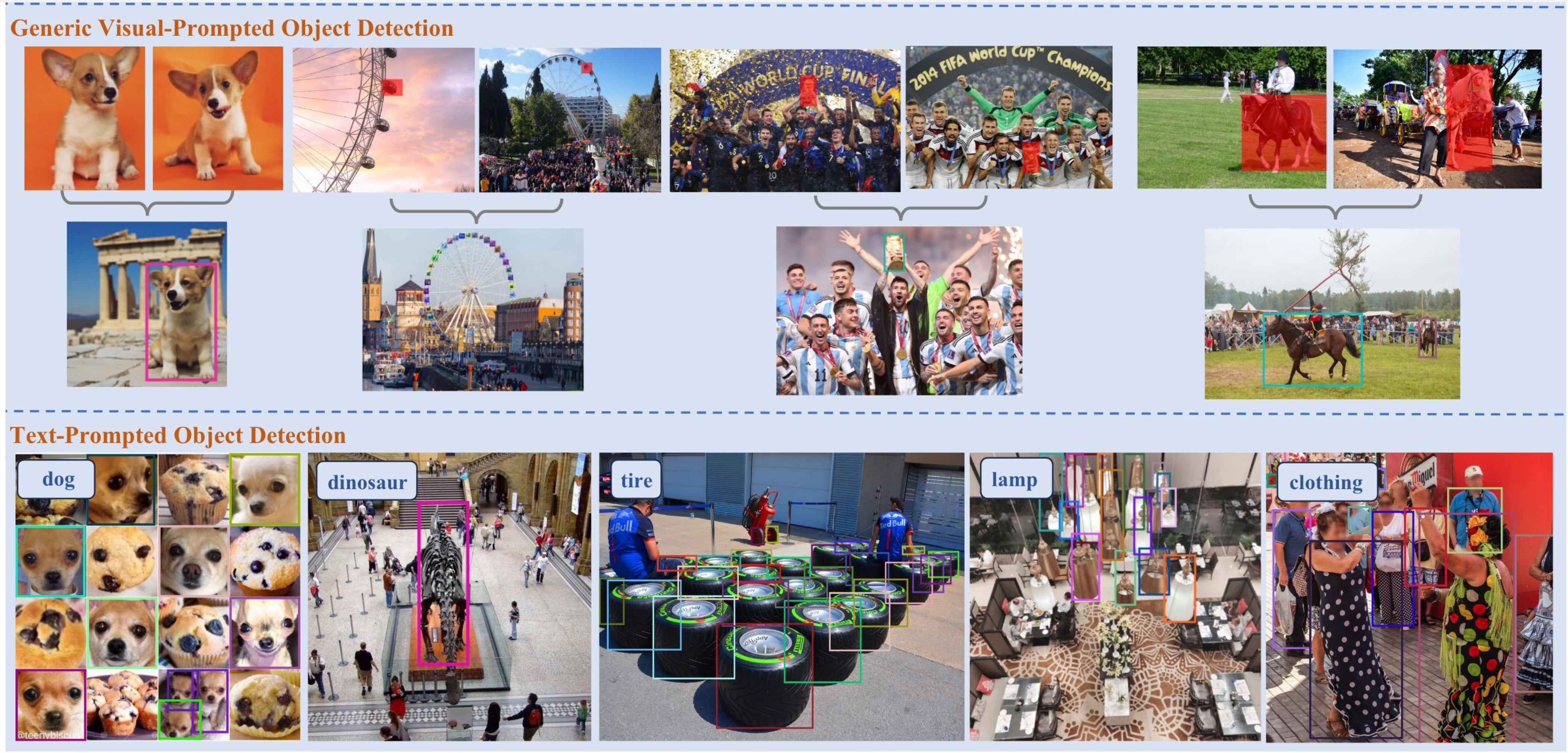
Text prompt v.s. Visual prompt on LVIS



- Text prompt is good at common and frequent object, while visual prompt succeed in rare and novel scenarios.

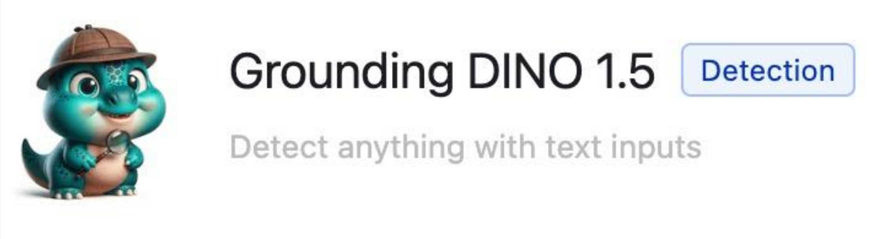
Joint prompt leads to generic object detection

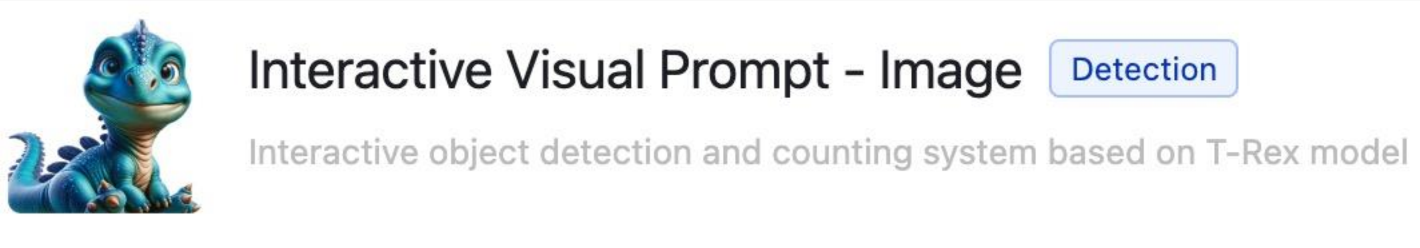
Joint prompt leads to generic object detection



Summary

- **Open-set** detection is the next BIG problem after closed-set detection
- **Prompt** is a new way to transform open-set detection
  - Text prompt: effective to cover head and middle concepts
  - Visual prompt: effective to cover more long tailed concepts
- **Grounded** understanding is key to multimodality intelligence





https://deepdataspace.com/

Thanks!

Qing Jiang

2024 6.14