

Ceptor: Vision-Language Model-Infused Diverse Guidance for Detecting Anything

Jinyang Li¹ Bin-Bin Gao² Weifu Fu² Jingnan Luo³

Hanqiu Deng² Yue Guo⁴ Jun Liu² Yong Liu² Chengjie Wang² Wenbing Tao^{1†}

¹Huazhong University of Science and Technology ²YouTu Lab, Tencent

³Southern University of Science and Technology ⁴Macau University of Science and Technology

Sep. 08 — Sep. 12, 2026, Malmo

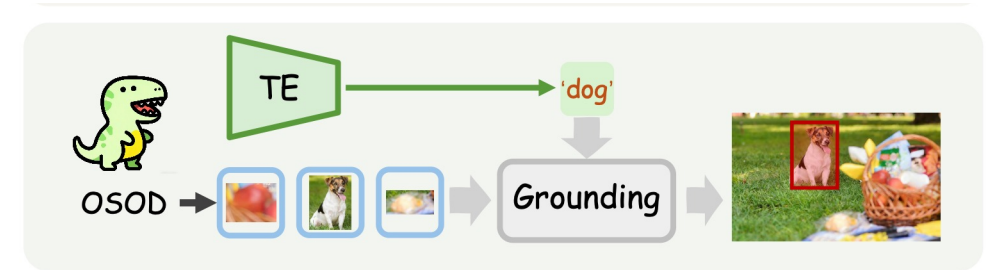
J. Li and B.-B. Gao—Equal Contribution.

Open-Set Object Detection

Open-set object detection aims to detect objects beyond fixed training categories.

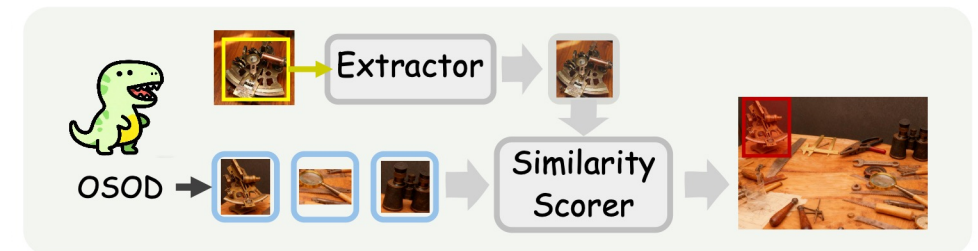
✓ **Text-prompt detectors** excel at high-abstraction semantics, but fail on low-abstraction objects

- ✗ long-tail categories;
- ✗ domain-specific entities;
- ✗ hard-to-describe instances;



✓ **Visual-prompt detectors** excel at low-abstraction objects, but show poor generalization

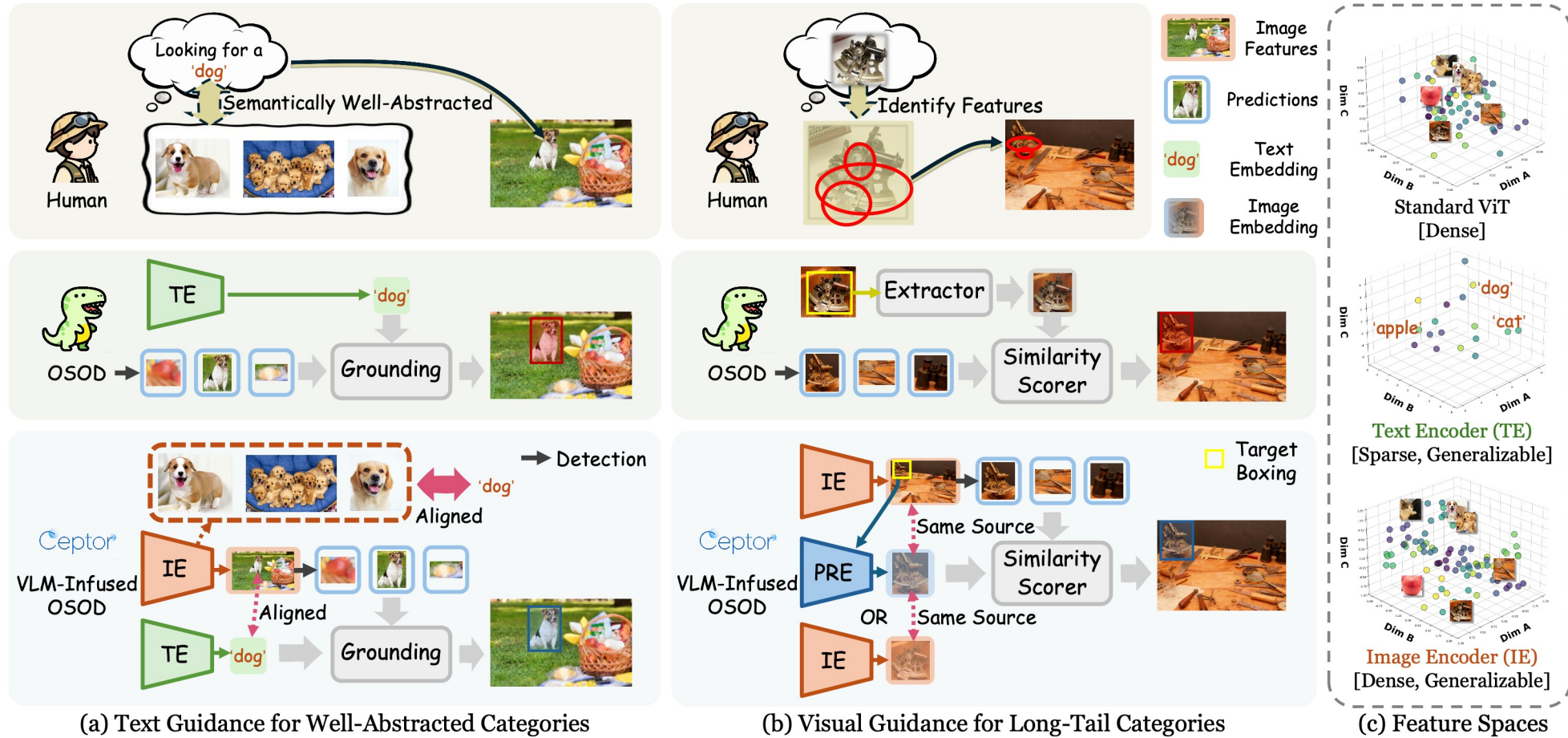
- ✗ point/box lack explicit category semantics;
- ✗ cropped regions lose contextual information;
- ✗ non-unified prompts lead to inefficient training;



Introduction

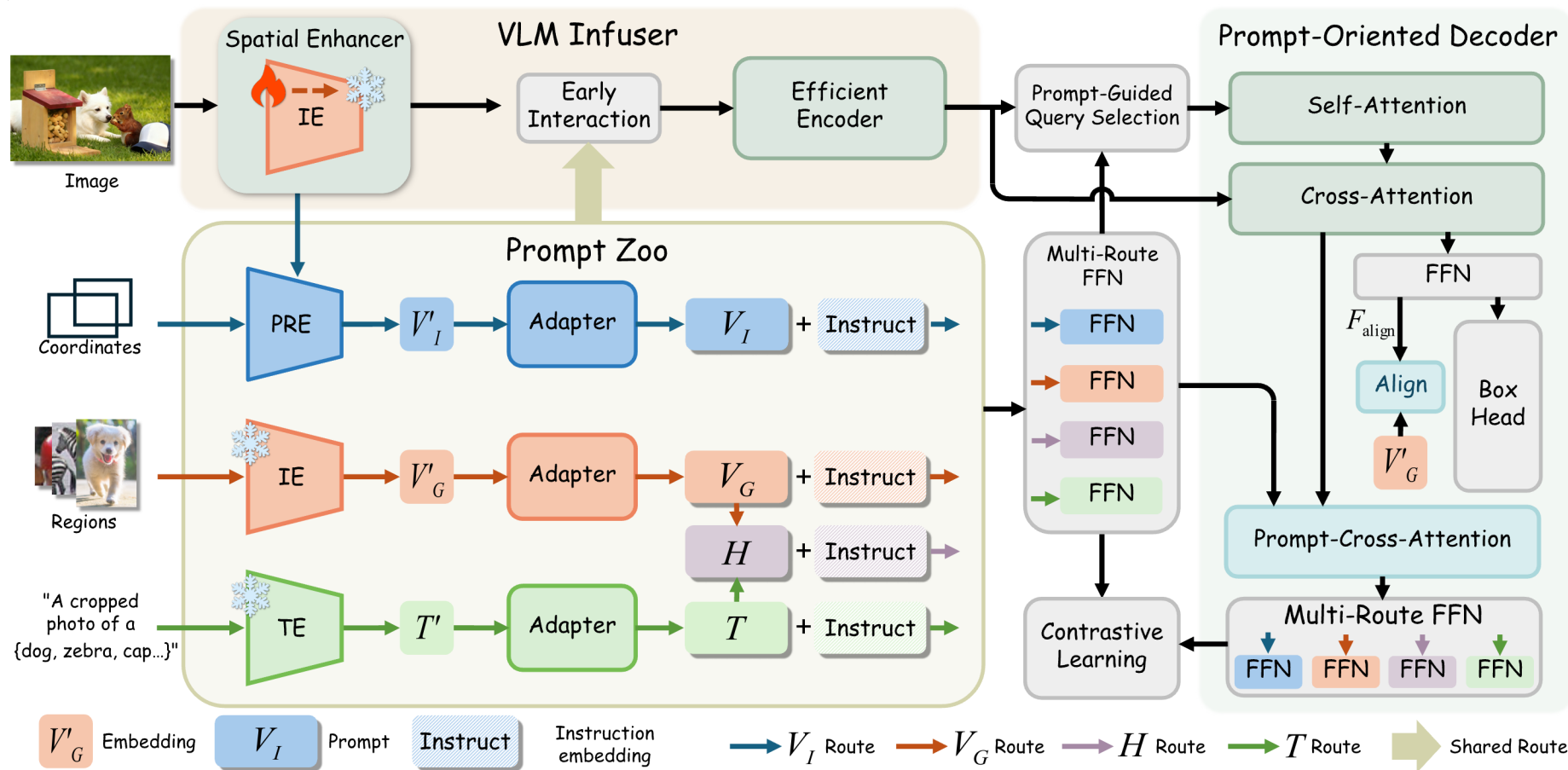
Inspired by Human Visual Perception

The human perceptual system seamlessly integrates multi-cues, including language instructions and visual exemplars, to identify objects across all abstraction levels.



Our Method

Ceptor:



Ceptor is Built on Deformable DETR with four core modules: **VLM Infuser** (transfers VLM generalization), **Prompt Zoo** (generates diverse prompts), **Harmonization Strategy**, and **Prompt-Oriented Decoder** (open-set localization + classification).

Our Method

VLM Infuser

Spatial Enhancer employs a simple feature pyramid over IE features following a light conv stream:

$$F_I = \text{IE}(I)$$

$$F_{\text{cat}}^l = \text{Concat}(\text{SFP}(F_I)^l, C(I)^l), \forall l \in \{1, \dots, L\}$$

$$F_{\text{enh}}^l = \text{Norm}(\text{Conv}_{1 \times 1}(F_{\text{cat}}^l))$$

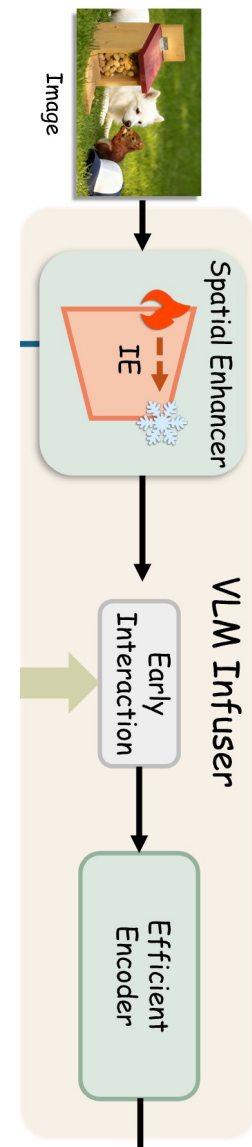
Early Interaction applies MHCA between the prompts and each level image features;

$$F_{\text{fused}}^l = \text{MHCA}(F_{\text{enh}}^l, P), \forall l \in \{1, \dots, L\}$$

$$F_{\text{top}} = \text{FFN}(\text{MHSA}(F_{\text{fused}}^L))$$

$$\{F_{\text{infuser}}^l\}_{l=1}^L = \mathcal{I}_{\text{cross-scale}}(\{F_{\text{fused}}^l\}_{l=1}^{L-1}, F_{\text{top}})$$

Efficient Encoder only enhance the top-level feature with MHSA and FFN;



Our Method

Prompt Zoo

construct four prompt types:

- ✓ **Interactive Visual Prompt**

box on the current image $\rightarrow$ PRE

- ✓ **Generic Visual Prompt**

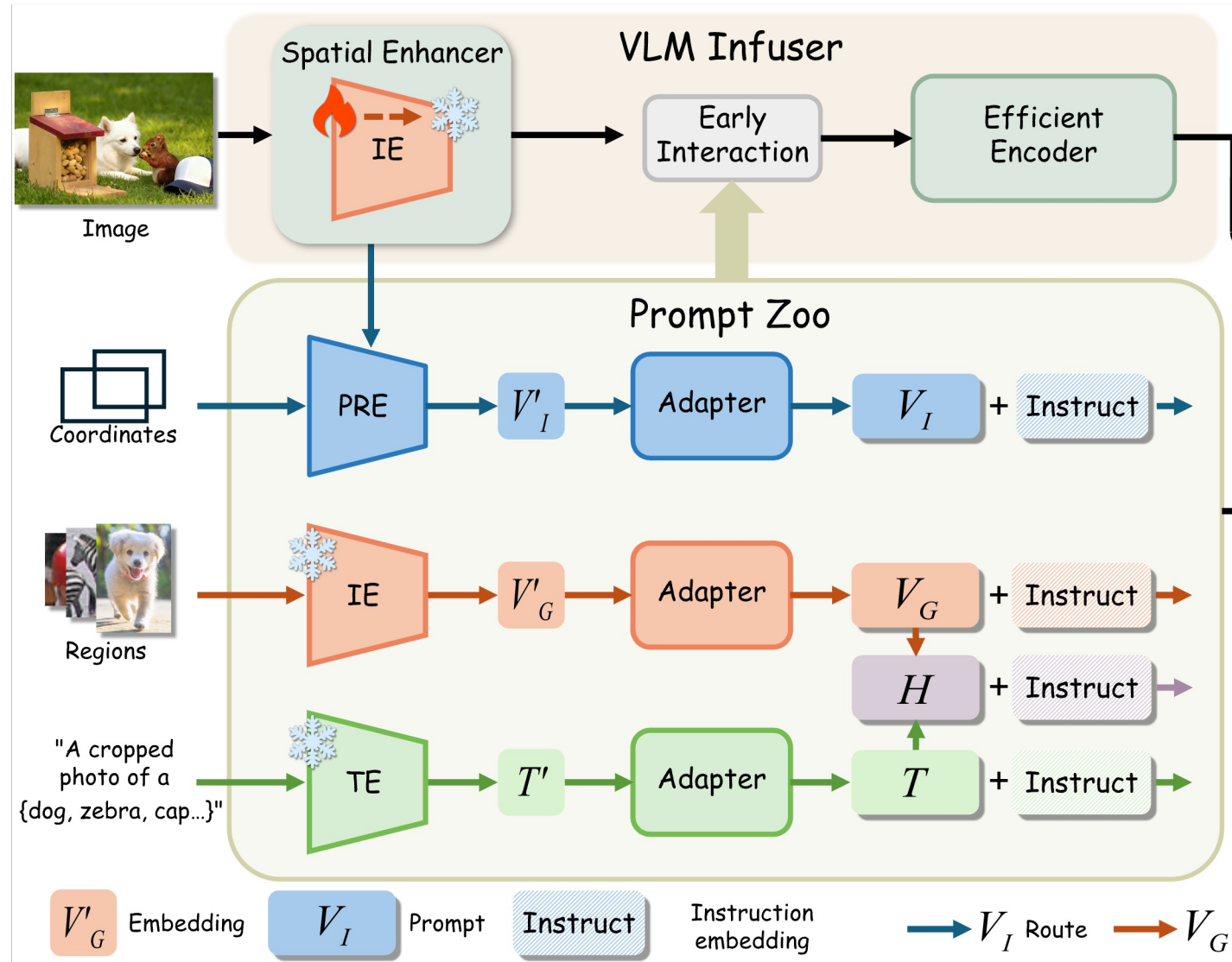
Frozen IE, N=16 crop regions, registered per class

- ✓ **Hybrid Prompt**

mean of text and generic visual prompt

- ✓ **Text Prompt**

Frozen TE, average 80 text prompt embeddings



Our Method

Harmonization Strategy

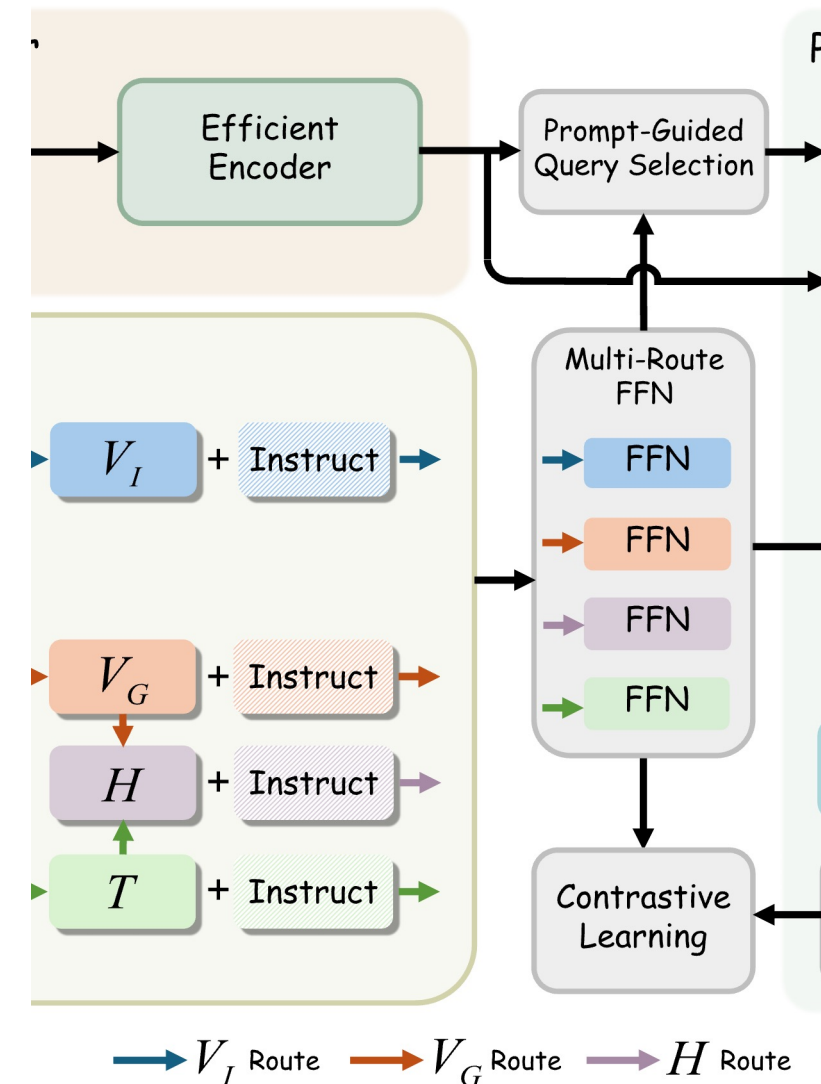
Eliminate multi-prompt conflicts for efficient joint optimization.

✓ Multi-Route FFNs

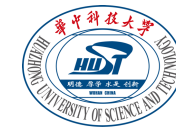
Four parallel, route-specific FFNs process individual prompt types.

✓ Learnable Instruction Embeddings

Four independent learnable embeddings act as auxiliary prompts.



Our Method



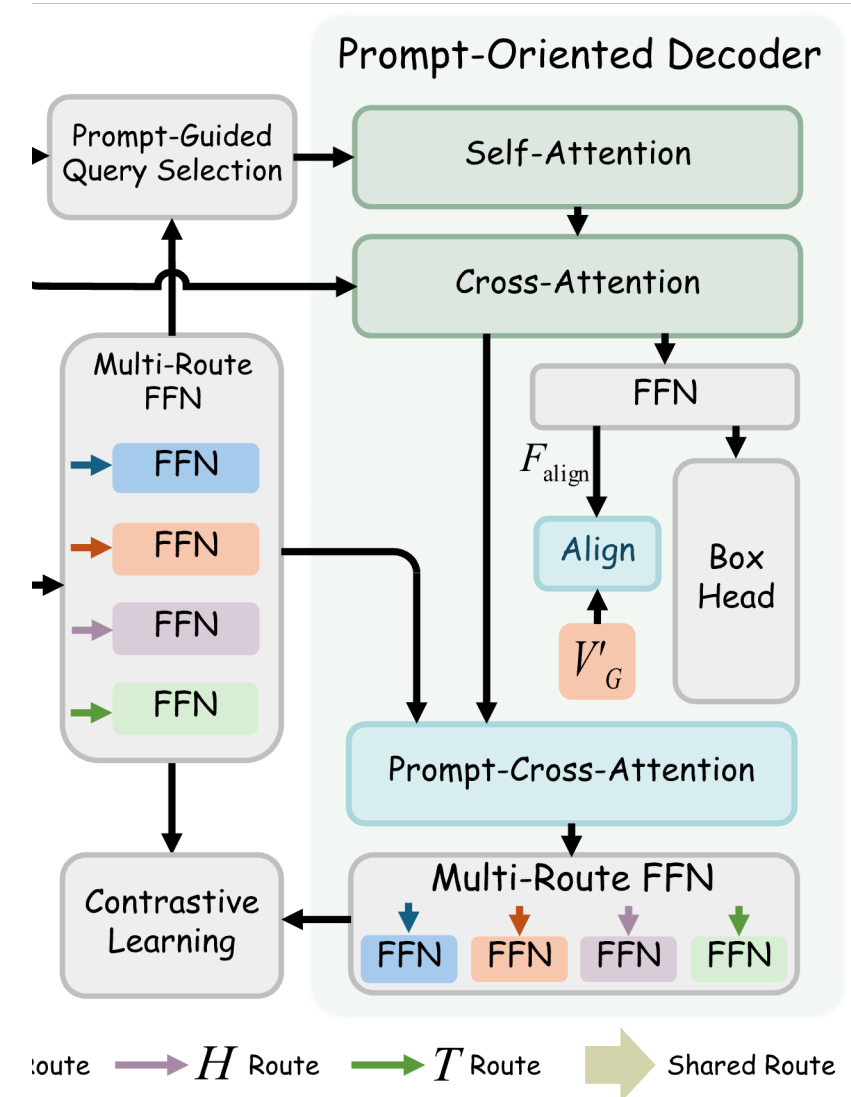
Prompt-Oriented Decoder

Alignment Mechanism enforces feature alignment between decoder predictions via DETR bipartite matching and frozen IE visual embeddings of the true classes.

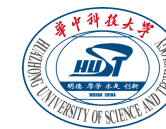
$$\mathcal{L}_{\text{align}} = \frac{1}{n \cdot d} \sum_{i=1}^n \sum_{j=1}^d (F_{i,j,\text{align}} - V'_{i,j,G})^2$$

Overall Loss Function

$$\mathcal{L} = \underbrace{\lambda_{\text{cls}} \mathcal{L}_{\text{cls}}}_{\text{Contranstive Loss}} + \underbrace{\lambda_{\text{L1}} \mathcal{L}_{\text{L1}} + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}}_{\text{Box Regression}} + \underbrace{\lambda_{\text{align}} \mathcal{L}_{\text{align}}}_{\text{Alignment}}$$



Experiments



Comparisons with State-of-the-Arts

Methods	Backbone	Data	COCO-val	LVIS-minival	LVIS-minival-r
Visual-I					
T-Rex2	Swin-T	7.3M	56.6	59.3	64.4
T-Rex2	Swin-L	7.3M	58.5	62.5	70.1
Ceptor	PEcore S	3.2M	65.1	66.9	77.0
Ceptor	PEcore L	3.5M	71.9	72.8	81.3
Visual-G					
T-Rex2	Swin-T	7.3M	38.8	37.4	29.9
T-Rex2	Swin-L	7.3M	46.5	47.6	45.4
YOLOE-11-S	—	1.4M	—	26.3	22.5
YOLOE-11-L	—	1.4M	—	33.7	28.1
Ceptor	PEcore S	3.2M	47.3	39.1	37.0
Ceptor	PEcore L	3.5M	54.0	47.8	47.2

Ceptor consistently outperforms T-Rex2 **across visual prompt types**, with comparable parameters and less than half training data.

Methods	Backbone	Data	COCO-val	LVIS-minival	LVIS-minival-r
Text					
Grounding DINO	Swin-T	5.4M	48.4	27.4	18.1
OV-DINO	Swin-T	2.4M	50.2	40.1	34.5
LLMDet	Swin-T	1.1M	55.6*	44.7	37.3
T-Rex2	Swin-T	7.3M	45.8	42.8	37.4
Ceptor	PEcore S	3.2M	51.7	43.1	40.6
LLMDet	Swin-L	1.1M	—	51.1	45.1
T-Rex2	Swin-L	7.3M	52.2	54.9	49.2
Ceptor	PEcore L	3.5M	54.8	54.0	54.7
Hybrid					
T-Rex2	Swin-T	7.3M	42.5	—	—
Ceptor	PEcore S	3.2M	50.5	44.1	40.9

Ceptor achieves **stronger open-vocabulary detection** with **text prompts**, and outperforms T-Rex2 by +8.0 AP on COCO using **hybrid prompts**.

Ablation Studies

Model	Visual-I	Visual-G	Text
Ceptor	72.4	25.6	32.2
w/o Infusion	66.7	24.9	30.4
w/o Harmonization	69.5	24.1	31.1
w/o VLM	59.7	21.4	30.5
w/o PRE (w/cross-attn)	70.3	—	—

Note: Metrics in AP_r . Each component provides cumulative gains, with the largest margins on rare categories.

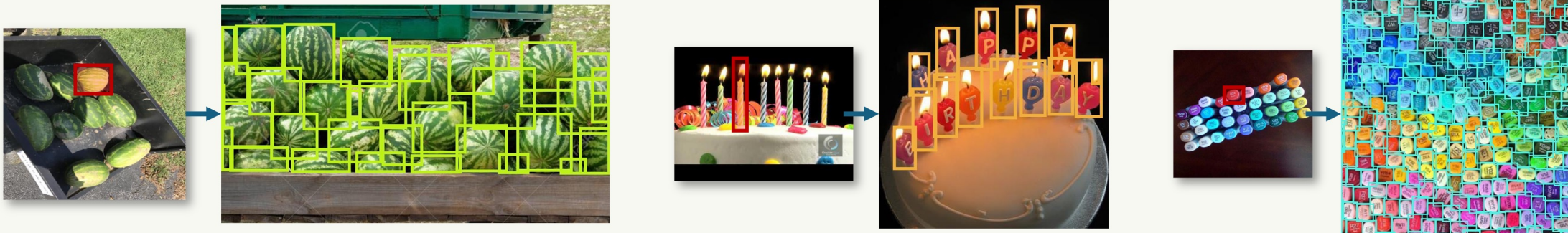
Ceptor achieves the highest performance across all prompt types; removing the VLM causing the largest drop, and Infusion, Harmonization, and PRE each providing crucial, cumulative accuracy gains.

Experiments

Qualitative Comparisons

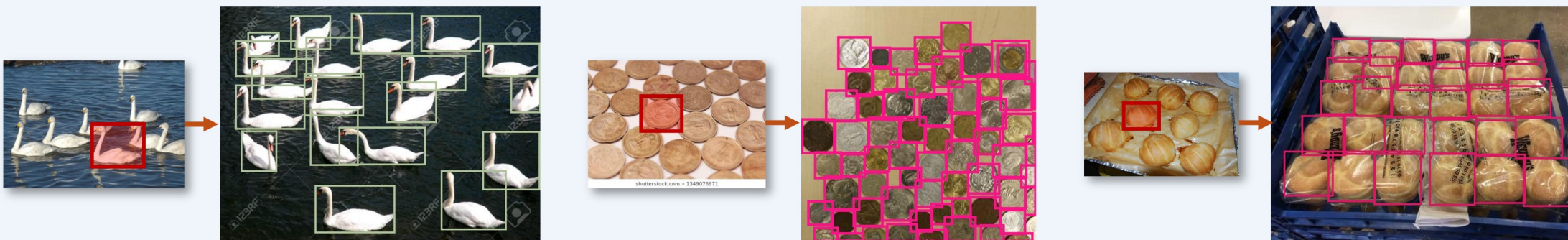
Single-Template Cross-Image Open-Set Detection (Interactive Visual Prompt):

→ : Prompt Generated by PRE



Single-Template Cross-Image Open-Set Detection (Generic Visual Prompt):

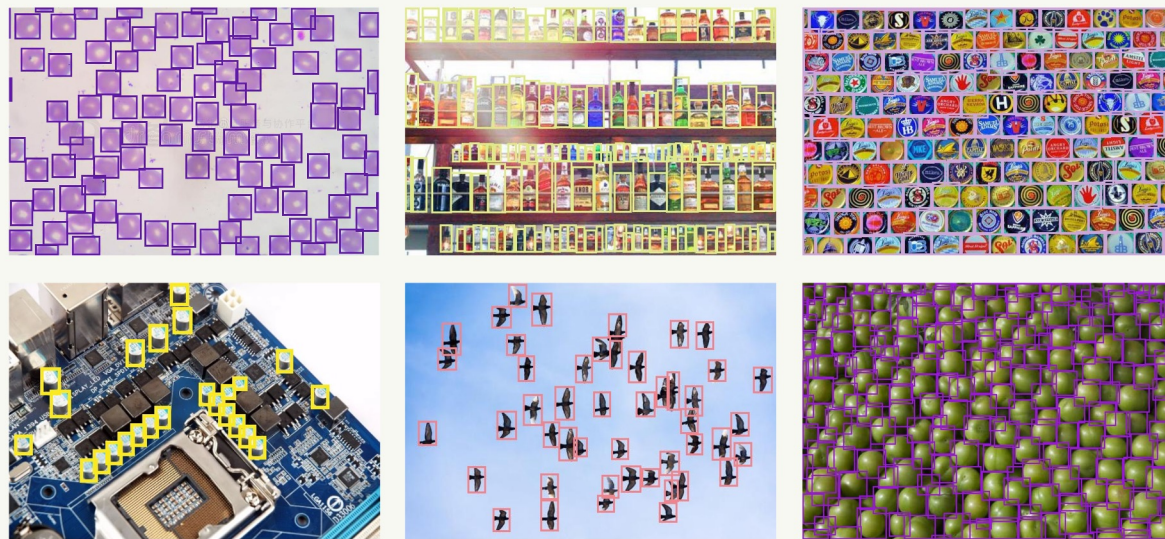
→ : Prompt Generated by IE



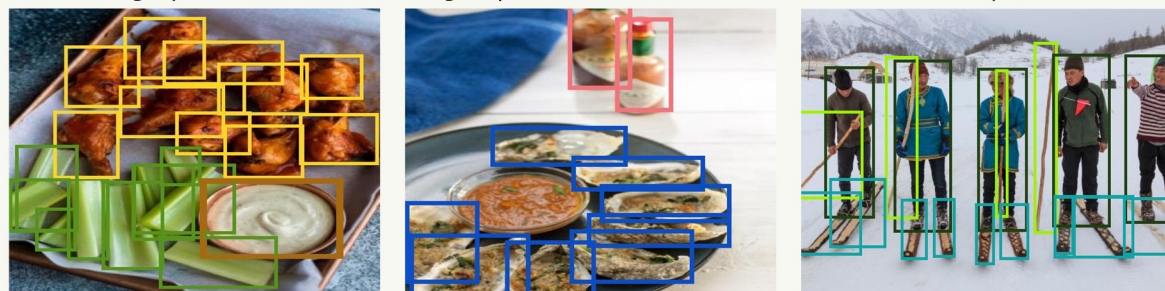
Experiments

Qualitative Comparisons

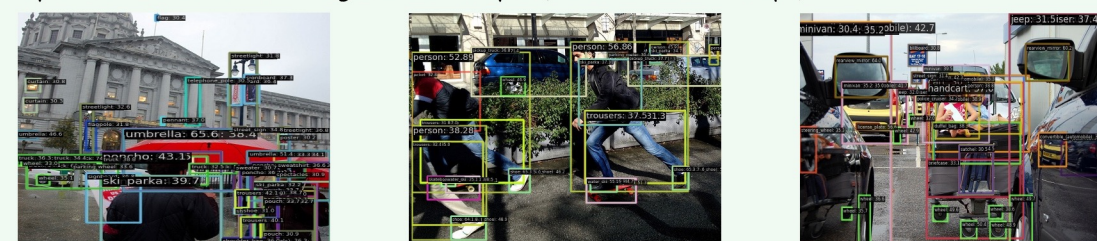
Single-Category Interactive Intra-Image Open-Set Detection (Interactive Visual Prompt):



Multi-Category Interactive Intra-Image Open-Set Detection (Interactive Visual Prompt):



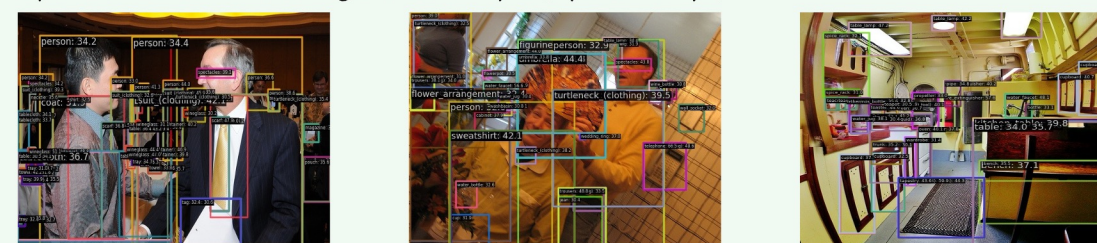
Open-Set Detection with Registered Prompts (Generic Visual Prompt):



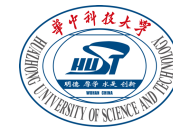
Open-Set Detection with Registered Prompts (Text Prompt):



Open-Set Detection with Registered Prompts (Hybrid Prompt):



- We propose a unified and universal open-set detector that uses diverse prompts, inspired by human visual search mechanisms;
- We propose a VLM Infuser that transfers pre-trained VLM generalization into open-set detection;
- We design a Pyramid RoI Extractor to enrich semantic, contextual, and spatial representations, paired with Harmonization Strategy for joint multi-route optimization.



Tencent
YouTu Lab

Thanks!