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

Jinyang Li[1★] Bin-Bin Gao[2★]

Weifu Fu[2] Jingnan Luo[3] Hanqiu Deng[2] Yue Guo[4] Jun Liu[2] Yong Liu[2] Chengjie Wang[2] Wenbing Tao[1†]

[1]Huazhong University of Science and Technology [2]YouTu Lab, Tencent [3]Southern University of Science and Technology [4]Macau University of Science and Technology

ECCV 2026 | MALMÖ | SEPT 8–12

Code & Models

1. Motivation

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

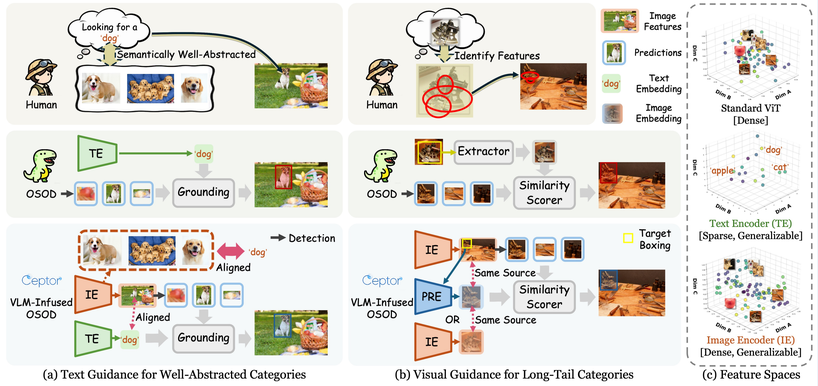
- ✗ long-tail categories;
- ✗ domain-specific entities;
- ✗ hard-to-describe instances;

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;

Our Ceptor: Inspired by Human Visual Perception

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



(a) Text Guidance for Well-Abstracted Categories (b) Visual Guidance for Long-Tail Categories (c) Feature Spaces

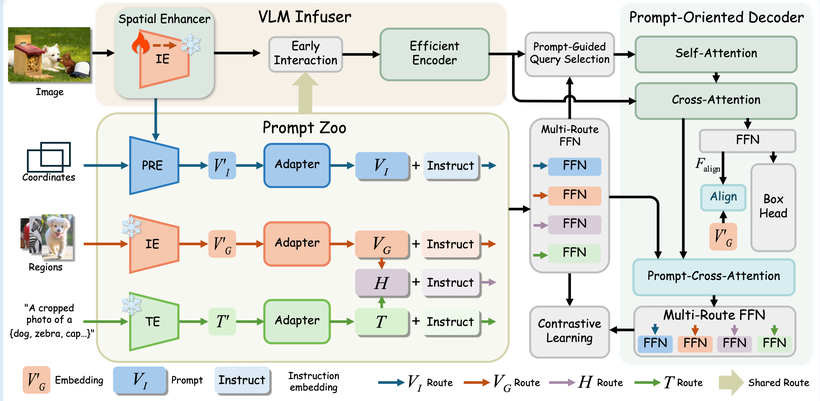
(a, b) Ceptor uses a VLM as *both* image backbone and prompt generator. (c) The **image encoder (IE)** space is denser and more generalizable than a plain ViT, and than the **text encoder (TE)** space, which is sparse because language is discrete.

Inspired by human visual perception, Ceptor adopts a unified VLM serving as both **vision extractor** and **prompt generator**, achieving **strong open-set detection performance with diverse prompts**, such as text, vision and hybrid prompts.

2. Contributions

- ✓ **Human-Inspired Detect-Anything Model:** propose a unified and universal open-set detector capable of handling diverse prompts, inspired by human visual perception mechanisms;
- ✓ **VLM Feature Infusion Strategy:** transfer pre-trained VLM generalization into open-set detection, ensuring intrinsic alignment between image and text prompt features;
- ✓ **Pyramid RoI Extractor & Prompt Harmonization:** design a Pyramid RoI Extractor to enrich semantic, contextual, and spatial representations, paired with Harmonization Strategies for joint multi-route optimization.

3. Our Method: Ceptor Framework



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

3.1 VLM Infuser

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

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

$$F^l_{\mathrm{cat}} = \mathrm{Concat}(\mathrm{SFP}(F_I)^l, C(I)^l), \forall l \in \{1, \cdots, L\}$$

$$F^l_{\mathrm{enh}} = \mathrm{Norm}(\mathrm{Conv}_{1\times1}(F^l_{\mathrm{cat}}))$$

Early Interaction applies MHCA between the prompts and each level image features; **Efficient Encoder** only enhance the top-level feature with MHSA and FFN; A cross-scale fusion module to strengthen category information across scales.

$$F^l_{\mathrm{fused}} = \mathrm{MHCA}(F^l_{\mathrm{enh}}, P), \forall l \in \{1, \cdots, L\}$$

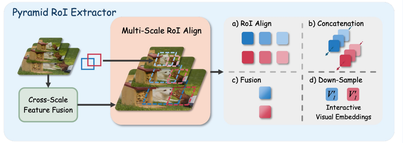
$$F_{\mathrm{top}} = \mathrm{FFN}(\mathrm{MHSA}(F^L_{\mathrm{fused}}))$$

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

3.2 Prompt Zoo

Single mode, single prompt type per forward pass.

V_I Interactive Visual Prompt	V_G Generic Visual Prompt
box on the current image → PRE	Frozen IE, N=16 crop regions, registered per class
T **Text Prompt**	H **Hybrid Prompt**
Frozen TE, average 80 text prompt embeddings	mean of V_G and T



We firstly use cross-scale feature fusion to smooth ViT features, and then perform **Multi-Scale RoI Align** to capture boxes of any position and scale.

3.3 Harmonization Strategies

Eliminate multi-prompt conflicts for efficient joint optimization via:

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

Instruction Embeddings: Four independent learnable embeddings act as auxiliary prompts.

3.4 Prompt-Oriented Decoder

Alignment Mechanism enforces feature alignment between decoder predictions (F_{align}) via DETR bipartite matching and frozen IE visual embeddings (V'_{G}) of the true classes.

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

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

4. Experiments

4.1 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 (+8.5)	62.5 (+7.6)	70.1 (+12.6)
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 (+7.1)
YOLOE-11-S	–	1.4M	– (+8.5)	26.3 (+1.7)	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 two 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 (+5.9)	42.8 (+0.3)	37.4 (+3.2)
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 (+2.6)	**54.9** (-0.9)	49.2 (+5.5)
Ceptor	**PEcore L**	**3.5M**	**54.8**	54.0	**54.7**
Hybrid					
T-Rex2	Swin-T	7.3M	42.5 (+8.0)	–	–
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.**

Method	Backbone	FPS	Mem.
Grounding DINO	Swin-T	11.0	5539
Ceptor	**PEcore S**	**12.4**	**3684**

Compared to Grounding DINO (Swin-T), Ceptor (PEcore-S) achieves:

- ✓ Faster Speed: **12.4 FPS** vs. 11.0 FPS
- ✓ Lower Memory: **3684 MB** vs. 5539 MB

4.2 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.

4.3 Qualitative Comparisons