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], and 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
{jinyangli, wenbingtao}@hust.edu.cn, csgaobb@gmail.com,
{ryanwfu, choasliu, jasoncjwang}@tencent.com

Abstract. Open-set object detection leverages language prompts to guide the perception of categories outside the training set. However, for categories that are difficult to describe or cannot be effectively abstracted semantically, models exhibit reduced receptiveness to guidance provided by textual descriptions. Recent approaches have explored visual prompting, but they often lack inherent generalization capabilities and suffer from limitations in semantic, spatial, and contextual awareness during prompt construction. Additionally, isolated prompt pathways hinder unified optimization. Inspired by the general process of human object search, we designed ***Ceptor***, a unified detector guided by diverse prompts for open-set object detection. Specifically, we propose **Infusion Strategies** to infuse our framework with the generalization capabilities of a vision-language model (VLM), leveraging its alignment properties and generalizable feature space to construct diverse prompts for different object types and varying category distributions. Furthermore, we introduce a novel prompt generation method, **PRE**, which effectively preserves generalization, contextual information, and spatial accuracy. **Harmonization strategies** are also incorporated to ensure coordinated optimization across various prompts. Ceptor achieves strong performance with various prompts across multiple datasets, demonstrating the effectiveness of our methodology. Models and code are released at `https://github.com/jinyanglii/Ceptor`.

Keywords: Open-Set Object Detection · Visual Prompt · Multimodal

⋆ Equal Contribution. This work was done when J. Li was an intern at YouTu Lab, Tencent, advised by B.-B. Gao.
† Corresponding Author

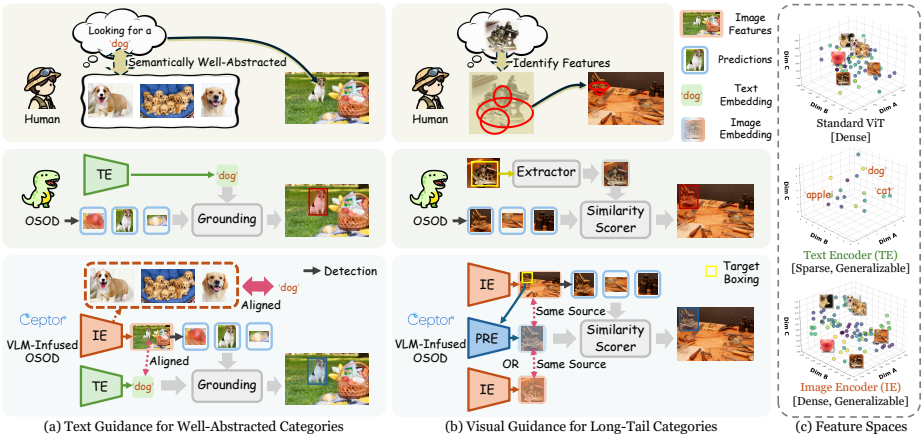

Fig. 1: Guided detection processes. (a),(b) Ceptor leverages a VLM as the image backbone and as the prompt generator. The VLM's generalization and vision–language alignment, together with the shared feature space between visual prompts and image features, enable effective object search. (c) Compared with a standard ViT [13] with limited generalization and the **TE** whose feature space can be sparse due to language discreteness, the **IE** feature space is denser and more generalizable for recognition.

1 Introduction

Open-set object detection (OSOD) addresses the primary generalization limitation of conventional object detection, unlocking potential applications in critical domains such as autonomous driving, industrial inspection, and medical imaging, and serving as vision experts for multimodal large language models [46].

Nevertheless, these open-vocabulary methods [17, 25, 29, 37, 47–49] tend to perform well primarily on categories whose text and image embeddings are well aligned in the pre-trained vision–language embedding space. However, their performance degrades significantly on objects with low semantic abstraction [20]. Such objects include long-tail categories underrepresented in vision–language pre-training, domain-specific entities, or instances that are inherently difficult to describe textually. Under these circumstances, models become less responsive to textual guidance.

Recently, methods such as T-Rex2 [20] have explored using coordinates to generate visual prompts for object localization to address these challenges. However, these approaches [7, 20, 42, 51] suffer from several limitations. (1) They intrinsically lack generalizability in both the perception network and the visual prompt design. (2) Exemplar-based coordinate prompting does not convey explicit category semantics, and instance-level feature extraction is inherently constrained. Meanwhile, methods relying on feature extraction from cropped regions often sacrifice contextual information while increasing computational cost. (3) In addition, text and visual prompting routes are not optimized within a unified network, leading to inefficient training. They also lack intermediate prompting mechanisms to adapt to different category distributions.

In contrast, prior studies [12, 33, 40] suggest that *humans, equipped with a powerful perceptual system, integrate multiple cues during visual search, including language-based instructions and visual templates*, a process closely aligned with our task. Specifically, as illustrated in Fig. 1, when searching for semantically familiar categories, humans can rely on the target name to retrieve highly abstracted semantic concepts in the brain for recognition [11,33,39,45]. Furthermore, humans are able to localize targets by identifying objects that are visually similar to a given exemplar [3, 5]. This capability is supported by the brain's high-dimensional signal space, which enables strongly generalizable feature extraction and categorization mechanisms [3, 12, 33, 40], allowing diverse features to be efficiently recognized.

These findings inspire our design, as illustrated in Fig. 1(a),(b). (1) For detection, we leverage the generalizable feature spaces of a contrastively pre-trained vision–language model (VLM), analogous to those of the human brain, as illustrated in Fig. 1(c). In particular, the dense feature space of the image encoder (IE) enables effective feature mapping even for long-tail objects. (2) We exploit the alignment between image and text embeddings from the VLM to guide object detection using text prompts. (3) We build visual prompts from VLM image features and leverage the shared representational origin between visual prompts and image features to facilitate detection, especially for long-tail categories.

To this end, we employ a VLM that serves both as the image backbone and as multiple prompt generators. On this basis, we propose **Infusion Strategies**, which enhance spatial perception and, through interaction and alignment mechanisms, transfer the generalization capability of the VLM to the detector. Consequently, image and prompt features reside in generalizable feature spaces while being mutually aligned or derived from the same source, thereby facilitating effective interaction and object localization.

We use the text encoder (TE) and image encoder (IE) of the VLM as prompt generators to produce four types of prompts: **text, interactive visual, generic visual, and hybrid prompts**. Text prompts generated by the TE provide high-level semantic guidance. For the interactive visual prompts, we design a **Pyramid RoI Extractor (PRE)** that leverages spatially enhanced image features to extract instance-level target representations based on query coordinates while preserving contextual information. For the generic visual prompt, we employ the IE to directly extract visual representations. In addition, the hybrid prompt is formed by combining text and generic visual prompts to provide complementary information. The interactive visual prompt supports interactive detection within the current image by using examples specified by boxes. The other three prompts allow users to register either textual descriptions or image regions of a target category, which can subsequently guide object detection in any image.

To optimize these prompt guidance approaches simultaneously within a unified network, we adopt **Harmonization Strategies** for multi-route guidance.

We propose ***Ceptor***, a unified and universal open-set detector with VLM infusion and diverse prompt guidance. Experimental results on the COCO [27]

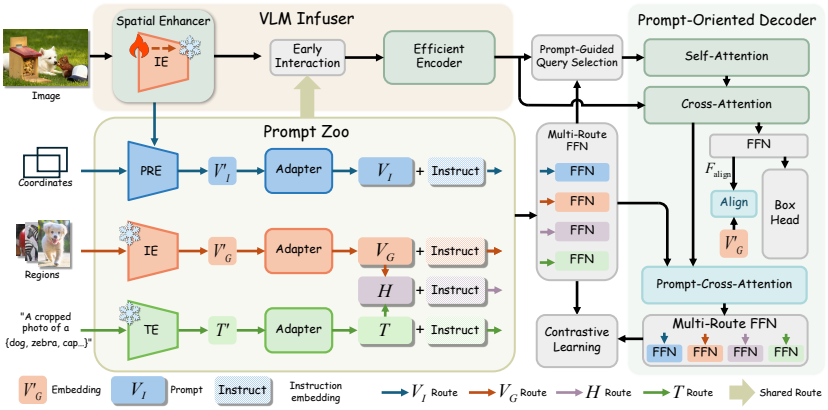

Fig. 2: Overview of Ceptor. Our model builds on Deformable DETR and has three core components: the VLM Infuser, the Prompt Zoo, and the Prompt-Oriented Decoder. These modules transfer VLM generalization, generate diverse prompts, and perform open-set localization and classification, respectively. Here, V_I, V_G, H, and T denote interactive visual, generic visual, hybrid, and text prompts.

and LVIS [18] datasets demonstrate that Ceptor outperforms state-of-the-art methods. Our main contributions are as follows:

- We propose a unified and diverse prompt-guided detector, a robust and universal detect-anything model inspired by the human object search process.
- We propose Infusion Strategies to transfer VLM capabilities for detection generalization. We ensure that image and prompt features are aligned or derived from the same source for effective guided interaction.
- We propose a Pyramid RoI Extractor to enhance the semantic, contextual, and spatial properties of prompt representations, and introduce Harmonization Strategies to jointly optimize multiple prompt routes.

2 Related Work

Text-Guided Open-Set Object Detection. The advancement of text-guided object detection has been driven by vision-language pre-training techniques. GLIP [25] established the foundation by reformulating object detection as a phrase grounding task. Grounding DINO [29,37] advanced this paradigm through a tightly-coupled fusion architecture that deeply integrates language and vision modalities. Building upon this foundation, LLMDet [14] incorporated large language models to generate detailed captions for enhanced semantic supervision. OV-DINO [43] introduced unified data integration to address noise in training data. OmDet-Turbo [52] optimized model efficiency for real-time performance, and Dynamic-DINO [31] employed dynamic Mixture-of-Experts inference. Besides, DetCLIP [47–49] developed a parallel concept learning approach with dictionary enrichment. Furthermore, object perception with text

prompts and long-tail categories has also been actively explored in segmentation [16, 35, 36], multi-object tracking [8–10, 24], and multimodal large language models (MLLMs) [2, 34]. The training strategies applicable to these domains are continuously evolving [15, 23].

Object Detection with Visual Prompts. Visual prompting plays a key role in open-set recognition, especially for objects that are difficult to describe textually or belong to long-tail categories. Early methods such as OV-DETR [50] used CLIP encoders to generate visual and text prompts, forming the basis of multimodal prompting. T-Rex2 [20] aligns text and visual prompts through contrastive learning, enabling unified support for multiple workflows and demonstrating strong zero-shot and long-tail-friendly open-set detection capability. Recent innovations include YOLOE's Semantic-Activated Visual Prompt Encoder [42], which utilizes decoupled branches to enhance semantic representation while maintaining computational efficiency. These works show the evolution of visual prompts into powerful mechanisms for open-set detection.

3 Ceptor

3.1 Overview

As shown in Fig. 2, our model builds upon Deformable DETR [54] and contains three core modules: a *VLM Infuser* for enhancing the transfer of the VLM's generalization capability, a *Prompt Zoo* for generating diverse prompts, and a *Prompt-Oriented Decoder (POD)* for performing open-set localization and classification. During forward propagation, different types of input cues are converted into corresponding prompts (**text prompts T, interactive visual prompts V_I, generic visual prompts V_G, and hybrid prompts H**) in the Prompt Zoo, as indicated by arrows of different colors. The input image is processed by the VLM Infuser to obtain enhanced image features, which are then used in POD to update queries. These queries interact with prompts processed by the Multi-Route Feed-Forward Network (Multi-Route FFN), thereby producing the final predictions. Additionally, alignment and Harmonization Strategies are incorporated to further improve model performance.

3.2 Infusion Strategies for Transferring VLM Generalization

VLM Infuser. We employ a VLM as the image backbone and prompt generator, bringing its generalization ability into the model. However, the plain ViT backbones of VLMs maintain single-scale representations and lack explicit spatial inductive biases, making them less suitable for Deformable DETR [54]-based methods that rely on multi-scale features for effective object detection [26, 32]. To address this, in the *Spatial Enhancer* shown in Fig. 2, the input image I is fed into the image encoder (IE), followed by a simple feature pyramid [26], denoted as SFP. In parallel, I is also processed by a set of lightweight convolutional networks C to extract auxiliary features. The two multi-scale feature streams are

fused at each level via concatenation (Concat), followed by a 1×1 convolution ($\text{Conv}_{1\times1}$) and a normalization layer (Norm). The formulation is as follows:

$$F_I = \text{IE}(I) \tag{1}$$

$$F_{\text{cat}}^l = \text{Concat}(\text{SFP}(F_I)^l, C(I)^l), \quad \forall l \in \{1, \cdots, L\} \tag{2}$$

$$F_{\text{enh}}^l = \text{Norm}(\text{Conv}_{1\times1}(F_{\text{cat}}^l)) \tag{3}$$

where F_I denotes the image features extracted by the image encoder E_I, F_{cat}^l denotes the concatenated features at level l, F_{enh}^l denotes the enhanced features produced by the Spatial Enhancer, and L is the number of feature levels.

For effective interaction, inspired by Grounding DINO [29], we perform *Early Interaction* by applying multi-head cross-attention (MHCA) between the prompts $P = \{V_I, V_G, H, T\}$ and the enhanced image features $\{F_{\text{enh}}^l\}_{l=1}^L$. In the *Efficient Encoder*, since the multi-scale features are derived from the same image features F_I produced by IE, we only enhance the top-level feature F_{fused}^L using multi-head self-attention (MHSA) and an FFN, which also reduces inference cost. Additionally, we adopt the cross-scale feature fusion module (CCFM) from RT-DETR [53], denoted as $\mathcal{I}_{\text{cross-scale}}$, to further strengthen category information across scales. These processes are formulated as follows:

$$F_{\text{fused}}^l = \text{MHCA}(F_{\text{enh}}^l, P), \quad \forall l \in \{1, \cdots, L\} \tag{4}$$

$$F_{\text{top}} = \text{FFN}\big(\text{MHSA}(F_{\text{fused}}^L)\big) \tag{5}$$

$$\{F_{\text{infuser}}^l\}_{l=1}^L = \mathcal{I}_{\text{cross-scale}}\big(\{F_{\text{fused}}^l\}_{l=1}^{L-1}, F_{\text{top}}\big) \tag{6}$$

where F_{infuser}^l denotes the output features of the VLM Infuser, F_{fused}^l and F_{top} are intermediate features, and L is the number of feature levels.

Aligned or Same-Source Representation Interaction. The VLM serves as both the image backbone and prompt generator, providing global features that ensure latent *consistency*, where image and prompt features are aligned or derived from the same source. The purpose of this consistency is to allow image features to focus more effectively on prompts not only during *Early Interaction* but also during the *Prompt-Cross-Attention* in the *Prompt-Oriented Decoder*, enabling more meaningful interactions. Through these dual interactions, we aim to inject the strong generalization capability of the VLM into the POD.

Alignment Mechanism. The VLM serves as the primary feature source on the input side of the model. To preserve Ceptor's generalization ability near the output side and enable better responses to prompt guidance, we transfer the generalizable knowledge of the VLM into the internal structure of Ceptor.

Specifically, we utilize generic visual embeddings $V_G' \in \mathbb{R}^{n\times d}$ extracted from the frozen image encoder (IE) using image regions. These embeddings correspond to the ground-truth categories of the current image I. Meanwhile, we select the instance-level features F_{align}, which are the *FFN* outputs of the *Prompt-Oriented Decoder* in Fig. 2, as the features to be aligned. Following the optimal bipartite matching in DETR between predicted objects and ground-truth objects [6], we obtain the matching pairs and align the predicted features F_{align} with the

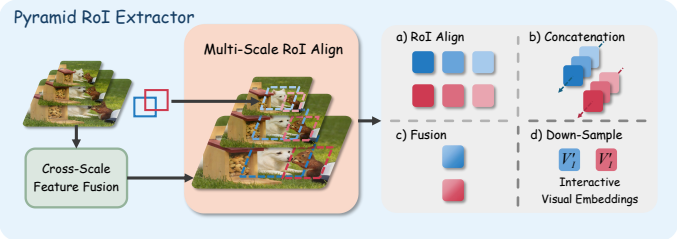

Fig. 3: Structure of the Pyramid RoI Extractor. We use cross-scale feature fusion to smooth ViT features, multi-scale RoI Align, and multi-scale fusion of region features to ensure effective extraction of contextual and instance-level information, as well as a dedicated downsampling network to preserve spatial information.

corresponding embeddings V'_G based on the matching results. Each feature is represented as a d-dimensional vector, where n denotes the number of ground-truth objects. The alignment loss $\mathcal{L}_{\text{align}}$ is defined as follows:

$$\mathcal{L}_{\text{align}} = \frac{1}{n \cdot d} \sum_{i=1}^{n} \sum_{j=1}^{d} \left(F_{i,j,\text{align}} - V'_{i,j,\text{G}} \right)^2 . \tag{7}$$

This alignment mechanism aims to distill the VLM's generalization capability into the model from the output side, helping the backbone IE retain its generalization ability. Moreover, it is designed to implicitly align outputs under different prompt guidance, making optimization more harmonious from the output side.

3.3 Diverse Guidance

As shown in Fig. 2, we construct four types of prompts in the *Prompt Zoo*, namely **text prompts, interactive visual prompts, generic visual prompts, and hybrid prompts**, denoted as T, V_I, V_G, and H, respectively, using different generation strategies. All prompts are derived from the VLM, which endows the prompt generator with inherent generalization capability and ensures that the prompts share a common origin with the image features, thereby facilitating object search. During each forward pass, the model operates in a single mode and is guided by only one type of prompt.

Interactive Visual Guidance. To enable interactive detection via bounding box selection while preserving contextual information, ensuring accurate instance-level feature extraction, and maintaining computational efficiency, we design a *Pyramid RoI Extractor (PRE)*, as shown in Fig. 3.

Specifically, we first apply a lightweight cross-scale feature fusion module (CCFM) to smooth the ViT [13]-extracted patch-level features $\{F^l_{\text{enh}}\}^L_{l=1}$, obtaining the feature pyramid $\{F^m_{\text{pre}}\}^M_{m=1}$, where M denotes the total number of pyramid levels. This process ensures that bounding boxes of arbitrary positions and scales, after being mapped onto $\{F^m_{\text{pre}}\}^M_{m=1}$, can accurately capture complete instance-level features. For each target specified by the mapped bounding

box, we then perform RoI Align [19] on each level F^m_{pre}. These region-level features inherently retain rich contextual information due to the global attention mechanism of ViT. Finally, the multi-scale region-level features corresponding to each query box are concatenated along the embedding dimension and fused via a 1×1 convolution. The fused features are subsequently downsampled through multi-layer convolutions to preserve spatial structure, producing interactive visual embeddings V'_I for prompt generation.

Generic Visual Guidance. VLM image embeddings are aligned with text embeddings T', and thus encode higher-level and more global semantics than V'_I. Accordingly, we use the IE to construct generic visual prompts V_G for cross-image generalizable search, enabling the model to search for objects of a given category across a large number of images.

Specifically, we employ the frozen IE to generate V_G. For category y, we randomly sample N images and crop regions containing objects belonging to category y. These cropped regions are then fed into the IE to extract features offline, which are subsequently aggregated into a single representation to obtain V_G. For each category, we repeat this process to obtain k aggregated features. During network forward propagation, for category y, only one aggregated feature is randomly selected to guide detection. We design these two types of visual prompts to exploit their respective strengths, offering the model greater flexibility and richer combinational possibilities, and adapting to a wider range of scenarios.

Text Guidance. Following text-prompted detection approaches [29, 31, 43], we utilize high-level semantic information to guide object detection. Moreover, we leverage the intrinsic vision–language alignment of the VLM, ensuring that the text prompts are aligned with the output features of the image backbone.

Hybrid Guidance. We obtain the hybrid prompt by averaging and normalizing the generic visual prompt and the text prompt, allowing the two modalities to jointly interact with image features and complement each other.

All four types of prompts have dedicated optimization pathways, while sharing the core interaction networks and jointly guiding the model through a unified and efficient training process.

3.4 Harmonization Strategies

Due to the diversity of prompts and their deep interaction with intermediate features in the model, potential conflicts may arise, negatively affecting optimization efficiency. To better manage the guidance from the Prompt Zoo, as shown in Fig. 2, we propose *Harmonization Strategies*, including *Multi-Route Feed-Forward Networks (Multi-Route FFN)* and *instruction embeddings*.

Multi-Route FFN is composed of four parallel feed-forward networks, each corresponding to and processing one type of prompt guidance. It serves as a set of fixed experts, enabling tailored feature enhancement and providing a larger optimization margin at the network level. Instruction embeddings are four learnable independent embeddings. They serve as auxiliary prompts, aiming to inform the model which type of prompt it is currently interacting with.

Table 1: Comparison of visual prompt guided open-set detection under Visual-I and Visual-G settings. ✓ and ✗ denote satisfied and not satisfied, respectively.

Method	Open Source	Backbone	Data Scale	COCO-val	LVIS-minival				LVIS-val			
				AP	AP_{all}	AP_r	AP_c	AP_f	AP_{all}	AP_r	AP_c	AP_f
Visual-I												
T-Rex2 [20]	✗	Swin-T	7.3M	56.6	59.3	64.4	63.5	54.6	62.6	71.9	63.7	57.3
T-Rex2 [20]	✗	Swin-L	7.3M	58.5	62.5	70.1	66.1	57.9	65.8	72.6	67.3	61.2
Ceptor	✓	$PE_{core}S$	3.2M	65.1	66.9	77.0	72.8	59.8	63.0	74.6	65.5	55.0
Ceptor	✓	$PE_{core}L$	3.5M	71.9	72.8	81.3	78.8	65.9	68.8	80.2	70.9	61.5
Visual-G												
T-Rex2 [20]	✗	Swin-T	7.3M	38.8	37.4	29.9	33.9	41.8	34.9	32.4	30.3	41.1
T-Rex2 [20]	✗	Swin-L	7.3M	46.5	47.6	45.4	46.0	49.5	45.3	43.8	42.0	49.5
YOLOE-v8-S [42]	✓	-	1.4M	-	26.2	21.3	27.7	25.7	-	-	-	-
YOLOE-v8-L [42]	✓	-	1.4M	-	34.2	33.2	34.6	34.1	-	-	-	-
YOLOE-11-S [42]	✓	-	1.4M	-	26.3	22.5	27.1	26.4	-	-	-	-
YOLOE-11-L [42]	✓	-	1.4M	-	33.7	28.1	34.6	33.8	-	-	-	-
Ceptor	✓	$PE_{core}S$	3.2M	47.3	39.1	37.0	38.1	40.4	30.9	28.8	28.3	34.6
Ceptor	✓	$PE_{core}L$	3.5M	54.0	47.8	47.2	47.5	48.1	40.5	40.5	38.7	42.6

3.5 Optimization

We adopt the L1 loss and GIoU loss [38] for localization. For the classification loss $\mathcal{L}_{\text{cls}}$, following Grounding DINO [29], we employ a contrastive loss, which computes dot products between each output detection query and the prompt features to obtain logits, followed by the focal loss. The alignment loss $\mathcal{L}_{\text{align}}$ is incorporated to further facilitate optimization, as mentioned in Sec. 3.2.

The box regression and classification losses are initially used for Hungarian matching between predicted outputs and ground truths. The final loss $\mathcal{L}$ is computed between the ground truths and the corresponding matched predictions and is formulated as follows:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \tag{8}$$

where $\mathcal{L}_{\text{L1}}$ denotes the L1 loss, $\mathcal{L}_{\text{GIoU}}$ denotes the GIoU loss, and λ_{cls}, λ_{L1}, λ_{GIoU}, and λ_{align} are the weighting coefficients. Additionally, auxiliary losses [1] are applied after each decoder layer.

4 Experiments

4.1 Setup

Dataset. Ceptor enables joint training on both detection and grounding datasets. For text prompt-guided pre-training, we use the detection datasets Objects365 [41] and V3Det [44], together with the grounding dataset GoldG [21], totaling 1.56M labeled images. For training with all types of prompt-guided supervision, we utilize Objects365, OpenImages [22], V3Det, and HierText [30], resulting in a

Table 2: Comparison of text prompt guided open-set object detection. ✓ and ✗ denote satisfied and not satisfied, respectively. Gray numbers indicate including COCO data in training.

Method	Open Source	Backbone	Data Scale	COCO-Val	LVIS-minival				LVIS-val			
				AP	AP_{all}	AP_r	AP_c	AP_f	AP_{all}	AP_r	AP_c	AP_f
Text-Guided Open-Set Object Detection Models												
GLIP [25]	✓	Swin-T	5.4M	46.7	26.0	20.8	21.4	31.0	17.2	10.1	12.5	25.5
Grounding DINO [29]	✗	Swin-T	5.4M	48.4	27.4	18.1	23.3	32.7	-	-	-	-
OmDet-Turbo [52]	✗	Swin-T	1.4M	42.5	30.3	-	-	-	-	-	-	-
DetCLIP [48]	✗	Swin-T	2.4M	-	35.9	33.2	35.7	36.4	28.4	25.0	27.0	28.4
DetCLIPv2 [47]	✗	Swin-T	16.4M	-	40.4	36.0	41.7	40.4	32.8	31.0	31.7	34.8
DetCLIPv3 [49]	✗	Swin-T	51.6M	47.2	47.0	45.1	47.7	46.7	38.9	37.2	37.5	41.2
Grounding DINO 1.5 Edge [37]	✗	EfficientViT-L	20M	42.9	33.5	28.0	34.3	33.9	27.3	26.3	25.7	29.6
Grounding DINO 1.5 Pro [37]	✗	ViT-L	20M	54.3	55.7	56.1	57.5	54.1	47.6	44.6	47.9	48.7
OV-DINO [43]	✓	Swin-T	2.4M	50.2	40.1	34.5	39.5	41.5	32.9	29.1	30.4	37.4
LLMDet [14]	✓	Swin-T	1.1M	55.6	44.7	37.3	39.5	41.5	34.9	26.0	30.1	44.3
LLMDet [14]	✓	Swin-L	1.1M	-	51.1	45.1	46.1	56.6	42.0	31.6	38.8	50.2
Dynamic-DINO [31]	✗	EfficientViT-L	1.6M	46.2	36.2	41.9	39.9	31.9	29.6	35.4	29.2	27.3
Multi-Prompt Guided Open-Set Detection Model												
T-Rex2 [20]	✗	Swin-T	7.3M	45.8	42.8	37.4	39.7	46.5	34.8	29.0	31.5	41.2
T-Rex2 [20]	✗	Swin-L	7.3M	52.2	54.9	49.2	54.8	56.1	45.8	42.7	43.2	50.2
YOLOE-v8-S [42]	✓	-	1.4M	-	27.9	22.3	27.8	29.0	-	-	-	-
YOLOE-v8-L [42]	✓	-	1.4M	-	35.9	33.2	34.8	37.3	-	-	-	-
YOLOE-11-S [42]	✓	-	1.4M	-	27.5	21.4	26.8	29.3	-	-	-	-
YOLOE-11-L [42]	✓	-	1.4M	-	35.2	29.1	35.0	36.5	-	-	-	-
Ceptor	✓	$PE_{core}S$	3.2M	51.7	43.1	40.6	41.1	45.4	34.3	28.2	31.4	40.1
Ceptor	✓	$PE_{core}L$	3.5M	54.8	54.0	54.7	53.8	54.0	45.5	42.8	44.2	48.1

total of 2.5M labeled images. When using $PE_{core}L$ as the backbone, we additionally incorporate 0.3M training samples from LCS [14, 28]. *All datasets used are publicly available, and no additional manual annotation is required.*

Implementation Details. The VLM used in our method is the ViT-based Perception Encoder ($PE_{core}S$ / $PE_{core}L$) [4]. The training details for Ceptor with $PE_{core}S$ are as follows. We adopt AdamW as the optimizer, with both the initial learning rate and weight decay set to 1e-4. During pre-training, we use a batch size of 8 and train for 30 epochs. The learning rate is decayed by a factor of 10 at epochs 19 and 28. For full prompt-guided training, we train for 20 epochs, with the learning rate similarly decayed at epochs 15 and 19. Throughout training, the Image Encoder (IE) and Text Encoder (TE) used for prompt generation remain frozen. During the text prompt-guided pre-training stage, the IE serving as the image backbone is fine-tuned with a learning rate multiplier of 0.05 to accommodate the increased input resolution of 1024×1024. In the full prompt-guided training stage, detection is trained with all prompt types. Specifically, training iterations with generic visual prompts and text prompts are alternated. Additionally, interactive visual prompts and hybrid prompts are each introduced once every five alternating cycles. The training details for Ceptor with $PE_{core}L$ are provided in the supplementary material.

4.2 Evaluation

We evaluate open-set detection under diverse prompt guidance settings on two comprehensive benchmarks: COCO [27] and LVIS [18]. COCO contains 80 common categories and serves as a standard benchmark for object detection. LVIS comprises 1203 categories (including 337 rare categories), making it suitable for

Table 3: Comparison of hybrid prompt guided open-set object detection. ✓ and ✗ denote satisfied and not satisfied, respectively.

Method	Open Source	Data Scale	COCO-val	LVIS-minival				LVIS-val			
			AP	AP_{all}	AP_r	AP_c	AP_f	AP_{all}	AP_r	AP_c	AP_f
T-Rex2 [20]	✗	7.3M	42.5	-	-	-	-	37.0	34.3	33.8	41.7
Ceptor	✓	3.2M	50.5	44.1	40.9	42.5	46.1	35.9	31.4	33.5	40.6

assessing long-tail and zero-shot capabilities. For all benchmarks, Average Precision (AP) is used as the evaluation metric. The specific settings for evaluating different guidance modes are detailed below.

Interactive Visual Guidance Route (Visual-I). For this route, Ceptor performs interactive, visual prompt-guided detection based on provided bounding boxes. For each category within an image, we randomly select one ground truth bounding box as the interactive visual prompt, which is then converted into a prompt for that category. This approach to interactive object detection supports a wide range of application scenarios, including automatic annotation, object counting, and visual search.

Generic Visual Guidance Route (Visual-G). For this route, Ceptor uses the registered visual prompts to guide object detection across a large number of images. Specifically, for each category, we extract generic visual embeddings from the training set images of each benchmark. We first randomly sample N images for each category that contains at least one instance. For each sampled image, we crop the regions corresponding to all ground-truth bounding boxes of that category, enlarged by a factor of 1.2, and use these as input to the frozen image encoder (IE) to obtain the image embeddings. We then compute the average of these embeddings from N images for each category and normalize the result. These averaged image embeddings are used as generic visual prompts. To maintain consistency with prior work, N is set to 16.

Text Guidance Route (Text). For this route, we utilize all category names from the benchmark. For each category name, we pair it with 80 templates and feed them into the frozen text encoder. The resulting text embeddings are averaged and normalized to obtain the text prompt. For grounding datasets, descriptive phrases are used as input.

Hybrid Guidance Route (Hybrid). The hybrid guidance route simply averages the generic visual and text prompts, followed by normalization, to obtain a bimodal prompt. This prompt is processed in the same way as the other prompts.

4.3 Main Results

Visual Prompt Guided Detection. We compare our method with T-Rex2 [20] and YOLOE [42], as shown in Tab. 1. Ceptor generally outperforms these methods on COCO [27] and LVIS [18] under both types of visual prompt, even though it is trained with less data than T-Rex2. As shown in Table 1, under a comparable model scale, Ceptor with $PE_{core}S$ achieves significant advantages in the

Table 4: Comparison of open-set detection performance under diverse prompts. Underline denotes that the hybrid use of two prompts leads to improved performance.

Method	Mode	COCO-val	LVIS-minival				LVIS-val			
		AP	AP_{all}	AP_r	AP_c	AP_f	AP_{all}	AP_r	AP_c	AP_f
Ceptor	Visual-I	65.1	66.9	77.0	72.8	59.8	63.0	74.6	65.5	55.0
	Visual-G	47.3	39.1	37.0	38.1	40.4	30.9	28.8	28.3	34.6
	Text	51.7	43.1	40.6	41.1	45.4	34.3	28.2	31.4	40.1
	Hybrid	50.5	44.1	40.9	42.5	46.1	35.9	31.4	33.5	40.6

Visual-I setting, outperforming T-Rex2 by +8.5 AP on COCO, +7.6 AP on LVIS-minival [21], and +12.6 AP on the rare categories. On LVIS-val, it also achieves +0.4 AP and +2.7 AP on the rare categories. With the larger $PE_{core}L$, Ceptor further outperforms T-Rex2 with Swin-L by +13.4 AP, +10.3 AP, and +11.2 AP on COCO, LVIS-minival, and rare categories, respectively, as well as +3.0 AP and +7.6 AP on LVIS-val and its rare categories. This highlights the effectiveness of PRE in extracting informative feature regions.

In the Visual-G setting, our method with $PE_{core}S$ achieves gains of +8.5 AP on COCO, +1.7 AP on LVIS-minival, and +7.1 AP on the rare categories. With $PE_{core}L$, our method outperforms the state-of-the-art model of comparable scale by +7.5 AP on COCO, +0.2 AP on LVIS-minival, and +1.8 AP on the rare categories. On LVIS-val, Ceptor exhibits less competitive performance, which we attribute to the gap in the scale of high-quality training data. The results under both types of visual guidance demonstrate the effective transfer of generalization, and that the two types of visual prompts are harmonious.

Text Prompt Guided Detection. As shown in Tab. 2, compared with T-Rex2 [20], another detector leveraging various prompts, our method achieves stronger open-vocabulary detection performance on COCO [27] and LVIS [18], even when using less training data. Specifically, Ceptor with $PE_{core}S$ outperforms competing methods by +5.9 AP on COCO, +0.3 AP on LVIS-minival, and notably +3.2 AP on the rare categories. When comparing the larger models, the gains reach +2.2 AP on COCO and +5.5 AP on the rare categories of LVIS-minival. On LVIS-val, Ceptor performs on par with T-Rex2. Moreover, compared with text-guided open-set object detection models, Ceptor also achieves top-tier performance. These results indicate that our method unlocks the generalization capabilities of the VLM, and that the text-prompt detection route is well designed.

Hybrid Prompt Guided Detection. We further present the results of hybrid prompts for generic object detection. As shown in Tab. 3, all methods are evaluated using their smaller-scale versions. Under this setting, our method outperforms competing approaches by +8.0 AP on COCO [27], while also achieving competitive performance on LVIS [18] under limited training data. These results indicate that, in contrast to T-Rex2 [20], which completely decouples the two prompts and combines them in a post-hoc manner, our hybrid prompts interact deeply with image features during training and are jointly optimized with mul-

Table 5: Ablation study on main components.

Mode	Model	COCO-val	LVIS-minival			
		AP	AP_{all}	AP_r	AP_c	AP_f
Visual-I	Ceptor	62.3	61.1	72.4	68.5	52.5
	w/o Infusion Strategies	61.0	57.5	66.7	64.0	50.1
	w/o Harmonization Strategies	61.3	59.2	69.5	66.0	51.3
	w/o VLM	61.5	53.6	59.7	59.9	47.0
	w/o PRE (w/ Cross-Attention)	59.3	59.3	70.3	65.4	51.9
Visual-G	Ceptor	48.1	31.5	25.6	30.2	33.7
	w/o Infusion Strategies	47.9	30.8	24.9	29.5	33.0
	w/o Harmonization Strategies	46.5	29.6	24.1	27.8	32.1
	w/o VLM	47.6	30.4	21.4	29.2	33.0
Text	Ceptor	50.2	33.7	32.2	31.9	35.7
	w/o Infusion Strategies	48.9	32.8	30.4	31.0	34.9
	w/o Harmonization Strategies	48.1	32.3	31.1	30.8	33.9
	w/o VLM	49.1	33.1	30.5	31.6	34.9

tiple prompts, effectively guiding the model for object detection and showing promising potential for future exploration.

Analysis Across Diverse Prompts. As shown in Tab. 4, we observe that the generic visual prompt achieves relatively balanced predictions between rare and common categories, exhibiting greater robustness to category scarcity. The interactive visual prompt performs particularly well on rare categories, even surpassing its results on common categories. On COCO [27], the hybrid prompt achieves a result that is balanced between the generic visual prompt and the text prompt, while on LVIS [18], there is a further performance improvement. These findings suggest that the two types of prompts are complementary and tend to focus on different objects. This illustrates the significance of multi-prompt learning and mixed-prompt strategies, which enable coverage of a wider range of object types.

4.4 Ablation Study

As shown in Tab. 5, we conduct ablation studies on the main components of our framework. All model variants are trained on Objects365 [41] and use the $PE_{core}S$ backbone for fair comparison. During pre-training, we use a batch size of 8 and train for 24 epochs, with the learning rate decayed by a factor of 10 at epoch 16. For full prompt-guided training, we train for 12 epochs, with the learning rate decayed at epoch 6. (1) We remove the Infusion Strategies from Ceptor while retaining the IE and Early Interaction, and replace the Efficient Encoder with a conventional deformable transformer encoder [54]. Experimental results show that removing the Infusion Strategies negatively impacts performance across all three modes, with the effect being particularly pronounced for rare categories. This demonstrates that transferring generalization capability is essential and

Fig. 4: RoI Align on feature regions. We map the bounding boxes to the feature map of the image and perform RoI Align accordingly, then compare the extracted features with the output embeddings of the text encoder for classification. RoI Align effectively preserves the VLM's generalization capability for classification.

further validates the effectiveness of our VLM Infuser design. (2) Eliminating the Harmonization Strategies also causes performance drops in all modes, indicating that harmonization provides benefits for joint optimization. (3) When the VLM model is replaced with a standard pre-trained ViT, performance drops are observed across all three modes. This suggests that both interactive visual prompts and generic visual prompts cannot be properly aligned with image features without using a unified, generalizable VLM feature space. Moreover, the alignment between image features and text prompts also becomes ineffective. These findings highlight the necessity of introducing the VLM-based scheme in our method. (4) The ablation study on PRE shows that PRE enables prompts to preserve instance-level information and demonstrates the effectiveness of sharing the same representational source as image features.

4.5 Analysis of RoI Align on Feature Regions

We select the target to be recognized by drawing a bounding box and map the bounding box onto the image features according to the ratio between the image size and the feature size. These boxes are then used to perform RoI Align on the image features from the VLM image encoder. The resulting region features are passed through the pooling and linear layers to obtain the region embeddings. These embeddings are then compared with the text embeddings produced by the VLM's text encoder for classification. As shown in Fig. 4, even small objects in the image can be accurately classified. This demonstrates not only the classification capability of the VLM and the alignment between image and text

embeddings preserved during RoI Align, but also the spatial precision of the image features.

5 Conclusion

Inspired by the process by which humans search for objects, we design Ceptor. Specifically, we propose Infusion Strategies to deeply migrate the capabilities of a vision-language model into our framework, leveraging its alignment properties and generalizable feature space to construct diverse prompts for handling different types of objects and addressing various category distribution scenarios. In addition, we introduce a novel prompt generation method, PRE, which effectively preserves generalization, contextual information, and spatial accuracy. We further incorporate harmonization strategies, ensuring harmonious optimization under multiple prompting schemes. Ceptor achieves strong performance with multiple prompts across multiple datasets, demonstrating the effectiveness of our methodology. Experimental results show that different prompts lead to varied predictions and possess distinct advantages, indicating that multi-prompt object detectors represent a promising direction. Our hybrid prompt mechanism is extensible and can support flexible combinations of prompts. We hope our work provides new insights for prompt-based object detection.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62576144 and in part by the Hubei Provincial Science and Technology Plan under Grant 2025BEB006.

References

1. Al-Rfou, R., Choe, D., Constant, N., Guo, M., Jones, L.: Character-level language modeling with deeper self-attention. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3159–3166 (2019)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. de Beeck, H.P.O., Torfs, K., Wagemans, J.: Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. Journal of Neuroscience **28**(40), 10111–10123 (2008)
4. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Bangalath, H., et al.: Perception Encoder: The best visual embeddings are not at the output of the network. In: Advances in Neural Information Processing Systems. vol. 38, pp. 60884–60937 (2026)
5. Bravo, M.J., Farid, H.: The specificity of the search template. Journal of vision **9**(1), 34–34 (2009)

6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229 (2020)
7. Chen, Q., Jin, W., Ge, J., Liu, M., Yan, Y., Jiang, J., Yu, L., Guo, X., Li, S., Chen, J.: Cp-detr: Concept prompt guide detr toward stronger universal object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2141–2149 (2025)
8. Chen, S., Yu, E., Li, J., Tao, W.: Delving into the trajectory long-tail distribution for muti-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19341–19351 (2024)
9. Chen, S., Yu, E., Tao, W.: Cross-view referring multi-object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2204–2211 (2025)
10. Chen, S., Yu, Y., Yu, E., Tao, W.: Reamot: A benchmark and framework for reasoning-based multi-object tracking. arXiv preprint arXiv:2505.20381 (2025)
11. Cheung, O.S., Gauthier, I.: Visual appearance interacts with conceptual knowledge in object recognition. Frontiers in psychology **5**, 793 (2014)
12. DiCarlo, J.J., Cox, D.D.: Untangling invariant object recognition. Trends in cognitive sciences **11**(8), 333–341 (2007)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
14. Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14987–14997 (2025)
15. Fu, W., Li, J., Gao, B.B., Li, J., Lin, Y., Deng, H., Tao, W., Liu, Y., Wang, C.: Pet-dino: Unifying visual cues into grounding dino with prompt-enriched training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13039–13048 (2026)
16. Fu, W., Nie, C., Sun, T., Liu, J., Zhang, T., Liu, Y.: Lvis challenge track technical report 1st place solution: Distribution balanced and boundary refinement for large vocabulary instance segmentation. arXiv preprint arXiv:2111.02668 (2021)
17. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2022)
18. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
20. Jiang, Q., Li, F., Zeng, Z., Ren, T., Liu, S., Zhang, L.: T-rex2: Towards generic object detection via text-visual prompt synergy. In: European Conference on Computer Vision. pp. 38–57 (2024)
21. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)

22. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)
23. Li, J., Nie, Q., Fu, W., Lin, Y., Tao, G., Liu, Y., Wang, C.: Lors: Low-rank residual structure for parameter-efficient network stacking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15866–15876 (2024)
24. Li, J., Yu, E., Chen, S., Tao, W.: Ovtr: End-to-end open-vocabulary multiple object tracking with transformer. In: The Thirteenth International Conference on Learning Representations (2025)
25. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
26. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296 (2022)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)
29. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55 (2024)
30. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1049–1059 (2022)
31. Lu, Y., Weng, M., Xiao, Z., Jiang, R., Su, W., Zheng, G., Lu, P., Li, X.: Dynamic-dino: Fine-grained mixture of experts tuning for real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20847–20856 (2025)
32. Mao, Y., Qin, Z., Zhou, J., Fan, B., Zhang, J., Zhong, Y., Dai, Y.: Learning spatial decay for vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7945–7953 (2026)
33. Peelen, M.V., Caramazza, A.: Conceptual object representations in human anterior temporal cortex. Journal of Neuroscience **32**(45), 15728–15736 (2012)
34. Peng, Z., Xu, Z., Liu, Q., Yang, X., Shen, W.: Hyperet: Efficient training in hyperbolic space for multi-modal large language models. Advances in Neural Information Processing Systems (2025)
35. Peng, Z., Xu, Z., Tang, F., Shen, W.: Cp-clip: Customized parameter generation for open-vocabulary semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 8394–8402 (2026)
36. Peng, Z., Xu, Z., Zeng, Z., Wen, C., Huang, Y., Yang, M., Tang, F., Shen, W.: Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4562–4572 (2025)

37. Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., et al.: Grounding dino 1.5: Advance the" edge" of open-set object detection. arXiv preprint arXiv:2405.10300 (2024)
38. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 658–666 (2019)
39. Riesenhuber, M., Poggio, T.: Neural mechanisms of object recognition. Current opinion in neurobiology **12**(2), 162–168 (2002)
40. Rust, N.C., DiCarlo, J.J.: Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. Journal of Neuroscience **30**(39), 12978–12995 (2010)
41. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8430–8439 (2019)
42. Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., Ding, G.: YOLOE: Real-time seeing anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24591–24602 (2025)
43. Wang, H., Ren, P., Jie, Z., Dong, X., Feng, C., Qian, Y., Ma, L., Jiang, D., Wang, Y., Lan, X., et al.: Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. arXiv preprint arXiv:2407.07844 (2024)
44. Wang, J., Zhang, P., Chu, T., Cao, Y., Zhou, Y., Wu, T., Wang, B., He, C., Lin, D.: V3det: Vast vocabulary visual detection dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19844–19854 (2023)
45. Wolfe, J.M.: Guided search 6.0: An updated model of visual search. Psychonomic bulletin & review **28**(4), 1060–1092 (2021)
46. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024)
47. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23497–23506 (2023)
48. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. Advances in Neural Information Processing Systems **35**, 9125–9138 (2022)
49. Yao, L., Pi, R., Han, J., Liang, X., Xu, H., Zhang, W., Li, Z., Xu, D.: Detclipv3: Towards versatile generative open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27391–27401 (2024)
50. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: European Conference on Computer Vision. pp. 106–122 (2022)
51. Zhang, J., Wang, P., Liu, C., Liu, W., Jin, D., Zhang, Q., Meng, E., Hu, Z.: Just a few glances: Open-set visual perception with image prompt paradigm. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9969–9976 (2025)
52. Zhao, T., Liu, P., He, X., Zhang, L., Lee, K.: Real-time transformer-based open-vocabulary detection with efficient fusion head. arXiv preprint arXiv:2403.06892 (2024)

53. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16965–16974 (2024)
54. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)