# OV-DQUO: Open-Vocabulary DETR with Denoising Text Query Training and Open-World Unknown Objects Supervision

**Junjie Wang**[1,2], **Bin Chen**[2,3,4,5*], **Bin Kang**[3], **Yulin Li**[1,2], **Weizhi Xian**[4], **Yichi Chen**[3], **Yong Xu**[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[2]International Research Institute for Artificial Intelligence, Harbin Institute of Technology, Shenzhen
[3]School of Computer Science and Technology, University of Chinese Academy of Sciences
[4]Chongqing Research Institute of HIT
[5]National Key Laboratory of Smart Farm Technologies and Systems
{jjwanghz, yulin}@stu.hit.edu.cn, {chenbin2020, laterfall}@hit.edu.cn, {kangbin23, chenyichi21}@mails.ucas.ac.cn

## Abstract

Open-vocabulary detection aims to detect objects from novel categories beyond the base categories on which the detector is trained. However, existing open-vocabulary detectors trained on base category data tend to assign higher confidence to trained categories and confuse novel categories with the background. To resolve this, we propose OV-DQUO, an Open-Vocabulary DETR with Denoising text Query training and open-world Unknown Objects supervision. Specifically, we introduce a wildcard matching method. This method enables the detector to learn from pairs of unknown objects recognized by the open-world detector and text embeddings with general semantics, mitigating the confidence bias between base and novel categories. Additionally, we propose a denoising text query training strategy. It synthesizes foreground and background query-box pairs from open-world unknown objects to train the detector through contrastive learning, enhancing its ability to distinguish novel objects from the background. We conducted extensive experiments on the OV-COCO and OV-LVIS benchmarks, achieving new state-of-the-art results of 45.6 AP50 and 39.3 mAP on novel categories, respectively.

**Code** — https://github.com/xiaomoguhz/OV-DQUO

## Introduction

Open-Vocabulary Detection (OVD) (Zareian et al. 2021) focuses on identifying objects from novel categories not encountered during training. Recently, Vision-Language Models (VLMs) (Sun et al. 2023; Li et al. 2023b) pretrained on large-scale image-text pairs, such as CLIP (Radford et al. 2021), have demonstrated impressive performance in zero-shot image classification, providing new avenues for OVD.

ViLD (Gu et al. 2021) is the first work to distill VLMs' classification knowledge into an object detector by aligning the detector-generated region embeddings with the corresponding features extracted from VLMs. Subsequent methods (Wu et al. 2023a; Wang et al. 2023; Wu et al. 2024b; Zang et al. 2022; Li et al. 2023a) have proposed more elaborately designed strategies to improve the efficiency of knowledge distillation, such as BARON (Wu et al. 2023a), which aligns bag-of-regions embeddings with image features extracted by VLMs. However, the context discrepancy limits the effectiveness of knowledge distillation (Zhu and Chen 2023). RegionCLIP (Zhong et al. 2022) is a representative pseudo-labeling method that employs VLM and RPN to generate region-text pairs from image-caption datasets (Sharma et al. 2018) for training open-vocabulary detectors. Later works (Chen et al. 2022; Zhao et al. 2024, 2022; Minderer et al. 2024) have further extended the implementation of pseudo-labeling. For instance, SAS-Det (Zhao et al. 2024) incorporates self-training paradigms into OVD. Nevertheless, these methods suffer from pseudo-label noise.

All of the above methods employ indirect utilization of VLMs, thus not unleashing their full potential. Current state-of-the-art methods (Wu et al. 2024a, 2023b; Kuo et al. 2023) typically employ a frozen VLM image encoder as the backbone to extract region features associated with prediction boxes for classifying novel category objects. Intuitively, the performance ceiling of such methods depends directly on the classification ability of VLMs. Therefore, current works mainly enhance a VLM's region recognition accuracy through fine-tuning (Wu et al. 2023b, 2024b) or self-distillation (Wu et al. 2024a). Yet, these methods overlook the fact that **detectors trained on base category data tend to assign higher confidence scores to trained categories and confuse novel categories with the background.**

To verify the impact of confidence bias on novel category detection, we first analyze the confidence score assigned by VLMs and detectors to base and novel categories, as shown in Figure 1**(a)**. It is evident that the detector assigns significantly lower confidence scores to novel category objects (e.g., umbrella) than to base categories (e.g., person). Furthermore, we observed a significant performance gap when using a VLM to classify Ground Truth (GT) boxes compared to detector predictions. However, this gap narrows when we manually eliminate the confidence bias by adjusting the prediction confidence of bounding boxes based on their Intersection over Union (IoU) with GT boxes. The experimental results reveal that **confidence bias constitutes one of the factors responsible for suboptimal performance in novel category detection.**

*Corresponding Author.

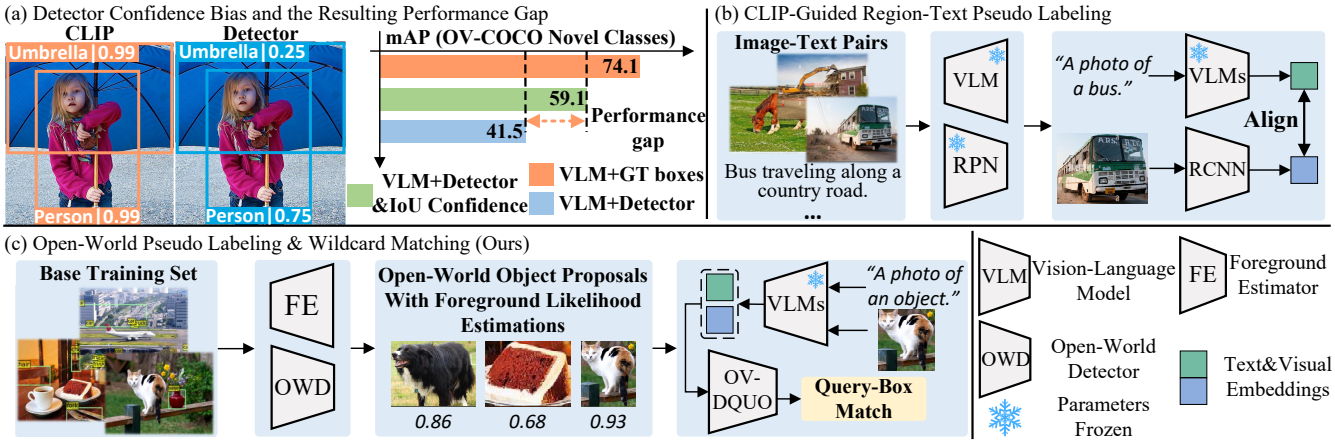



Figure 1: **(a)** Detector confidence bias is a primary reason for suboptimal detection performance on novel categories. **(b)** Existing pseudo-labeling methods mainly focus on establishing region-text alignment from external caption datasets whereas ignoring the confidence bias. **(c)** Instead, this work directly tackles this confidence bias issue by utilizing the open-world detector to discover novel unknown objects during training and learning to match them with wildcard text embeddings.

Based on the above findings, we propose OV-DQUO, an OVD framework with denoising text query training and open-world unknown objects supervision. Unlike existing pseudo-labeling methods that aim to enable a detector to acquire region-text alignment from VLMs (Figure 1**(b)**), we identify that a frozen VLM serves as an effective region classifier and is not the performance bottleneck of current advanced OVD models. In contrast, we aim to address the confidence bias in OVD models, a challenge that leads to performance degradation in novel category detection.

As shown in Figure 1**(c)**, to address the confidence bias between base and novel categories, we propose the open-world pseudo-labeling and wildcard matching methods. This approach enables a detector to learn to use text embeddings with general semantics to match unknown objects recognized by open-world detectors, preventing them from being regarded as background during training. Since the open-world detector cannot identify all potential novel objects, we developed a denoising text query training method to reduce the detector's confusion between novel categories and the background. It synthesizes foreground and background query-box pairs from open-world unknown objects, thereby enabling a detector to better distinguish novel objects from the background through contrastive learning. Finally, to reduce the impact of confidence bias on the region proposal selection, we propose a Region of Query Interests (RoQIs) selection method that integrates region-text similarity with confidence scores for proposal selection, achieving a more balanced recall of base and novel category objects. The main contributions of this paper are summarized as follows:

- Inspired by the open-world detection task of recognizing unknown objects, we propose the OV-DQUO framework, which aims to mitigate the confidence bias of the OVD model in detecting novel categories.
- We introduce a wildcard matching method that enables the detector to learn from pairs of text embeddings with general semantics and unknown objects recognized by the open-world detector, thereby alleviating the confidence bias between base and novel categories.
- We propose a denoising text query training strategy that allows a detector to perform contrastive learning from synthetic query-box pairs, improving its ability to distinguish novel objects from the background.
- OV-DQUO consistently outperforms existing state-of-the-art methods on the OV-COCO and OV-LVIS OVD benchmarks and demonstrates excellent performance in cross-dataset detection on COCO and Objects365.

## Related Works

**Open-Vocabulary Detection** is a paradigm proposed by OVR-CNN (Zareian et al. 2021), which aims to train models to detect objects from arbitrary categories, including those not seen during training. State-of-the-art methods (Kuo et al. 2023; Wu et al. 2023b, 2024a) leverage a frozen VLM image encoder as the backbone to extract features and perform OVD. Compared to pseudo-labeling (Bangalath et al. 2022; Zhou et al. 2022; Zhong et al. 2022; Zhao et al. 2022; Ma et al. 2024; Song and Bang 2023) and knowledge distillation-based methods (Wu et al. 2023a; Wang et al. 2023; Wu et al. 2024b; Zang et al. 2022; Li et al. 2023a), these approaches directly benefit from the large-scale pre-training knowledge of VLMs and can better generalize to novel objects. F-VLM (Kuo et al. 2023) pioneered the discovery that VLMs retain region-sensitive features useful for object detection. It freezes the VLM and uses it as a backbone for feature extraction and region classification. CORA (Wu et al. 2023b) also uses a frozen VLM but fine-tunes it with a lightweight region prompt layer, enhancing region classification accuracy. CLIPself (Wu et al. 2024a) reveals that the ViT version of VLM performs better on image crops than on dense features, and explores aligning dense features with image crop features through self-distillation. However, we identify that these methods suffer from the confidence

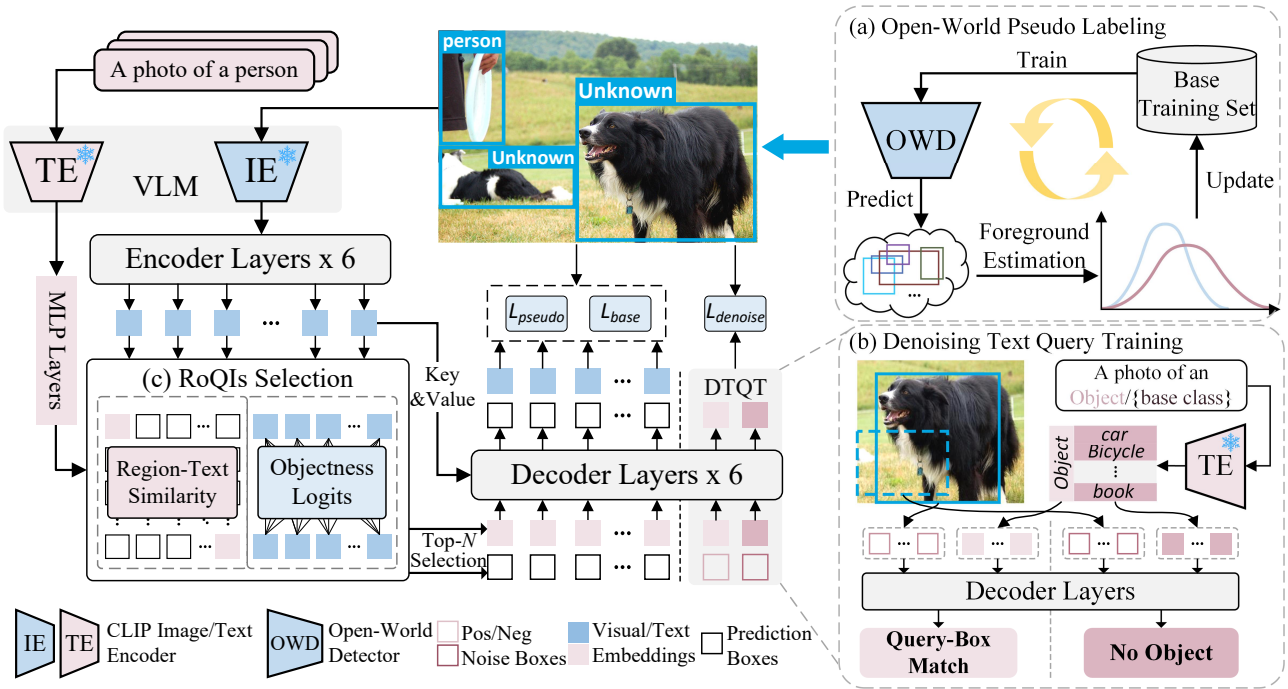



Figure 2: **Overview of OV-DQUO. (a)** Open-world pseudo labeling pipeline, which iteratively trains the detector, generates unknown object proposals, estimates foreground probabilities, and updates the training set. **(b)** Denoising text query training, which enables contrastive learning with synthetic noisy query-box pairs from open-world unknown objects. **(c)** RoQIs selection module, which takes into account both objectness and region-text similarity for selecting regions of interest.

bias issue, resulting in suboptimal OVD performance.

**Open-World Detection (OWD)** is a paradigm proposed by ORE (Joseph et al. 2021), which aims to achieve two goals: (1) recognizing both known category objects and the unknown objects not present in the training set, and (2) enabling incremental object detection learning through newly introduced external knowledge. OW-DETR (Gupta et al. 2022) attempts to identify potential unknown objects based on feature map activation scores, as foreground objects typically exhibit stronger activation responses compared to the background. PROB (Zohar, Wang, and Yeung 2023) performs distribution modeling on the model output logits to identify unknown objects and decouples the identification of background, known objects, and unknown objects. Based on the observation that foreground regions exhibit more variability while background regions change monotonously, MEPU (Fang et al. 2023) employs Weibull modeling on the feature reconstruction error of these regions and proposes the Reconstruction Error-based Weibull (REW) model. REW assigns likelihood scores to region proposals that potentially belong to unknown objects. These methods inspire us to leverage open-world detectors to address the confidence bias issue in OVD.

## Method

In this section, we introduce OV-DQUO, a novel OVD framework designed to mitigate confidence bias in detecting novel categories. Figure 2 offers an overview of OV-DQUO. We begin with a brief review of the OVD setup and conditional matching methods. Then, we detail the open-world pseudo-labeling pipeline and the corresponding wildcard matching method, which form our key approach for mitigating the confidence bias between base and novel categories. Subsequently, we elaborate on the denoising text query training strategy that enhances the model's ability to distinguish novel objects from the background. Finally, we detail the RoQIs selection method, which achieves a more balanced recall of base and novel category objects.

### Preliminaries

**Task Formulation.** In this study, we adhere to the classical open-vocabulary problem setup as outlined in OVR-CNN (Zareian et al. 2021). In this setup, only partial class annotations of the dataset are available during the training phase, commonly referred to as base classes, denoted by the symbol $\mathcal{C}^{\text{base}}$. During the inference stage, the model is tasked with recognizing objects from both the base classes and the novel classes (denoted as $\mathcal{C}^{\text{novel}}$, where $\mathcal{C}^{\text{base}} \cap \mathcal{C}^{\text{novel}} = \varnothing$) that were not encountered during training, while the names of the novel classes are provided as cues during inference.

**Conditional Matching.** A DETR-style object detector consists of three parts: a backbone network, an encoder, and a decoder. The encoder refines feature maps extracted by the backbone and generates region proposals. The decoder refines a set of object queries with their associated region proposals into the final box and classification predictions.

OV-DETR (Zang et al. 2022) and CORA (Wu et al. 2023b) modify the decoder with the conditional matching method to achieve OVD. Specifically, each object query $q_i$ is assigned a label $c_i$ by classifying its associated region proposal $b_i$

$$c_i = \underset{c \in \mathcal{C}^{\text{base}}}{\text{argmax}} \ \text{cosine}\left(v_i, t_c\right), \tag{1}$$

where $v_i$ represents the region feature of $b_i$, obtained by performing RoI Align (He et al. 2017) on the feature map from a frozen VLM, and $t_c$ denotes the text embedding of class $c$. cosine denotes the cosine similarity. Then, the class-aware object query $q_i^*$ is given by

$$q_i^* = q_i + \text{MLP}\left(t_{c_i}\right), \tag{2}$$

where $q_i$ denotes the vanilla object query. The decoder iteratively refines each object query with its corresponding region proposal $(q_i^*, b_i)$ into $(\hat{m}_i, \hat{b}_i)$, where $\hat{b}_i$ represents the refined box coordinates and $\hat{m}_i$ is a sigmoid probability scalar indicating that the object within $\hat{b}_i$ matches the query category $c_i$. During inference, the frozen VLM is responsible for classifying the prediction box $\hat{b}_i$, and the classification score for each category is multiplied by the corresponding matching probability $\hat{m}_i$ to account for box quality

$$\text{P}(\hat{b}_i \in c) = \hat{m}_i \, \text{cosine}\left(\hat{v}_i, t_c\right). \tag{3}$$

## Open-World Pseudo Labeling & Wildcard Matching

Since only base category annotations are available during OVD model training, novel category objects are treated as background, leading to the confidence bias between base and novel categories. OV-DQUO provides a novel solution by leveraging open-world detectors to discover potential novel unknown objects, as shown in Figure 2**(a)**.

**Open-world detector training and proposal generation.** Open-world detector OLN (Kim et al. 2022) has been shown to have superior recall for novel categories compared to RPN. Therefore, we utilize it to identify potential novel category objects and generate pseudo-labels. Specifically, we first train the OLN using $\mathcal{C}^{\text{base}}$ data. After training, we apply it to infer the training set and generate open-world object proposals. Given an input image $I \in \mathbb{R}^{3 \times H \times W}$, OLN outputs a series of unknown objects $U = \{u_1, u_2, \ldots, u_n\}$, where each $u_i = (o_i, s_i)$. Here, $o_i$ denotes the coordinates of an unknown object, and $s_i$ indicates the localization quality. Subsequently, to mitigate the noisy boxes introduced by open-world pseudo-labeling, we use a Foreground Estimator (FE) (Fang et al. 2023) to estimate the probability that each $u_i$ is foreground as soft labels.

**Foreground likelihood estimation.** FE distinguishes foreground and background regions by feature reconstruction error, as background regions often have simpler patterns with lower reconstruction errors, while foreground regions are more complex with higher errors. Specifically, we first train a feature reconstruction network in an unsupervised setting. Then, we collect the feature reconstruction errors for foreground and background regions based on the $\mathcal{C}^{\text{base}}$ annotations. After that, we can model the probability distributions for the foreground and background separately (denoted as $\mathcal{D}_{fg}$ and $\mathcal{D}_{bg}$) by applying maximum likelihood estimation to the following equation

$$\mathcal{D}(\eta|a, c) = ac\left[1 - \exp\left(-\eta^c\right)\right]^{a-1} \exp\left(-\eta^c\right) \eta^{c-1}, \tag{4}$$

where $a$ and $c$ represent the shape parameters of the distribution, while $\eta$ represents the feature reconstruction errors. With $\mathcal{D}_{fg}$ and $\mathcal{D}_{bg}$, we can estimate the foreground likelihood $w_i$ for each open-world unknown object $u_i$ as follows

$$w_i = \frac{\mathcal{D}_{fg}\left(\eta_{o_i}\right)}{\mathcal{D}_{fg}\left(\eta_{o_i}\right) + \mathcal{D}_{bg}\left(\eta_{o_i}\right)}, \tag{5}$$

where $\eta_{o_i}$ represents the reconstruction error of $o_i$. Then each $u_i \in U$ is updated with its corresponding foreground likelihood estimation, resulting in $u_i = (o_i, s_i, w_i)$.

**Training set update with open-world proposals.** For each $I \in \mathcal{C}^{\text{base}}$, its corresponding $U$ is used to update the annotations after being filtered by GT annotations, $s_i$, and some heuristic criteria. Subsequently, the entire process can be iterated upon to generate more open-world unknown object proposals.[1]

**Wildcard Matching.** open-world pseudo labels enable an OVD model to avoid treating potential novel objects as background during training. However, incorporating such pseudo labels into the OVD training framework raises the following challenge: open-world unknown objects lack category details. OV-DQUO offers an elegant method that uses text embeddings with general semantics to match these objects.

Let $t_{ow}$ denotes the text embedding for unknown objects, obtained by encoding the text "a photo of a {wildcard}" using the text encoder of the VLM. The wildcard can be words like "object" or "thing", etc., which can consistently exhibit a certain degree of visual-text similarity with any foreground visual object. It is important to note that our proposed wildcard matching method is not a substitute for conditional matching, but rather a complementary approach. During the training process, if a region proposal $b_i$ generated by the encoder has an IoU with unknown objects greater than the threshold $\tau$, we add $t_{ow}$ to its associated vanilla object query $q_i$. Otherwise, we add the text embedding of the base category with the maximum similarity $t_{c_i}$ to $q_i$

$$q_i^* = \begin{cases} q_i + \text{MLP}\left(t_{ow}\right) & \text{if IoU}(b_i, U) > \tau, \\ q_i + \text{MLP}\left(t_{c_i}\right) & \text{otherwise}, \end{cases} \tag{6}$$

where $U$ represents the set of open-world pseudo labels associated with the input image $I \in \mathcal{C}^{\text{base}}$. Then, the decoder iteratively refines each object query with its corresponding region proposal $(q_i^*, b_i)$ into $\hat{y}_i = (\hat{m}_i, \hat{b}_i)$. To achieve text query conditional matching, we constrain each ground-truth box to match only the predictions associated with object queries of the same category. Given the set of open-world unknown objects $U$ and the prediction set $Y^{ow} = \{\hat{y}_1^{ow}, \hat{y}_2^{ow}, \ldots, \hat{y}_n^{ow} \mid q_i^* = q_i + \text{MLP}\left(t_{ow}\right)\}$, we leverage the Hungarian matching algorithm to find the optimal pairing $\hat{\sigma}_{ow}$ with minimal cost between the two sets

$$\hat{\sigma}_{ow} = \text{HM}(\mathcal{L}_{\text{cost}}, U, Y^{ow}), \tag{7}$$

[1]Details can be found in https://arxiv.org/pdf/2405.17913

where HM denotes the Hungarian matching algorithm. The cost function $\mathcal{L}_{\text{cost}}$ is defined as

$$\mathcal{L}_{\text{cost}}(y, \hat{y}) = \mathcal{L}_{\text{match}}(m, \hat{m}) + \mathcal{L}_{\text{bbox}}(b, \hat{b}), \tag{8}$$

where $\mathcal{L}_{\text{match}}$ denotes the Focal loss (Lin et al. 2017), while $\mathcal{L}_{\text{bbox}}$ consists of L1 and GIoU (Zhai, Cheng, and Wang 2020) loss. The optimal pairing results for base categories $(\hat{\sigma}_c \mid c \in \mathcal{C}^{\text{base}})$ can be obtained similarly. The loss for open-world unknown objects is defined as follows:

$$\mathcal{L}_{\text{pseudo}} = \boldsymbol{w}\mathcal{L}_{\text{match}}(\mathbf{m}^{ow}, \hat{\mathbf{m}}^{ow}_{\hat{\sigma}_{ow}}), \tag{9}$$

where $\boldsymbol{w}$ is a 1-dimensional vector that denotes the foreground likelihood estimation of $U$. The loss for base categories can be expressed as follows:

$$\mathcal{L}_{\text{base}} = \sum_{c \in \mathcal{C}^{\text{base}}} \mathcal{L}_{\text{match}}\left(\mathbf{m}^c, \hat{\mathbf{m}}^c_{\hat{\sigma}_c}\right) + \mathcal{L}_{\text{bbox}}\left(\mathbf{b}^c, \hat{\mathbf{b}}^c_{\hat{\sigma}_c}\right). \tag{10}$$

During inference, the behavior of wildcard matching aligns with the conditional matching method.

## Denoising Text Query Training

The open-world pseudo-labeling and wildcard matching methods help alleviate the confidence bias between base and novel categories. However, the open-world detector cannot recognize all potential novel objects. Therefore, as shown in Figure 2**(b)**, OV-DQUO introduces another method, denoising text query training, which improves the ability of OVD models to distinguish novel objects from the background.

Specifically, given an open-world unknown object $u_i$, $2N$ noise boxes are generated based on its coordinates $o_i$

$$\tilde{o}_i = \begin{cases} o_i + \lambda_1 \cdot \epsilon(o_i) & \text{if } 0 \le i < N, \\ o_i + \lambda_2 \cdot \epsilon(o_i) & \text{otherwise,} \end{cases} \tag{11}$$

Where $\lambda_1 \sim \text{Uniform}(0, 1)$ and $\lambda_2 \sim \text{Uniform}(1, 2)$ denote the noise scales. The function $\epsilon(\cdot)$ computes the basic offset, defined as half the width and height of the input box. The first $N-1$ boxes exhibit a higher IoU with $o_i$ (blue box in Figure 2**(b)**). In contrast, the boxes from $N$ to $2N-1$ have a little IoU with $o_i$ (blue dashed box in Figure 2**(b)**). Then, we synthesize foreground query-box pairs by adding the correct text embedding $t_{ow}$ to the vanilla object query $\tilde{q}_i$ of the first $N-1$ boxes. Besides, we synthesize background query-box pairs by randomly adding incorrect text embeddings from base categories $(t_c | c \in \mathcal{C}^{\text{base}})$ to the object query of the later $N$ to $2N-1$ boxes based on a probability $\rho$

$$\tilde{q}^*_i = \begin{cases} \tilde{q}_i + \text{MLP}\left(t_c\right) & \text{if } N \le i < 2N \text{ and } \lambda_1 < \rho, \\ \tilde{q}_i + \text{MLP}\left(t_{ow}\right) & \text{otherwise.} \end{cases} \tag{12}$$

During training, the decoder simultaneously refines the vanilla part $(q^*_i, b_i)$ and the denoising part $(\tilde{q}^*_i, \tilde{o}_i)$, using an attention mask for isolation to prevent information leakage. The denoising training loss is defined as follows:

$$\mathcal{L}_{\text{denoise}} = \sum_{i=0}^{2N} w_i \mathcal{L}_{\text{match}}\left(\mathbb{I}_{(0<i<N)},\ \tilde{m}_i\right), \tag{13}$$

where $\mathbb{I}$ is the indicator function, which equals 1 if $0 < i < N$ and 0 otherwise. $\tilde{m}_i$ denotes the match probability of the denoising part. $w_i$ is the foreground likelihood estimation. The overall training objective for this framework is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pseudo}} + \mathcal{L}_{\text{base}} + \beta\mathcal{L}_{\text{denoise}}, \tag{14}$$

where $\beta$ denotes the denoising loss weight.

## Region of Query Interests Selection

Open-world pseudo-labeling and denoising text query training methods mitigate the confidence bias in detecting novel categories. However, this bias also impedes the intermediate region proposal selection process, resulting in a preference for base category objects as region proposals. Consequently, as illustrated in Figure 2**(c)**, OV-DQUO introduces a RoQIs selection method to improve this process.

OV-DQUO relies on the objectness score $o$ to select region proposals. $o$ denotes the probability of an object being present in a region, which favors the base category objects due to confidence bias. Since CLIP's classification scores exhibit less bias (Figure 1**(a)**), we use them to mitigate this issue. However, CLIP's scores are insensitive to localization quality, which can affect detection performance. To address this, we combine CLIP's scores with $o$ using a geometric mean. Given the objectness score vector $\mathbf{o}$, CLIP's region features $\mathbf{v}$, text embeddings of class names $\mathbf{t}$, we compute the new criterion $\varphi$ for selecting region proposals

$$\varphi = (\text{Max}(\mathbf{v} \cdot \mathbf{t}^\top))^\alpha \cdot \mathbf{o}^{(1-\alpha)}, \tag{15}$$

where $\text{Max}$ means the maximum similarity of each region feature to all text embeddings. $\alpha$ is the weighted geometric mean parameter. With criterion $\varphi$, we can select the regions of interest $B^*$ from the region proposal set $B$

$$B^* = \text{gather}(B, \varphi, N), \tag{16}$$

where $\text{gather}$ denotes the operation of selecting the top $N$ proposals from $B$ according to $\varphi$.

# Experiments

## Datasets and Evaluation Metrics

**OV-COCO benchmark.** Following (Zareian et al. 2021), we divide the 80 classes in the COCO dataset (Lin et al. 2014) into 48 base classes and 17 novel classes. The training set comprises 107,761 images of base category annotations, and the test set includes 4,836 images with both base and novel category annotations. In this benchmark, we report the box mean Average Precision (mAP) at IoU threshold 0.5, which is denoted as $\text{AP}_{50}$. $\text{AP}_{50}$ of novel categories ($\text{AP}^{\text{Novel}}_{50}$) is the widely used major metric to evaluate the OVD performance on OV-COCO benchmark.

**OV-LVIS benchmark.** Following standard practice (Gu et al. 2021), we remove categories with rare tags in the LVIS (Gupta, Dollar, and Girshick 2019) training set. Models are trained on 461 common classes and 405 frequent classes, which contain 100,170 images. After training, the models are evaluated on the validation set, which includes the common, frequent, and rare classes, containing 19,809 images. In this benchmark, we report the mAP of boxes averaged on IoUs from 0.5 to 0.95. The mAP of rare categories ($\text{mAP}_r$) is the widely used major metric to evaluate the OVD performance on OV-LVIS benchmark.

## Implementation Details

OV-DQUO is built on the closed-set detector DINO (Zhang et al. 2023) and is configured with 1,000 object queries, 6 encoder layers, and 6 decoder layers. In the OV-COCO benchmark, we use CLIP models RN50 and RN50x4 (Wu et al.

| Method | Supervision | Backbone | $AP_{50}^{Novel}$ |
|---|---|---|---|
| ViLD (Gu et al. 2021) | CLIP | RN50 | 27.6 |
| Detic (Zhou et al. 2022) | Caption | RN50 | 27.8 |
| OV-DETR (Zang et al. 2022) | CLIP | RN50 | 29.4 |
| ProxyDet (Jeong et al. 2024) | Caption | RN50 | 30.4 |
| RegionCLIP (Zhong et al. 2022) | Caption | RN50 | 31.4 |
| RTGen (Chen et al. 2024) | Caption | RN50 | 33.6 |
| BARON-KD (Wu et al. 2023a) | CLIP | RN50 | 34.0 |
| CLIM (Wu et al. 2024b) | CLIP | RN50 | 36.9 |
| SAS-Det (Zhao et al. 2024) | CLIP | RN50 | 37.4 |
| RegionCLIP (Zhong et al. 2022) | Captions | RN50x4 | 39.3 |
| CORA (Wu et al. 2023b) | CLIP | RN50x4 | 41.7 |
| PromptDet (Song and Bang 2023) | Caption | ViT-B/16 | 30.6 |
| RO-ViT (Kim et al. 2023b) | CLIP | ViT-L/16 | 33.0 |
| CFM-ViT (Kim et al. 2023a) | CLIP | ViT-L/16 | 34.1 |
| BIND (Zhang et al. 2024) | CLIP | ViT-L/16 | 41.5 |
| CLIPSelf (Wu et al. 2024a) | CLIP | ViT-L/14 | 44.3 |
| **OV-DQUO (Ours)** | CLIP | RN50 | 39.2 |
| **OV-DQUO (Ours)** | CLIP | RN50x4 | **45.6** |

(a) OV-COCO benchmark

| Method | Supervision | Backbone | $mAP_r$ |
|---|---|---|---|
| ViLD (Gu et al. 2021) | CLIP | RN50 | 16.3 |
| OV-DETR (Zang et al. 2022) | CLIP | RN50 | 17.4 |
| BARON-KD (Wu et al. 2023a) | CLIP | RN50 | 22.6 |
| RegionCLIP (Zhong et al. 2022) | Caption | RN50x4 | 22.0 |
| $CORA^+$ (Wu et al. 2023b) | Caption | RN50x4 | 28.1 |
| SAS-Det (Zhao et al. 2024) | CLIP | RN50x4 | 29.1 |
| CLIM (Wu et al. 2024b) | CLIP | RN50x64 | 32.3 |
| F-VLM (Kuo et al. 2023) | CLIP | RN50x64 | 32.8 |
| RTGen (Chen et al. 2024) | Caption | Swin-B | 30.2 |
| BIND (Zhang et al. 2024) | CLIP | ViT-L/16 | 32.5 |
| Detic (Zhou et al. 2022) | Caption | Swin-B | 33.8 |
| CFM-ViT (Kim et al. 2023a) | CLIP | ViT-L/14 | 33.9 |
| RO-ViT (Kim et al. 2023b) | CLIP | ViT-H/16 | 34.1 |
| CLIPSelf (Wu et al. 2024a) | CLIP | ViT-L/14 | 34.9 |
| ProxyDet (Jeong et al. 2024) | Caption | Swin-B | 36.7 |
| CoDet (Ma et al. 2024) | Caption | ViT-L/14 | 37.0 |
| **OV-DQUO (Ours)** | CLIP | ViT-B/16 | 29.7 |
| **OV-DQUO (Ours)** | CLIP | ViT-L/14 | **39.3** |

(b) OV-LVIS benchmark

Table 1: Comparison with state-of-the-art open-vocabulary object detection methods. Caption supervision indicates that the method learns from extra image-text pairs, while CLIP supervision refers to transferring knowledge from CLIP.

| Method | COCO | | Objects365 | |
|---|---|---|---|---|
| | AP | $AP_{75}$ | AP | $AP_{75}$ |
| Supervised Baseline (Gu et al. 2021) | 46.5 | 50.9 | 25.6 | 28.0 |
| ViLD (Gu et al. 2021) | 36.6 | 39.6 | 11.8 | 12.6 |
| DetPro (Du et al. 2022) | 34.9 | 37.4 | 12.1 | 12.9 |
| BARON (Wu et al. 2023a) | 36.2 | 39.1 | 13.6 | 14.5 |
| F-VLM (Kuo et al. 2023) | 37.9 | 41.2 | 16.2 | 17.5 |
| CoDet (Ma et al. 2024) | 39.1 | 42.3 | 14.2 | 15.3 |
| **OV-DQUO (Ours)** | **39.2** | **42.5** | **18.4** | **19.6** |

Table 2: Cross-dataset evaluation on COCO and Objects365.

| # | Open-World Supervision | Denoising Text Query Training | RoQIs Selection | $AP_{50}^{Novel}$ | $AP_{50}^{All}$ |
|---|---|---|---|---|---|
| 1 | - | - | - | 41.7 | 46.4 |
| 2 | ✓ | ✗ | ✗ | 43.3 | 47.3 |
| 3 | ✓ | ✓ | ✗ | 45.0 | 47.9 |
| 4 | ✗ | ✗ | ✓ | 42.7 | 46.6 |
| 5 | ✓ | ✓ | ✓ | **45.6** | 48.1 |

Table 3: Ablation study on main effective components.

2023b) as the backbone networks. In the OV-LVIS benchmark, we employ self-distilled CLIP models ViT-B/16 and ViT-L/14 (Wu et al. 2024a) as the backbone networks. We train OV-DQUO using 8 GPUs with a batch size of 4 per GPU for 30 epochs, employing the AdamW optimizer with a learning rate of $1e{-}4$. The threshold $\tau$ for wildcard matching is set at 0.5. The threshold $\rho$ for denoising training is set at 0.25. The geometric mean parameter $\alpha$ for RoQIs selection is set to 0.45. The loss weights for class, bbox, and GIoU are set to 2.0, 5.0, and 2.0, respectively. The weight $\beta$ for denoising loss is set to 2.0.

## Benchmark Results

**OV-COCO.** Table 1(a) provides details on the performance of OV-DQUO on the OV-COCO benchmark. To ensure a fair comparison, we list the external training resources and backbone architectures utilized by each method, as these factors vary across methods and significantly impact performance. When comparing with methods trained on CLIP RN50, OV-DQUO outperforms the previously best-performing method SAS-Det by 1.8 $AP_{50}$ in novel classes. When comparing with methods trained on CLIP RN50x4, OV-DQUO outperforms the previously best-performing method CORA by 3.9 $AP_{50}$. Even when compared to the current state-of-the-art method CLIPself, which has a larger backbone (ViT-L), OV-DQUO still maintains a lead of 1.3 $AP_{50}$.

**OV-LVIS.** Table 1(b) summarizes the main results of OV-DQUO on the OV-LVIS benchmark. Since the LVIS dataset encompasses considerably more categories than COCO (1203 vs. 80), we replaced the backbone network with those of stronger classification capabilities, ViT-B/16 and ViT-L/14 (Wu et al. 2024a), in the OV-LVIS experiments. It is worth noting that this does not lead to an unfair comparison, as OV-DQUO still consistently outperforms all state-of-the-art methods. When comparing with methods trained on CLIP ViT-L/14 without external training resources, OV-DQUO surpasses the previously best method CLIPself by 4.4 $mAP_r$. When comparing with the current most advanced method CoDet using external image-caption data supervision, OV-DQUO still achieves a lead of 2.3 $mAP_r$.

**Cross-Dataset Evaluation.** Given that the open-vocabulary detector may encounter data from various domains in open-world applications, we further evaluate OV-DQUO under a cross-dataset setting. Table 2 presents the primary results of transferring OV-DQUO trained on OV-LVIS to the valida-

| Iterations | $AR_{50}^{All}$ | $AP_{50}^{Novel}$ | $AP_{50}^{All}$ |
|---|---|---|---|
| - | 80.2 | 41.7 | 46.4 |
| 1 | 85.7 | 44.0 | 47.9 |
| **2** | 86.5 | **45.0** | 47.9 |
| 3 | 87.1 | 44.8 | 48.5 |

(a) Ablation study on iterations of open-world pseudo-labeling

| Wildcard Text | $AP_{50}^{Novel}$ | $AP_{50}^{All}$ |
|---|---|---|
| `"Salient Object"` | 44.4 | 47.0 |
| `"Target"` | 44.5 | 47.5 |
| `"Thing"` | 44.9 | 47.2 |
| **"Object"** | **45.0** | 47.9 |

(b) Ablation study on different choices of wildcard text

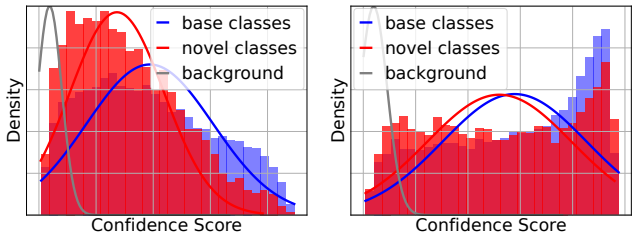
| $\gamma$ | $AP_{50}^{Novel}$ | $AP_{50}^{All}$ |
|---|---|---|
| 0.0 | 43.0 | 46.2 |
| **0.5** | **45.0** | 47.9 |
| 1.0 | 44.4 | 47.3 |
| 2.0 | 44.1 | 46.7 |

(c) Ablation study on scaling foreground score

| $\beta$ | $AP_{50}^{Novel}$ | $AP_{50}^{All}$ |
|---|---|---|
| 1.0 | 44.8 | 47.4 |
| **2.0** | **45.0** | 47.9 |
| 3.0 | 44.4 | 47.7 |
| 4.0 | 44.4 | 47.5 |

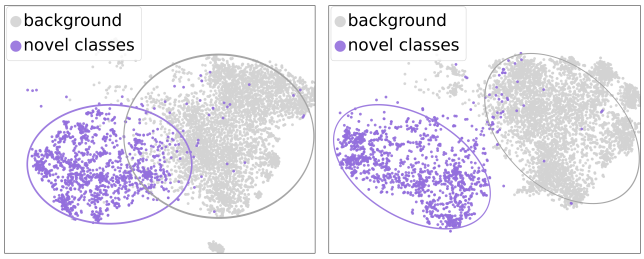
(d) Ablation study on denoising loss weight

Table 4: Ablation studies of hyperparameters in OV-DQUO on OV-COCO benchmark. Our default settings are marked in bold.




(a) Baseline Detector (b) OV-DQUO

Figure 3: Visualization of confidence score distributions.




(a) Baseline Detector (b) OV-DQUO

Figure 4: T-SNE visualization of embedding distributions.

tion sets of COCO and Object365 (Shao et al. 2019). The experiments demonstrate that OV-DQUO achieves competitive results on the COCO dataset and surpasses the previous leading method, F-VLM, by 2.2 AP on the Object365 dataset, demonstrating robust cross-dataset generalization.

## Ablation Study

**Ablation Study on Main Components.** As shown in Table 3, with the RN50x4 backbone, the baseline OV-DQUO achieves 41.7 $AP_{50}^{Novel}$ (#1). Additional supervision from open-world unknown objects raises this to 43.3 $AP_{50}^{Novel}$ (#2). Furthermore, incorporating denoising text query training yields an additional 1.7 $AP_{50}^{Novel}$ performance gain (#3), demonstrating its effectiveness in enhancing the discriminability between novel categories and backgrounds. Finally, RoQIs selection contributes an additional 0.6 $AP_{50}^{Novel}$ (#5).

**Iterations of Open-World Pseudo-Labeling.** Table 4(a) presents the ablation study on pseudo-labeling iterations. We calculated the recall for objects in the COCO training set after each pseudo-labeling iteration for reference. The experimental results indicate that OV-DQUO achieves optimal performance when iterations equal to 2. Although recall improves with additional iterations, the introduced noise begins to impair model performance on novel categories.

**Choice of Wildcard Text.** As shown in Table 4(b), we investigate the impact of matching various wildcard texts with open-world unknown objects, including "Salient Object", "Target", "Thing", and "Object". Experimental results show that, compared to intricate wildcards ("Salient Object"), simple and general wildcards ("Thing", "Object") yield better results.

**Scaling Foreground Score.** Table 4(c) presents the ablation study on scaling the foreground score. We utilize the power function $(w_i)^\gamma$ to scale the foreground likelihood score for each unknown object, where $\gamma$ controls the degree of scaling. When $\gamma$ is set to 0, it functions as an ablation for the FE module. Results indicate that setting $\gamma$ to 0 significantly degrades performance due to the release of pseudo-label noise. The optimal performance is attained when $\gamma$ is set to 0.5.

**Denoising Loss Weight.** Table 4(d) presents the ablation study on the weight of the denoising loss $\beta$. Experimental results show that changing the weight of the denoising loss does not significantly affect performance. Moreover, the best results on novel categories are achieved when the denoising loss weight equals the classification loss weight, i.e., $\beta = 2$.

**Visualization Analysis of OV-DQUO.** We visualize the prediction results of OV-DQUO and the baseline detector (Wu et al. 2023b) on the OV-COCO benchmark, including their confidence score distributions and T-SNE results of output embeddings. As shown in Figure 3, compared to the baseline detector, OV-DQUO exhibits a more balanced prediction confidence distribution between novel and base classes. Furthermore, the confidence distribution predicted by OV-DQUO for both base and novel classes shows less overlap with the background distribution. As shown in Figure 4, the embeddings predicted by OV-DQUO exhibit superior separability from background embeddings compared to the baseline detector when detecting novel class objects.

## Conclusions

In this paper, we reveal that confidence bias constrains the novel category detection of existing OVD methods. Inspired by OWD tasks that identify unknown objects, we introduce an OV-DQUO framework to address this bias, which achieves new state-of-the-art results on various OVD benchmarks. Although integrating OVD with OWD into a unified end-to-end framework is promising, it remains underexplored here and reserved for future research.

## Acknowledgments

This work was supported by the Science and Technology Project of Shenzhen (GXWD-20220811170603002, KJZD-20230923114600002), the China Postdoctoral Science Foundation (2024MD754244), the General Program of the Natural Science Foundation of Chongqing (CSTB2024NSCQ-MSX047), and the Chongqing Postdoctoral Foundation Special Support Program (2023CQBSHTB3119).

## References

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35: 33781–33794.

Chen, F.; Zhang, H.; Yang, Z.; Chen, H.; Hu, K.; and Savvides, M. 2024. RTGen: Generating Region-Text Pairs for Open-Vocabulary Object Detection. *arXiv preprint arXiv:2405.19854*.

Chen, P.; Sheng, K.; Zhang, M.; Lin, M.; Shen, Y.; Lin, S.; Ren, B.; and Li, K. 2022. Open vocabulary object detection with proposal mining and prediction equalization. *arXiv preprint arXiv:2206.11134*.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14084–14093.

Fang, R.; Pang, G.; Zhou, L.; Bai, X.; and Zheng, J. 2023. Unsupervised Recognition of Unknown Objects for Open-World Object Detection. *arXiv preprint arXiv:2308.16527*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Gupta, A.; Narayan, S.; Joseph, K.; Khan, S.; Khan, F. S.; and Shah, M. 2022. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9235–9244.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Jeong, J.; Park, G.; Yoo, J.; Jung, H.; and Kim, H. 2024. ProxyDet: Synthesizing Proxy Novel Classes via Classwise Mixup for Open-Vocabulary Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2462–2470.

Joseph, K.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5830–5840.

Kim, D.; Angelova, A.; Kuo, W.; Kuo, W.; and Kuo, W. 2023a. Contrastive Feature Masking Open-Vocabulary Vision Transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15556–15566.

Kim, D.; Angelova, A.; Kuo, W.; Kuo, W.; and Kuo, W. 2023b. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11144–11154.

Kim, D.; Lin, T.-Y.; Angelova, A.; Kweon, I. S.; and Kuo, W. 2022. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2): 5453–5460.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2023. Open-Vocabulary Object Detection upon Frozen Vision and Language Models. In *The Eleventh International Conference on Learning Representations*.

Li, L.; Miao, J.; Shi, D.; Tan, W.; Ren, Y.; Yang, Y.; and Pu, S. 2023a. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6501–6510.

Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; and He, K. 2023b. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23390–23400.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Ma, C.; Jiang, Y.; Wen, X.; Yuan, Z.; and Qi, X. 2024. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36.

Minderer, M.; Gritsenko, A.; Houlsby, N.; and Houlsby, N. 2024. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Song, H.; and Bang, J. 2023. Prompt-guided transformers for end-to-end open-vocabulary object detection. *arXiv preprint arXiv:2303.14386*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Wang, L.; Liu, Y.; Du, P.; Ding, Z.; Liao, Y.; Qi, Q.; Chen, B.; and Liu, S. 2023. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11186–11196.

Wu, S.; Zhang, W.; Jin, S.; Liu, W.; and Loy, C. C. 2023a. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15254–15264.

Wu, S.; Zhang, W.; Xu, L.; Jin, S.; Li, X.; Liu, W.; and Loy, C. C. 2024a. CLIPSelf: Vision Transformer Distills Itself for Open-Vocabulary Dense Prediction. In *The Twelfth International Conference on Learning Representations*.

Wu, S.; Zhang, W.; Xu, L.; Jin, S.; Liu, W.; and Loy, C. C. 2024b. Clim: Contrastive language-image mosaic for region representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6117–6125.

Wu, X.; Zhu, F.; Zhao, R.; and Li, H. 2023b. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7031–7040.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, 106–122. Springer.

Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14393–14402.

Zhai, H.; Cheng, J.; and Wang, M. 2020. Rethink the IoU-based loss functions for bounding box regression. In *2020 IEEE 9th joint international information technology and artificial intelligence conference (ITAIC)*, volume 9, 1522–1528. IEEE.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.; and Shum, H.-Y. 2023. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *The Eleventh International Conference on Learning Representations*.

Zhang, H.; Zhao, Q.; Zheng, L.; Zeng, H.; Ge, Z.; Li, T.; and Xu, S. 2024. Exploring Region-Word Alignment in Built-in Detector for Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16975–16984.

Zhao, S.; Schulter, S.; Zhao, L.; Zhang, Z.; G, V. K. B.; Suh, Y.; Chandraker, M.; and Metaxas, D. N. 2024. Taming Self-Training for Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13938–13947.

Zhao, S.; Zhang, Z.; Schulter, S.; Zhao, L.; Vijay Kumar, B.; Stathopoulos, A.; Chandraker, M.; and Metaxas, D. N. 2022. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, 159–175. Springer.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16793–16803.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368. Springer.

Zhu, C.; and Chen, L. 2023. A survey on open-vocabulary detection and segmentation: Past, present, and future. *arXiv preprint arXiv:2307.09220*.

Zohar, O.; Wang, K.-C.; and Yeung, S. 2023. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11444–11453.