


# MarvelOVD: Marrying Object Recognition and Vision-Language Models for Robust Open-Vocabulary Object Detection

Kuo Wang<sup>1</sup>, Lechao Cheng<sup>2\*</sup>, Weikai Chen<sup>3</sup>, Pingping Zhang<sup>4</sup>, Liang Lin<sup>1,5</sup>, Fan Zhou<sup>6</sup>, and Guanbin Li<sup>1,5\*</sup>

<sup>1</sup> School of Computer Science and Engineering, Sun Yat-sen University

<sup>2</sup> Hefei University of Technology

<sup>3</sup> Tencent America

<sup>4</sup> Dalian University of Technology

<sup>5</sup> Peng Cheng Laboratory

<sup>6</sup> School of Computer Science and Engineering, National Engineering Research Center of Digital Life, Sun Yat-sen University

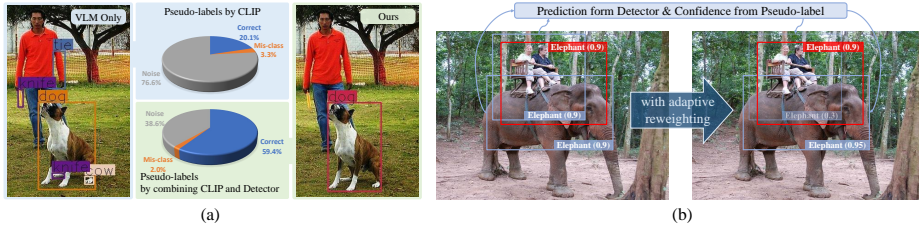
wangk229@mail2.sysu.edu.cn, chenglc@hfut.edu.cn, chenwk891@gmail.com, zhpp@dlut.edu.cn, linliang@ieee.org, {isszf, liguanbin}@mail.sysu.edu.cn

**Abstract.** Learning from pseudo-labels that generated with VLMs (Vision Language Models) has been shown as a promising solution to assist open vocabulary detection (OVD) in recent studies. However, due to the domain gap between VLM and vision-detection tasks, pseudo-labels produced by the VLMs are prone to be noisy, while the training design of the detector further amplifies the bias. In this work, we investigate the root cause of VLMs’ biased prediction under the OVD context. Our observations lead to a simple yet effective paradigm, coded MarvelOVD, that generates significantly better training targets and optimizes the learning procedure in an online manner by marrying the capability of the detector with the vision-language model. Our key insight is that the detector itself can act as a strong auxiliary guidance to accommodate VLM’s inability of understanding both the “background” and the context of a proposal within the image. Based on it, we greatly purify the noisy pseudo-labels via Online Mining and propose Adaptive Reweighting to effectively suppress the biased training boxes that are not well aligned with the target object. In addition, we also identify a neglected “base-novel-conflict” problem and introduce stratified label assignments to prevent it. Extensive experiments on COCO and LVIS datasets demonstrate that our method outperforms the other state-of-the-arts by significant margins. Codes are available at <https://github.com/wkfdb/MarvelOVD>.

**Keywords:** Pseudo Labeling · Open Vocabulary Object Detection

---

\* Corresponding authors.



**Fig. 1:** Improvements achieved by incorporating the detector for pseudo-label generation and the following learning phase. **(a)** The distribution of pseudo-labels generated by CLIP and our method. “Mis-class” means boxes labeled as wrong categories and “noise” indicates boxes that should not be considered as pseudo-labels. The VLM (CLIP) has low “mis-class” rate but fails to distinguish noisy boxes. Our method discriminates the noises by combining the characteristics of the detector, and hence significantly improves the quality of the pseudo labels. **(b)** The red box indicates pseudo-label and the blue boxes represent the matched training boxes. Adaptive proposal reweighting computes independent weights according to the prediction of detector and the confidence from pseudo-label, leading the training to focus on more reliable instances (e.g. the lower right training box).

## 1 Introduction

Open-vocabulary object detection (OVD) [35] is receiving increasing attention due to its capability of detecting novel objects at test time. In a typical OVD setting, only a fraction of the target categories is annotated (referred as the base categories) while the goal of OVD is to recognize a set of novel classes at inference time. The objects of novel categories can appear in the training images but do not receive any annotations. To enhance the generalizability of the OVD detector, recent works have proposed to incorporate the vision-language models (VLMs) [13, 24, 36], which have been verified with excellent zero-shot recognition capacity, to improve the existing OVD pipeline.

A common practice for the OVD task with known novel concepts is to generate pseudo-labels using the VLMs (e.g. CLIP [24]) in an offline manner [38]. However, because of the domain shift between the contrastive language-image pretraining and object detection tasks, VLMs trained with image-level data inevitably introduce noisy annotations when applied to the cropped partial images. We demonstrate an in-depth analysis of the pseudo-labels generated by the CLIP model in Figure 1(a). We denote the valid novel object proposal mistakenly classified as other categories as “*Mis-class*” and boxes that should not be considered to contain a novel object as “*Noise*”. In fact, the mis-classification rate of the VLM-based method is rather low (only 3.3%). The main source of error stems from its incapability of distinguishing “noisy” boxes (error rate 76.6%), e.g., the dog leg in Figure 1(a), that should not be considered as a valid object of interest.

The key reasons for the difficulty of the VLMs to recognize noisy proposals are two-fold. 1) The lack of contextual information to understand the locally cropped images. Instead of trained with image patches, the CLIP model is fed

with complete images with paired texts. Therefore, it is not able to leverage the image context outside the input proposal, which may be crucial to interpret the semantics of the candidate box. For instance, in Figure 1(a), the man’s arm is falsely classified by the CLIP model as “tie”, as it fails to recognize that the seemingly “tie” object is in fact connected to a human body. 2) The unawareness of the “background” elements. The CLIP model generates the category prediction by computing the similarity between the query image feature and the text embedding of the candidate categories. Since “background” is relatively defined according to interested foreground categories, there is no pre-defined text embedding to represent the concept of “background” during the inference. However, the CLIP model still has to provide a prediction even when the input content is not related to any of the target categories. In Figure 1(a), the dog leg falls in this case – it is identified as “cow” only because it appears more like cow than any other category, leading to noisy boxes.

Unlike the CLIP model, the RoI align technique in detectors naturally provides rich contextual information for local regions. Moreover, the detector is aware of the concept of “background” during inference. As a result, the noisy boxes that confuse VLMs can be clarified by the detector as “background” with high confidence. Inspired by this key observation, we propose MarvelOVD, a dedicated framework for open-vocabulary detection that can yield high-quality pseudo-labels and marvelous performance by combining the merits of the object detector and the vision-language models. In particular, we leverage the “context and background” awareness of the detector as strong auxiliary guidance to comprehensively improve the *pseudo-label generation* pipeline and the *training* procedure.

For pseudo-label generation, the predicted category confidence is based on a weighted sum of the outputs from the detector and the VLM, favoring the reliable classification of the VLM models while ruling out noisy boxes using the detector. To accelerate the training, we pre-generate the VLM predictions on all the candidate boxes and dynamically mine credible pseudo-labels under the guidance of the detector at each training iteration. The complementary capabilities of the detector and the VLM significantly improve the accuracy of pseudo-labels, even at the early training stage. Moreover, as the detector improves during training, the quality of the generated pseudo-labels increases as well, which eventually boosts the final performance as shown in Figure 1(a).

Conventional training design of object detectors [10] equally treats each proposal that matches with one training target. Such a design is not suitable for learning from pseudo-labels. Specifically, the generated pseudo-box may deviate a lot from the bounding box of the real novel object. Therefore, as shown in Figure 1(b), the overlaps between the training boxes and the actual novel object usually present a large variance. This means that these training boxes should not make equal contributions to the final loss, even if they match the same pseudo-label. To this end, instead of weighting the pseudo-labels [38], we adaptively compute individual weights for each training box that is matched with a pseudo-label. As shown in Figure 1(b), training boxes with inaccurate positions

will receive smaller weights and vice versa. Note that the training boxes are produced with a stratified label assignment strategy, which eliminates the conflicts between the pseudo-labels and base annotations and thus prevents the negative influence of noisy pseudo-label on the performance of base category detection.

## 2 Related Work

*Vision-Language Pre-training* Vision-Language pre-training, aiming to align visual and textual representations, employs contrastive learning on large-scale image-caption pairs [6, 12, 14]. Significant research has enhanced various downstream tasks using Vision-Language models [15, 16, 23]. Notably, models like CLIP [24] and ALIGN [13] have leveraged billion-scale image-text pairs for vision-language representation learning, achieving remarkable success in zero-shot image classification and image-text retrieval. This success has inspired the application of Vision-Language Models (VLMs) in enhancing the range and accuracy of dense recognition tasks, such as object detection [8, 26, 28, 30] and semantic segmentation [25, 33, 40]. Nonetheless, VLMs, typically pre-trained by considering the entire image, face domain gaps in dense prediction tasks. Our research investigates the application of VLMs for object detection in localized regions, aiming to broaden the detector’s cognitive scope without relying on manual annotations.

*Open Vocabulary Object Detection* Open Vocabulary Object Detection (OVD) aims to extend the detector’s recognition capability to classes not present in the training data using auxiliary data or models. The concept was initially introduced by OVR-CNN [35], which empowered the detector to recognize diverse object concepts by leveraging vision and text encoders pre-trained on image-text pairs. Subsequently, several OVD methods have been developed. Recent research in OVD has explored various forms of auxiliary data, including transfer learning with image-text pairs [2, 3, 19], knowledge distillation from pre-trained Vision-Language Models [1, 4, 8, 28, 29], pseudo-label generation from image classification data [1, 41], and pretraining with grounding data [18, 37]. Except for CNN-based detectors, transformer-based open vocabulary detectors [17, 30, 34] have also been widely exploited. In addition to the standard OVD scenario, which assumes that novel categories are unknown during training, some existing works [5, 38] have also explored open vocabulary detection with prior knowledge about potential novel concepts. Typically, in such cases, pseudo-labels for novel categories are generated using Vision-Language Models (VLMs) before training. However, the domain gaps between vision-language pre-training and object detection introduce noisy pseudo-labels, significantly constraining the performance of existing methods. To address this challenge, we identify a critical limitation in applying VLMs to localized regions and propose a solution that involves integrating the detector’s capabilities to effectively mitigate this noise, leading to substantially improved performance.

### 3 Preliminaries

Open-vocabulary object detection aims to train detection models by leveraging a dataset  $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$  and auxiliary weakly supervised data (e.g. image-text pairs, VLMs, etc.), where  $\mathbf{x}_i$  represents the image and  $y_i$  includes the location and category of the objects contained in the image. Different from conventional detection tasks, the annotation of images only covers the base categories  $\mathcal{C}^B$ , while the OVD task requires the detector to additionally detect novel categories  $\mathcal{C}^N$  at the test time. Note that  $\mathcal{C}^B \cap \mathcal{C}^N = \emptyset$ , and the label space  $\mathcal{C}^N$  is already known during the training.

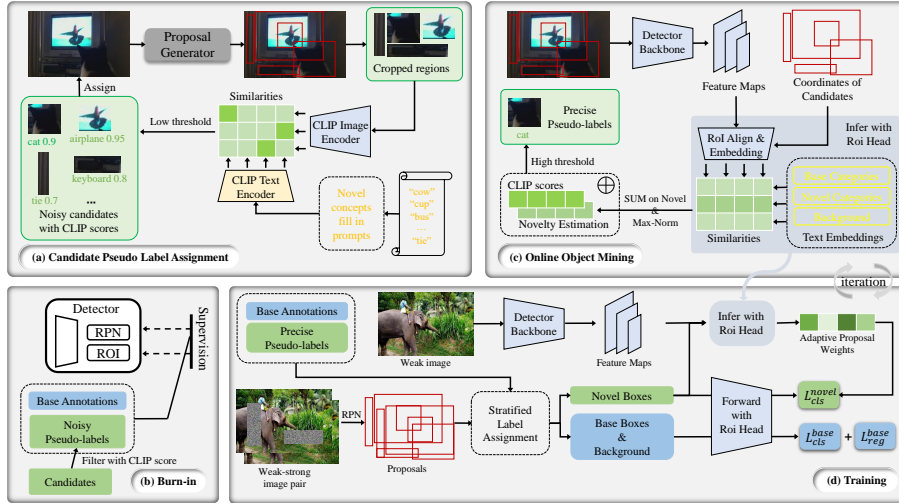
To achieve the detection in an open-vocabulary label space, the classification head in the detector is designed to compare the similarity between the region-based visual embeddings and the text embeddings [35]. In particular, the region-embeddings  $\mathbf{R} = \{\mathbf{r}_i\}_{i=1}^{N_r}$  are obtained through RoI Align and the following feature extractor, where  $N_r$  is the number of regional boxes in the images. The text embeddings are composed of  $\mathbf{C} = \{\mathbf{c}_{bg}\} \cup \{\mathbf{c}_i\}_{i=1}^{N_c}$ , where  $\mathbf{c}_i$  is obtained by feeding the category name with template prompts (e.g., “a photo of {category name} in the scene”) to the pre-trained text encoder and  $N_c$  is number of categories. The  $\mathbf{c}_{bg}$  is initialized as a learnable embedding. Based on the regional and text embeddings, the probability of the region  $\mathbf{r}_i$  being classified as category  $\mathbf{c}_j$  is defined as

$$p_{i,j} = \frac{\exp(\mathbf{r}_i \cdot \mathbf{c}_j)}{\exp(\mathbf{r}_i \cdot \mathbf{c}_{bg}) + \sum_{k=1}^{N_c} \exp(\mathbf{r}_i \cdot \mathbf{c}_k)}. \quad (1)$$

Comparing the similarity between the region and text embedding enables the detector to recognize objects in an unlimited label space.

### 4 MarvelOVD

To facilitate the learning of semantics associated with open categories without manual annotation, existing approaches often employ pre-trained Vision-Language Models (VLMs) to discover potential novel objects [5, 7, 38, 41] and generate pseudo-labels for subsequent training. The typical procedure involves training a proposal generator using base annotations to identify localized regions that may contain novel objects, followed by the generation of pseudo-labels based on VLM inference results within these cropped regions. However, as VLMs are pre-trained on entire images, their application to localized regions inevitably introduces noisy pseudo-labels, leading to disruptions in the learning process for novel categories. To enhance the learning of novel concepts, we present MarvelOVD, which dynamically integrates the detector’s capabilities into the pseudo-label generation process while optimizing the subsequent learning stages. Figure 2 provides an overview of our framework, with detailed explanations in the following sections.



**Fig. 2:** The framework of our method, which improves the quality of pseudo-labels while optimizing the following learning process by dynamically incorporating the detector during the training. We first assign candidate boxes to the images with CLIP and a proposal generator. Then we select noisy pseudo-labels according to the CLIP scores to burn-in the detector. After burn-in, the detector initially obtains the capacity to recognize novel concepts. Based on it, we dynamically estimate the novelty of each candidate box and combine the corresponding CLIP prediction to select precise pseudo-labels. We adopt stratified label assignment to generate training boxes, while the loss weights for the novel training boxes are independently computed based on the detector’s prediction.

#### 4.1 Candidate Pseudo-Label Assignment

Our proposed MarvelOVD dynamically generates better pseudo-labels at each training iteration under the guidance of both the detector and pre-trained VLMs. Since predicting the cropped regions with VLMs during training requires unaffordable time overhead, we alternatively assign candidate pseudo-labels to each image before the training and then select precise ones to train the detector at each iteration.

Following the existing method [38], we first train a class-agnostic proposal generator with the base annotation to produce regional boxes for each image. Image patches are then cropped according to the regions and fed into the CLIP image-encoder to obtain the regional-visual-embeddings. At the same time, we utilize the corresponding CLIP text-encoder and template text prompts to encode each novel category. After that, a similarity matrix is computed via dot product to describe the similarity between each visual-embedding and text-embedding. Finally, softmax is applied to obtain the distribution over novel categories for each region. Based on it, the conventional method [7,38] post-processes the boxes and selects high-confidence pseudo-labels to train the detector. The

problem is that the domain gap easily leads CLIP to produce high confidence predictions on noisy regions, which greatly limits the performance of the existing methods. In contrast, we record the CLIP predictions and assign candidate boxes to the image using a low threshold (e.g. 0.5), and then dynamically select precise pseudo-labels from the candidates under the guidance of the detector, which is introduced in detail in the following section.

## 4.2 Online Pseudo-label Mining

The candidate pseudo-labels can be categorized into two groups: ground-truth boxes (those tightly enclosing actual novel objects) and noisy boxes (those that should not be designated as novel objects). The primary objective is to eliminate noisy candidates while retaining ground-truth boxes as reliable pseudo-labels for training. The key distinction between the detector and CLIP in their inference of localized regions lies in the contextual information and the “background” concept. The RoI Align mechanism employed by the detector excels at extracting contextual features from the boxes, a capability that is absent when cropping images based on coordinates alone. Additionally, the detection task incorporates a specific task-specific category known as “background”, a concept that CLIP remains unaware of when inferring localized regions. Failing to recognize the “background” objects and the lack of contextual features are the root causes of CLIP model generating wrong predictions for noisy boxes. Though the CLIP model struggles with the noisy boxes, its predictions for the ground-truth boxes are highly accurate. Leveraging this insight, we propose to utilize the detector’s predictions for these candidates to estimate whether a box encloses real novel objects. Subsequently, we combine CLIP’s classification results to select high-quality pseudo-labels.

*Burn-in.* To estimate the novelty of the candidate boxes, the detector first needs to learn “what is novel”. To achieve this, we first utilize the top-1 CLIP score (top score of the distribution predicted with CLIP’s image-encoder and text-encoder) and a fixed threshold 0.8 (best threshold in previous work [38]) to initially select pseudo-labels to burn-in the detector for  $\omega$  steps. After the burn-in phase, the model will initially gain the ability to distinguish between base objects, novel objects, and the background.

*Online Object Mining.* Online object mining officially begins after the burn-in phase. We draw ideas from semi-supervised learning [27] to derive weak-strong image pairs for the training, which enhances the learning for pseudo-labels. In particular, we first predict the candidate boxes with the detector on weakly augmented features. Based on it, we compute a novelty score for each candidate as follows:

$$z_i = \frac{\sum_{k \in \mathcal{C}^N} \exp(\mathbf{r}_i \cdot \mathbf{c}_k)}{\sum_{j \in \mathcal{C}^B \cup \mathcal{C}^N \cup \{\mathbf{c}_{bg}\}} \exp(\mathbf{r}_i \cdot \mathbf{c}_j)} \quad (2)$$

where  $\mathbf{r}$  is the vision-embedding calculated by the detector and  $\mathbf{c}$  is the text-embedding of categories.  $\mathcal{C}^B$  and  $\mathcal{C}^N$  are the sets of base and novel categories,

respectively. The novelty score  $z_i$  relatively estimates the novelty of candidate boxes with respect to the base category and background. However, its value varies drastically with different degrees of convergence. To tackle this problem, we further apply max-norm to the novelty scores of candidates to obtain stable estimations:

$$s_i^{det} = \frac{z_i}{\max\{z_1, z_2, \dots, z_{N_r}\}} \quad (3)$$

where  $N_r$  represents the number of the candidates. Benefiting from the contextual reasoning capacity of the detector and the awareness of background, the novelty estimation  $s^{det}$  computed by the detector can more precisely distinguish the ground-truth/noise candidates. Combined with the accurate classification prediction generated by the CLIP model, we finally calculate the confidence score for each candidate box as follows:

$$s_i = \lambda s_i^{CLIP} + (1 - \lambda) s_i^{det} \quad (4)$$

In the above equation,  $s_i^{CLIP}$  means the top-1 CLIP score and  $\lambda \in [0, 1]$  is a scalar that controls the dependency of two different models. We utilize a fixed threshold  $\delta$  to select high-quality pseudo-labels. The training is then derived on both weakly and strongly augmented images.

The incorporation of the detector significantly reduces the confidence of the noisy candidates, which greatly improves the accuracy of the selected pseudo-labels, even at the initial training phase. Moreover, as the model converges with the training, the novelty estimation  $s^{det}$  would be more accurate, which results in pseudo-labels of higher quality and ultimately boosts the model’s detection performance on novel categories.

### 4.3 Training

In this section we describe our improvements to the conventional training design of the detector [10]. All the proposed methods in this section are applied both in the burn-in stage and the following online-object-mining stage.

*Stratified Label Assignment.* The learning of novel concepts should not affect the model’s performance in recognizing base objects. However, an easily overlooked phenomenon is that the mAP for base categories drops when novel pseudo-labels are applied for the training. The reason is that the novel pseudo-labels may overlap with the base annotation, resulting in “base-novel-conflicts” in the IoU-based label assignment. To tackle this problem we propose stratified label assignment, which first assigns proposals with base annotations by IoU-matching, and boxes that are marked as background in the first step are secondly matched with the pseudo-labels. Experiments demonstrate that stratified label assignment helps achieve high accuracy of detecting novel objects without compromising the performance on estimating base categories.



*Adaptive Proposal Reweighting.* Since the localization quality of the pseudo-label is limited, the box center may be far away from the ground-truth object center. As a result, training boxes that matched with the mislocalized pseudo-label share extremely unbalanced overlaps with the ground-truth object. However, the conventional training design of detectors [10] equally derives training losses on those unbalanced boxes, which hampers the learning process. To resolve this issue, we propose adaptive proposal reweighting to assign independent loss weights to each training box that matches with pseudo-label.

The loss function to train the detector with adaptive proposal reweighting is computed as:

$$\mathcal{L} = \frac{1}{N} \left( \sum_{i=1}^{n^{base}} l(b_i^{base}, \mathcal{G}^{base}) + \gamma \sum_{i=1}^{n^{novel}} w_i \cdot l(b_i^{novel}, \mathcal{G}^{novel}) \right) \quad (5)$$

where  $N = n^{base} + n^{novel}$  ( $n^{base}$  includes background box) is the total number of training boxes,  $\gamma$  is the overall weight for novel concept learning and  $w_i$  represents the independent weights for each novel training box. In particular, we follow the design of Eq. 4 to define the individual weight  $w_i$  as:

$$w_i = \lambda' s_i + (1 - \lambda') r_i \quad (6)$$

In Eq. 6,  $s_i$  indicates the confidence of the corresponding pseudo-label and  $r_i$  is a reliability score estimated for each matched training box. Estimating the reliability score  $r_i$  is crucial and challenging. We empirically find that background score predicted on the weakly augmented images keeps close negative correlation to the overlaps with actual object and define  $r_i = 1 - b_i$ , where  $b_i$  is the background score predicted according Eq 1. We also examine other indicators for comparison, more details are shown in supplementary materials. With adaptive reweighting, training boxes with higher overlap to real novel objects will be given greater weights and vice versa, thus de-biasing the learning procedure for novel concepts and further improving the performance.

## 5 Experiments

In this section, we evaluate our MarvelOVD framework against standard benchmarks, comparing it with current state-of-the-art approaches. Additionally, our ablation studies provide in-depth analyses of the primary issues leading to noises in traditional CLIP-based pseudo-label generation and detail how our framework effectively addresses these challenges.

### 5.1 Datasets

Our primary experiments utilize the COCO-2017 dataset [21] in an open vocabulary setting [35], dividing 48 base and 17 novel categories for evaluation. Annotations for base categories are provided, while only category names are

available for novel classes. We calculate  $AP_{50}^{Novel}$ ,  $AP_{50}^{Base}$ ,  $AP_{50}^{All}$ , representing the mean Average Precision at an IoU of 0.5 for novel, base, and all categories, respectively. Additionally, the LVIS-v1 dataset [9] is employed in standard Open Vocabulary Detection (OVD) settings [8], treating 337 rare categories as novel and the rest as base. For LVIS, we report box Average Precision (AP) averaged over IoUs from 0.5 to 0.95 for rare (novel), common, frequent, and all categories, denoted as  $AP_r$ ,  $AP_c$ ,  $AP_f$ , and  $AP$ .

## 5.2 Implementation Details

We utilize ViT-B/32 CLIP as the pre-trained Vision-Language Model (VLM) and its text-encoder for encoding category concepts. Consistent with existing approaches [38], our experiments on the COCO dataset employ Mask-RCNN [10] with ResNet50-FPN [11,20] as the base detector. For training, we initially select noise pseudo labels with CLIP scores above 0.8 and use this setup for the burn-in phase for  $\omega = 0.5k$  iterations. Subsequently, we set  $\lambda, \lambda' = 0.5$  to integrate the detector with CLIP and  $\delta = 0.9$  for generating precise pseudo-labels. We set  $\gamma = 2$  as the overall weights for novel concept learning. Training is conducted on 4 GPUs, with a total batch size of 16 across 90k iterations (including 0.5k for burn-in), using a starting learning rate of 0.02, reduced by a factor of 10 at 60k and 80k iterations. The image input size adheres to the standard configuration, with the short side ranging from [640, 800] and the long side under 1333. Additionally, we apply common weak-strong augmentations from semi-supervised object detection literature [22,32] in pseudo-label learning. For the LVIS dataset, we replicate Detic’s experimental setup [41] and apply our method to the CenterNet2 [42] baseline. The model is trained on 4 GPUs, while maintaining the total batch size unchanged. All experiments are conducted in Detectron2 [31], with further details provided in the supplementary materials.

## 5.3 Main Results

Our approach integrates the detector to refine pseudo-label generation and the subsequent learning phase, significantly reducing noises in pseudo-labels and training boxes. As shown in Table 1, our method outperforms the baseline method [38] substantially in inferring both base and novel categories. The improvement in base categories stems from stratified label assignment, ensuring undisturbed learning of base categories despite pseudo-labels. The detector’s integration effectively mitigates CLIP’s inability in distinguishing noise in localized regions, enhancing pseudo-label quality and, in turn, the detector’s ability to identify novel objects. Our method is also compared with other state-of-the-art open-vocabulary detection techniques utilizing pseudo-labeling in Table 1. While existing methods often rely on auxiliary data or supervision, like internet-sourced image-text pairs [5, 7], pseudo-region-text pair pre-training [39], or auxiliary image-level labels [1], our approach addresses and resolves fundamental issues in pseudo-label generation and conventional training designs [10], achieving significant gains without extra data or supervision. Results on the LVIS

**Table 1:** Comparison with state-of-the-art methods on COCO2017 dataset.

Methods	Training Source	$AP_{50}^{Novel}$	$AP_{50}^{Base}$	$AP_{50}^{All}$
RegionCLIP [39]	box-level labels in $\mathcal{C}_B$ , internet sourced image-text pairs, pretraining with pseudo box-level labels	31.4	57.1	50.4
Gao <i>et al.</i> [7]	box-level labels in $\mathcal{C}_B$ , internet sourced image-text pairs, pseudo-box labels in $\mathcal{C}_N$ generated by ALBEF	30.8	46.1	42.1
PromptDet [5]	box-level labels in $\mathcal{C}_B$ , internet sourced image-text pairs, pseudo-box labels in $\mathcal{C}_N$ generated by CLIP	26.6	-	50.6
OADP [28]	box-level labels in $\mathcal{C}_B$ , knowledge distillation from CLIP, pseudo-box labels in $\mathcal{C}_N$ generated by CLIP	35.6	55.8	50.5
Rasheed <i>et al.</i> [1]	box-level labels in $\mathcal{C}_B$ , internet sourced image-text pairs, image-level labels for $\mathcal{C}_B \cup \mathcal{C}_N$ , pseudo-box labels in $\mathcal{C}_N$	36.6	54.0	49.4
VL-PLM [38]	box-level labels in $\mathcal{C}_B$ ,	32.3	54.0	48.3
MarvelOVD(Ours)	box-level pseudo-labels in $\mathcal{C}_N$ generated with CLIP	<b>38.9</b>	56.4	51.8

**Table 2:** Comparison with state-of-the-art methods on LVIS-v1 dataset. All the methods are derived under the same base detector and experimental settings.

Methods	$AP_r$	$AP_c$	$AP_f$	AP
VLDet [19]	22.4	-	-	34.4
Detic [41]	24.6	32.5	35.6	32.4
Rasheed <i>et al.</i> [1]	25.2	33.4	35.8	32.9
MarvelOVD(Ours)	<b>26.0</b>	34.2	36.9	34.2

dataset (Table 2) compares our method against the common CenterNet2 baseline [41]. Contrary to existing methods [1, 41] that use additional classification data with image-level labels for enhanced novel object detection, our method exploits potential novel objects from original training data. The results in Table 2 indicate that our method also excels in large-scale label spaces.

## 6 Ablation Study

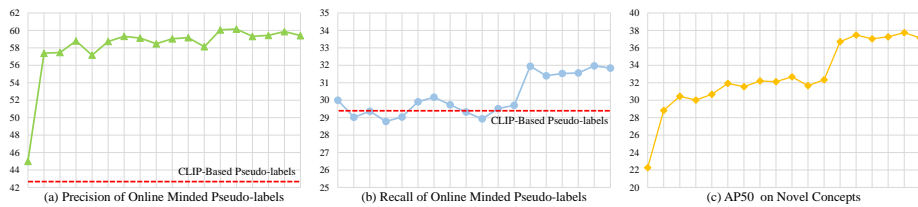
We conduct experiments on the COCO dataset to assess the effectiveness of our method’s key components, with additional ablations detailed in supplementary materials.

### 6.1 Effects of Each Component

Table 3 demonstrates the contribution of each proposed algorithmic component to the final performance. Initially, we change the global training setting of VL-PLM [38], including overall novel loss weight  $\gamma$  and the data augmentations.

**Table 3:** Roadmap from existing method to our framework.

	$AP_{50}^{Novel}$	$AP_{50}^{Base}$	$AP_{50}^{All}$
Supervised by base annotations	-	56.4	-
VL-PLM [38]	32.7	54.0	48.5
VL-PLM(set $\gamma = 2$ ) [38]	32.5	54.0	48.4
+Weak-Strong augmentation	34.2	53.9	49.1
+ <i>Stratified Label Assignment</i>	34.4	56.4 $\uparrow$	50.5
+ <i>Online object Mining</i>	37.8 $\uparrow$	56.5	51.3
+ <i>Adaptive Proposal Reweighting</i>	38.9 $\uparrow$	56.6	51.8

**Fig. 3:** Visualization of the quality of our dynamically generated pseudo-labels, with red dashed lines indicating the quality of the original CLIP-based pseudo-labels.

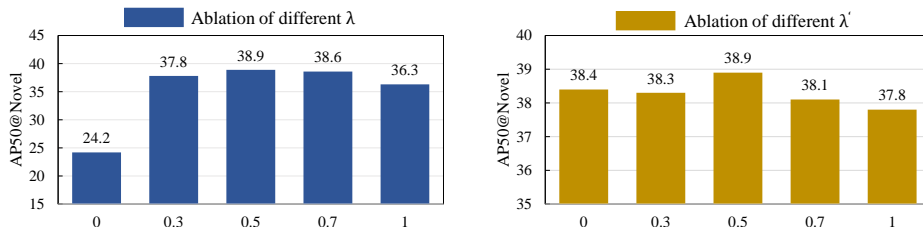
A larger novel loss weight doesn’t affect the performance while Weak-Strong augmentations enhance the learning of pseudo-labels. We then implement stratified label assignment to resolve conflicts between novel pseudo-labels and base annotations. This adjustment restores base category detection to supervised performance levels without impacting novel category detection. After that, we introduce online object mining to purify the pseudo labels, which brings a significant improvement in detection accuracy on the novel categories, indicating the effectiveness of our method in offsetting CLIP’s localized limitations by leveraging the detector’s capabilities. Based on it, applying adaptive proposal reweighting further enhances the average precision for novel categories. The promotion comes from the independent weights computed by adaptive reweighting, which enforces the model to focus on boxes with larger overlaps with actual novel objects. In summary, our method offers a less biased pipeline for pseudo-label-based novel concept learning, which not only effectively purifies the training targets but also optimizes the learning procedure, and significantly enhances performance in both base and novel categories without requiring extra data or pretraining.

## 6.2 Analysis of Pseudo Labels

We conduct an in-depth analysis of pseudo-labeling quality across different training stages using the COCO evaluation set, which comprises 2064 images featuring novel objects. Pseudo-labels with an IoU score above 0.5 relative to the ground

**Table 4:** Effects of different thresholds and burn-in steps. The default setting is  $\omega = 0.5k$  for burn-in and threshold  $\delta = 0.9$  for online pseudo-label mining.

$\delta$	0.8	0.85	0.9	0.95	$\omega$	0.5k	1k	2k	5k
$AP_{50}^{Novel}$	37.0	38.2	38.9	38.4	$AP_{50}^{Novel}$	38.9	38.7	38.7	38.5

**Fig. 4:** Effects of different dependence controller  $\lambda$  and  $\lambda'$ .

truth (GT) novel object and correctly categorized are classified as True Positives (TPs). The findings are depicted in Figure 3. For equitable comparison, we align the threshold of CLIP-based pseudo-labels ( $\delta = 0.95$ ) with our dynamic pseudo-labels ( $\delta = 0.8$ ) to maintain comparable recall rates. Initially, the burn-in stage imparts novel object discrimination ability, leading to a stable improvement in pseudo-label quality right after this phase. Subsequently, the enhanced pseudo-labels refine the detector’s ability to recognize novel objects, manifesting as a large increase in pseudo-label precision in the early stages of training. As illustrated in Figure 3, as training progresses, the detector increasingly differentiates novel boxes from background and base objects, dynamically enhancing pseudo-label quality.

### 6.3 Effects of different thresholds and burn-in steps

We evaluate the performance impact of various thresholds and burn-in steps, detailed in Table 4. Our base setting uses a threshold  $\delta = 0.9$  and  $\omega = 0.5k$  burn-in steps. Notably, the threshold for pseudo-label selection markedly affects model performance; while 0.8 is optimal for our baseline, a threshold of about 0.9 proves more effective for our method due to less biased pseudo-labeling confidence. Regarding burn-in steps, which guide initial learning from CLIP-generated pseudo-labels, their effect on final performance is minimal. As the model converges, the quality of these pseudo-labels improves, indicating that different initial settings eventually yield similar performance outcomes.

### 6.4 Dependency analysis of $\lambda$ and $\lambda'$

We calculate the confidence of each candidate pseudo-label using Eq. 4 and the training weights for each novel box via Eq. 6, where  $\lambda$  and  $\lambda'$  determines

the reliance on different models or measurements. The effects of varying  $\lambda$  and  $\lambda'$  are documented in Figure 4. Optimal performance is observed both at 0.5, and a range of [0.3, 0.7] yields comparable outcomes. Specifically, extreme values were also tested:  $\lambda = 0$  implies reliance solely on the detector post burn-in, which results in poor performance, underscoring the importance of CLIP’s role in distinguishing novel categories. Conversely, at  $\lambda = 1$ , pseudo-label generation reverts to conventional methods [5, 7, 38] that rely entirely on the CLIP model. While performance decreases in this setting, it still surpasses the baseline, suggesting that adaptive proposal reweighting effectively counters the impact of noisy boxes. By setting  $\lambda' = 1$ , adaptive reweighting changes to the original training design [10] with weighted pseudo-labels [38], and it limits the model’s learning of novel concepts, resulting in significant performance degradation.

## 7 Conclusion

In this paper, we address the limitations of pre-trained Vision-Language Models (VLMs) in generating accurate pseudo-labels for localized regions by integrating object detectors’ capabilities. The key issue with VLMs is their lack of contextual awareness and inability to differentiate “background”, leading to biased pseudo-labels. By leveraging the detector’s contextual feature extraction and background discrimination abilities, we significantly improve pseudo-label quality through online object mining and optimize the learning process with adaptive proposal reweighting. Our extensive experiments show that this approach not only enhances the detector’s novel object recognition but also outperforms state-of-the-art methods without additional data or supervision, offering an efficient and effective solution for learning open vocabulary concepts by pseudo-labels.

**Acknowledgments.** This work was supported in part by the Key-Area Research and Development Program of Guangdong Province (NO. 2021B0101420004), in part by the National Natural Science Foundation of China (NO. 62322608, NO. 62106235, NO. 62325605), in part by the Guangxi Science and Technology Plan Project (NO. GuikeAD23026034), in part by the Shenzhen Science and Technology Program (NO. JCYJ20220530141211024), and in part by the Open Project Program of the Key Laboratory of Artificial Intelligence for Perception and Understanding, Liaoning Province (AIPU, No. 20230003).

## References

1. Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems* **35**, 33781–33794 (2022)
2. Bravo, M.A., Mittal, S., Brox, T.: Localized vision-language matching for open-vocabulary object detection. In: *DAGM German Conference on Pattern Recognition*. pp. 393–408. Springer (2022)

3. Chen, P., Sheng, K., Zhang, M., Lin, M., Shen, Y., Lin, S., Ren, B., Li, K.: Open vocabulary object detection with proposal mining and prediction equalization. arXiv preprint arXiv:2206.11134 (2022)
4. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: 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. pp. 14084–14093 (2022)
5. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Prompt-det: Towards open-vocabulary detection using uncurated images. In: European Conference on Computer Vision. pp. 701–717. Springer (2022)
6. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems* **26** (2013)
7. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: European Conference on Computer Vision. pp. 266–282. Springer (2022)
8. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
9. 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)
10. 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)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Jayaraman, D., Grauman, K.: Zero-shot recognition with unreliable attributes. *Advances in neural information processing systems* **27** (2014)
13. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
14. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
16. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* **34**, 9694–9705 (2021)
17. Li, L., Miao, J., Shi, D., Tan, W., Ren, Y., Yang, Y., Pu, S.: Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6501–6510 (2023)
18. 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)

19. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. arXiv preprint arXiv:2211.14843 (2022)
20. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
21. 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: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
22. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
23. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **32** (2019)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
25. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091 (2022)
26. Shi, H., Hayat, M., Wu, Y., Cai, J.: Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9611–9620 (2022)
27. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* **33**, 596–608 (2020)
28. Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., Liu, S.: Object-aware distillation pyramid for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11186–11196 (2023)
29. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15254–15264 (2023)
30. Wu, X., Zhu, F., Zhao, R., Li, H.: 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. pp. 7031–7040 (2023)
31. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
32. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3060–3069 (2021)
33. Yun, S., Park, S.H., Seo, P.H., Shin, J.: Ifseg: Image-free semantic segmentation via vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2967–2977 (2023)



34. 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. Springer (2022)
35. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14393–14402 (2021)
36. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133 (2022)
37. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems* **35**, 36067–36080 (2022)
38. Zhao, S., Zhang, Z., Schuler, S., Zhao, L., Vijay Kumar, B., Stathopoulos, A., Chandraker, M., Metaxas, D.N.: Exploiting unlabeled data with vision and language models for object detection. In: European Conference on Computer Vision. pp. 159–175. Springer (2022)
39. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)
40. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
41. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision. pp. 350–368. Springer (2022)
42. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. arXiv preprint arXiv:2103.07461 (2021)