Learning Region-Word Alignment with Attentive Masking for Open-Vocabulary Object Detection

Masoumeh Zareapoor, Pourya Shamsolmoali, Yue Lu
Shanghai Jiao Tong University, East China Normal University
{yrui, ylu}@cee.ecnu.edu.cn

Abstract

Open-vocabulary object detection (OVDet) aims to detect novel categories based on textual descriptions, allowing models to generalize beyond the categories seen during training. However, achieving robust open-vocabulary detection poses significant challenges in aligning text descriptions with specific image regions and capturing spatial relationships between related regions. Most existing methods focus on aligning regions with categorical labels, often overlooking interactions between neighboring regions, limiting their ability to form a precise correspondence between text descriptions and image content. We propose AlignDet, which incorporates an attentive masking strategy to address these challenges. By masking irrelevant regions in the image, our model focuses on the most relevant areas for each text concept, leading to fine-grained region-word correspondences. Additionally, our soft association strategy allows multiple regions to align with a single text concept, capturing spatial relationships between neighboring or related regions of the image more effectively. Extensive experiments demonstrate that our model consistently surpasses existing methods across various benchmarks.

1 Introduction

Traditional object detection models are limited by their reliance on large-scale, class-specific annotations and their inability to generalize beyond a fixed set of predefined categories [26, 39, 43]. This dependence on predefined labels makes these models rigid and inflexible in real-world scenarios where the number and variety of object categories are vast and often unknown. Open-vocabulary object detection (OVDet) has emerged as a promising direction, moving beyond the closed-set paradigm of conventional object detection [1, 8, 40, 46]. OVDet leverages vision-language models (VLMs), which integrate language as a supervisory signal, reducing the reliance on comprehensive object annotations. This shift allows models to generalize beyond a fixed set of object categories and detect novel objects based on textual descriptions. Recent work has made progress by using pre-trained VLMs, such as CLIP and ALIGN [12, 16], which are trained on large-scale image-text pairs. For example, ViLD [8] uses embeddings from CLIP to perform object detection without the need for class-specific annotations. While these models perform well at recognizing general image-level features, they struggle with object-level detection, making them less effective for detecting small or occluded objects [33]. However, achieving robust open-vocabulary detection is not only to classify objects from textual descriptions but also to accurately align these descriptions with specific image regions. This requires fine-grained region-text alignment and the ability to capture spatial relationships between related regions. Recognizing the importance of region-level detection, GLIP and MDETR [14, 17] shifted the focus from image-level recognition to phrase grounding. These models integrate text and image features early in the detection pipeline, aligning text descriptions with specific regions in the image. However, these methods introduce increased model complexity, making them computationally intensive and difficult to scale for large datasets [21]. Other methods [6, 7, 11, 17, 37] attempted to enhance OVDet by using large-scale image-text pairs

38th Conference on Neural Information Processing Systems (OWA@NeurIPS 2024)

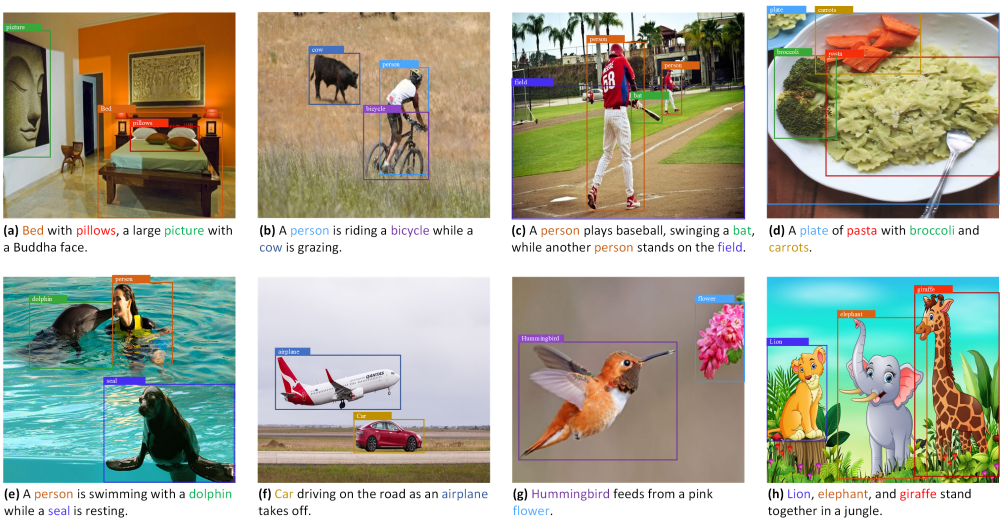

Figure 1: Region-word alignment for open-vocabulary object detection on CC12M [3]. The model aligns textual concepts with their respective regions, and detailed explanations are provided in 4.

from the web, generating pseudo-labels to provide region-level supervision where manual annotations are unavailable. However, the quality of these pseudo-labels is often limited, as the detectors that generate them are trained on a narrow set of human-annotated data. As a result, these models can struggle to generalize to unseen categories, and the high computational cost of processing large-scale, high-resolution images further limits their scalability. Despite advancements in phrase grounding and pseudo-labeling, achieving accurate region-word alignment remains a challenge. Recent efforts [20, 33, 36, 37] focus on directly improving region-word correspondence by aligning text concepts with image regions, enhancing detection accuracy across various scales. However, many of these methods still neglect the interactions between multiple regions related to a single text description, which is crucial for complex scenes where several regions may represent one object or concept.

We propose AlignDet, an end-to-end framework for open-vocabulary object detection that avoids the need for expensive annotations or distilling from classification-based models. The core idea is to use an attentive masking strategy that focuses on relevant image regions, ignoring irrelevant ones (see Fig. 2). Specifically, we treat image region features as one set and word embeddings as another, using dot-product similarity to compute region-word alignment scores. By using these scores, the model can determine the most relevant regions for each textual concept. Considering that a single text concept may correspond to multiple regions, our model allows each textual concept to be associated with all relevant regions. Our main contributions are: ❶ proposing AlignDet, an end-to-end open-vocabulary detection framework that leverages large image-text pairs; ❷ employing a unified training approach integrating grounding data [14], detection data [28], and image-text pairs [3]; ❸ we adopt an attentive masking strategy to select appropriate image regions for each textual concept, guiding the contrastive learning process and enabling precise alignment between visual and textual information. ❹ To capture inter-regional interactions that is often overlooked in prior work, our model uses a soft assignment strategy that handles multiple regions corresponding to a single textual concept. Our experiments on benchmark datasets validate the effectiveness of AlignDet. With a standard ATSS [43] detector and Swin-T [19] backbone, AlignDet achieves a decent zero-shot AP of 34.1% on the LVIS [9], outperforming DetCLIP and GLIPv2 by 20% and 16%, respectively, and showing competitive performance with VLDet (34.1%). In the rare categories, our model surpasses CODet with a +1.1 improvement, highlighting its strength in detecting underrepresented objects.

2 Related Work

Vision-language pre-training has gained significant traction by aligning image and text representations from large-scale image-text datasets [23, 27, 30, 44]. Recent advancements such as CLIP [23] and ALIGN [12] have shown impressive zero-shot performance in image classification. These vision-language models (VLMs) use contrastive learning [15] on large-scale image-text pairs from the

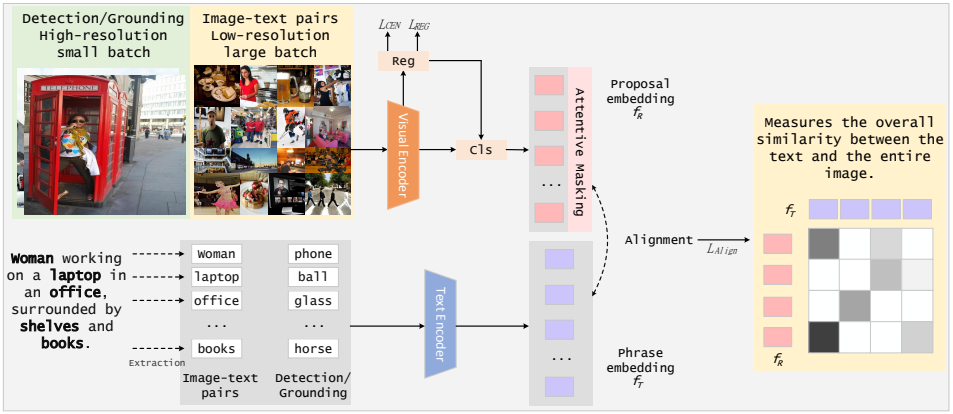

Figure 2: This architecture diagram illustrates open-vocabulary object detection using visual and textual information. The model employs a text encoder to convert textual descriptions into phrase embeddings (f_T), and a visual encoder to extract proposal embeddings (f_R), which identify potential regions of interest within the images from images. These embeddings are aligned through a matching process, facilitating contrastive learning and enabling effective word-region alignment.

web to align image and text embeddings. Inspired by the ability of VLMs, open-vocabulary object detection (OVDet) has emerged as a natural extension, aiming to recognize objects from unseen categories using image-text pairs. This trend marks a shift toward improving the generalization ability of object detectors by leveraging multi-modal knowledge. The early adoption of VLMs in OVDet was demonstrated by [40], which pretrained VLMs on image-caption pairs and transferred these features to a supervised object detector. This method laid the foundation for using VLMs in object detection. Further, ViLD [8] advanced the field by distilling the knowledge of CLIP into object detectors. By aligning the detector's visual embeddings with CLIP's image embeddings, ViLD enabled the recognition of novel object categories using text prompts.

Following these pioneering efforts, the focus shifted towards improving region-word alignment, a critical aspect of OVDet that enables models to associate textual descriptions with specific image regions accurately [18, 22, 32, 33]. Early methods by Zhong et al.[45] and Gao et al. [7] used pre-trained VLMs like CLIP to generate pseudo-region annotations, which were then used as training data for object detectors. While this improved text-image alignment, it also highlighted the necessity for enhanced localization [18]. Building on this trend, DetCLIP [37] integrated region proposal networks with CLIP, providing one of the first models capable of refining object detection without extensive retraining. DetCLIPv2 [36] takes this further by scaling up OVDet pre-training through improved word-region alignment, highlighting the growing importance of aligning textual descriptions with image regions for effective detection. In parallel, DePro [5] introduced automatic prompt learning to effectively integrate CLIP class embeddings into fine-grained detection, and Detic [46] enhanced performance on novel classes by supervising the largest region proposals with image classification labels. However, these approaches rely heavily on expensive annotations or complex two-stage architectures, which limit their scalability.

A notable breakthrough in this area is BARON [33], which shifted the focus from individual region alignment to aggregating multiple regions, enhancing detection performance by considering broader image context. CoDet [20] uses co-occurrence guided region-word alignment, leveraging the relationships between object regions within an image to achieve more precise and contextually aware detections. Similarly, [22] introduced region-level attention mechanisms, enabling a better understanding of spatial relationships by considering neighboring regions for improved alignment. [42] provided an in-depth analysis of how existing detection models handle region-word alignment tasks, revealing the limitations in current approaches. Moreover, the importance of fine-grained region-word correspondence is also emphasized in [2], which underscores that OVDet models need to improve their ability to differentiate between visually similar objects and their textual descriptions, a gap that previous models struggled to address. Unlike these models, we propose a model that directly align image regions and words (phrases) through an attentive masking strategy.

3 Proposed Model

Our model integrates data from various sources, including grounding, detection, and visual-textual pairs, to create a robust open-world object detection framework. We begin by presenting a unified data representation for training with diverse supervision types (Section 3.1). For visual-textual pairs without instance annotations, we use an attentive masking strategy to identify relevant image regions for each textual concept (Section 3.2). Finally, we detail our training objectives in Section 3.3.

3.1 Data Representation

In our proposed model, each data sample is represented as a triplet $(x_I, \{b_i\}_{i=1}^N, \{c_j\}_{j=1}^S)$, in which, $x_I \in \mathbb{R}^{3\times h\times w}$ represents the image, $\{b_i | b_i \in \mathbb{R}^4\}_{i=1}^N$ denotes the group of N bounding boxes, and $T = \{c_j\}$ is the set of concept names or textual embeddings. This unified representation allows the model to flexibly handle various types of data and effectively address the unique challenges posed by each type. For detection, our model distinguishes between semantically similar categories. Instead of solely relying on categorical labels (e.g., dog), we incorporate negative samples in the form of semantically related but incorrect categories (e.g., wolf, fox). This helps the model learn the visual distinctions between similar objects, improving its performance on novel or ambiguous categories. In grounding tasks, the model associates specific words or phrases in a caption with regions in the image. We use hierarchical concept embeddings, where positive samples $\{c_j\}_{pos}$ contain objects explicitly mentioned in the caption, and negative samples are selected from unrelated concepts (those not present in the caption). For image-text pairs without instance-level annotations, $\{b_i\}_{i=1}^N = \emptyset$, the model only has access to the caption and noun phrases extracted from it. This setup allows the model to rely on the text description to infer relevant regions in the image.

Similar to previous approaches [18, 20, 36, 37], our model architecture consists of an image encoder and a text encoder. The image encoder processes the input image x_I to generate region proposals $R = \{r_i\}_{i=1}^K$ along with their region features $f_R^i \in \mathbb{R}^{K\times D}$, where K is the number of region proposals and D is the feature dimension. The text encoder processes the text concepts $\{c_1, c_2, \ldots, c_S\}$, producing a set of embeddings $f_T \in \mathbb{R}^{S\times D}$. Once both the image region features and text embeddings are obtained, we calculate a similarity matrix $W \in R^{K\times S} = f_R \cdot [f_T]^T$, which captures the alignment between each image region and each text concept. For tasks with available annotations, we define an alignment matrix $Y \in \{0, 1\}^{K\times S}$, that indicates the correct alignment between the image regions and text concepts. However, for image-text pairs without instance-level annotation, we introduce the following strategy to address the challenge of identifying relevant regions.

3.2 Image-Text Pairs Representation

When working with large-scale image-text datasets, such as web-crawled data where instance-level annotations are missing, it becomes challenging for the model to identify which regions in the image are relevant to the text. Simple contrastive learning methods which rely on global image-text matching (comparing the entire image to the entire text), often fail to capture the fine-grained relationships [7, 18, 36, 42]. This becomes especially problematic when multiple objects are present, and only a subset is relevant to the text description. To address this issue, we propose an Attentive Masking (AM) strategy that enables the model to focus on the most relevant image regions for each word or concept in the text. Given an image-text pair (x_I, x_T), where x_I is the image and x_T is the associated text, we extract a set of noun phrases $T = \{c_j\}_{j=1}^S$ from the text. The model processes this pair using the following steps. The image encoder generates region proposals $R = \{r_i\}_{i=1}^K$ and their features $f_R \in \mathbb{R}^{K\times D}$, where K is the number of regions in the image. The text encoder extracts text embeddings $f_T \in \mathcal{R}^{S\times D}$ for the concepts $\{c_j\}_{j=1}^S$, where S is the number of text concepts (tokens). To align these image regions with the text concepts, we compute the similarity between each image region and all text embeddings. Specially, for the i-th region, the highest similarity across all text embeddings $(c_1, c_2, \ldots, c_S)$ is selected

$$W_i = \max_{j=1}^{S} w_{ij}, \quad \text{where} \quad w_{ij} = \langle f_T^j, f_R^i \rangle \tag{1}$$

to represent the correlation/alignment with the text. WThese similarity scores $\{W_1, W_2, \ldots, W_K\}$ are used to create a mask ($M_k \in \{0, 1\}$) that indicates whether the i-th image region is relevant ($M_k = 1$); or not ($M_k = 0$). After masking irrelevant regions, we align the remaining regions with

the text concepts. In many cases, a single text concept may correspond to multiple image regions. To handle this, we modify the alignment strategy to allow soft associations between text concepts and image regions. Instead of selecting only one region per concept, we compute a softmax-weighted sum over the similarity scores of all regions for each text concept

$$W(x_I, x_T) = \frac{1}{S} \sum_{j=1}^{S} \sum_{i=1}^{k} A_{j,i} \cdot w_{j,i}, \tag{2}$$

where, $A_{j,i} = \frac{\exp(w_{j,i}/\tau)}{\sum_{t=1}^{K} \exp(w_{j,t}/\tau)}$ is the softmax weight over region similarities for the j-th concept. This method allows each concept to align with multiple regions, improving the handling of neighboring region effects.

Contrastive Learning. Based on the obtained region-word alignment, we employ contrastive learning to refine the model's ability to associate image regions with corresponding text concepts. During training, we consider a batch of N image-text pairs $\{x_I^i, x_T^i\}_{i=1}^N$ and their corresponding representations $\{f_R^i, f_T^i\}_{i=1}^N$. The contrastive loss encourages correct word-region alignments by maximizing the similarity between correctly matched pairs while minimizing the similarity between incorrect or mismatched pairs. This loss for a batch of image-text pair is defined as

$$\mathcal{L}_{\text{cont}} = -\frac{1}{N} \log \frac{\exp(W(x_I^i, x_T^i)/\tau)}{\sum_{j=1}^{N} \exp(W(x_I^j, x_T^j)/\tau)} \tag{3}$$

This focuses on text-to-image alignment, ensuring that the i-th text x_T^i is more similar to its corresponding image x_I^i than to other images. τ is a temperature hyperparameter that controls the sharpness of the softmax distribution. Incorporating this alignment strategy allows the model to capture fine-grained relationships between the image regions and text concepts, making it effective in cases where multiple objects are present, but only a subset of them is relevant to the text.

Proposal Selection. The goal of proposal selection is to identify the most informative regions (proposals) in an image to compute similarities with the given textual concepts. Several methods (like RPN [24] and FCOS [31]) can achieve this by using objectness scores found in object detectors. However, they lack direct consideration of textual information, which is crucial for open-vocabulary object detection [33]. To ensure that the selected regions are valuable for contrastive learning, we use a scoring function that considers both visual and textual information: $o_k = \sigma(W_v f_p[k] + W_t f_T[j] + b)$, where, $f_p[k]$ is the feature of the k-th region proposal, and $f_T[j]$ is the embedding of the j-th text concept. $\sigma(x) = 1/(1 + e^{-x})$ ensuring the output score o_k falls between 0 and 1. By learning the combination of visual and textual information, the model can more accurately identify regions that are both visually distinctive and contextually relevant to the text. After computing o_k for all proposals, we can rank the regions and select the top-K proposals based on their scores.

3.3 Training Objective

Our model is built on the ATSS detector, enhanced with a transformer-based text encoder that provides textual embeddings to enrich the detection process, following the approach in [36, 37]. Additionally, the training objectives for our model are tailored to different tasks: detection, grounding, and image-text pair alignment. For detection, the total loss is a combination of the matching/alignment loss (L_{alg}), regression loss (αL_{reg}), and centerness loss (βL_{cent}). For grounding data, we use only the alignment loss (L_{alg}). This is because grounding annotations often have inaccurate bounding boxes, making regression and centerness losses less reliable. When dealing with text-image pairs, we incorporate a contrastive loss (λL_{const}) to align visual regions with textual descriptions. For these tasks, we use specific loss functions like focal loss for L_{alg}, cross-entropy (CE) for L_{center}, and for the regression loss L_{reg} we used GIoU [25]. In our training process, high-resolution images with small batch sizes are used for detection and grounding tasks, while for visual-textual pairs, we use low-resolution images with large batch sizes to enhance efficiency and increase the number of negative samples.

4 Experiments

To explore a large-scale generalized setting for open-vocabulary object detection, we conduct experiments across multiple datasets. Specially, for detection, we use Objects365v2 [28] dataset, containing

	Model	AP	APr	APc	APf
1	class	27.9	26.8	28.2	28.5
2	IoU	29.8	29.1	30.2	30.3
3	cent.	30.5	28.5	30.8	30.4
4	o_k, b=0.4	28.7	27.5	29.6	30.8
5	o_k, b=0	**31.6**	**29.6**	**31.9**	**31.7**
6	o_k, b=0.1	31.2	30.4	31.5	31.1

(a) Comparison of different proposal selection strategies. Our model with $b = 0$ achieves the optimal results.

Model	AP	APr	APc	APf
bbox	30.1	28.9	37.3	30.8
one-to-one	**31.6**	**29.6**	**31.9**	**31.7**
one-to-many	31.2	31.5	30.9	31.5

(b) Region-text matching strategies. One-to-one, pairing each text concept with the closest region, which shows the best fixed AP. The one-to-many strategy shows significant improvement in rare categories.

Nb.	AP	APr	APc	APf
25	30.2	29.6	30.4	30.6
50	30.9	30.3	30.5	31.2
100	**31.6**	**29.6**	**31.9**	**31.7**
200	31.1	29.5	30.7	31.4

(c) Number of proposals, where $k = 100$ achieves the best performance.

λ	AP	APr	APc	APf
0.05	30.7	29.6	29.8	31.2
0.2	**31.6**	**29.6**	**31.9**	**31.7**
0.5	30.8	29.3	31.2	30.4
1	29.1	27.4	28.7	29.5

(d) Effect of contrastive loss weight. ($\lambda = 0.2$) gives the best results.

η	AP	APr	APc	APf
1	30.5	28.7	30.1	30.0
0.5	**31.6**	**29.6**	**31.9**	**31.7**
0.2	31.2	29.5	31.3	30.8
0.05	28.7	27.3	28.9	29.5

(e) Effect of temperature parameter. $\eta = 0.5$ shows the best results.

Table 1: Ablation experiments using the Swin-T backbone trained on Objects365+CC3M dataset. Fixed AP (%) is reported on LVIS minival5k for rare (r), common (c), and frequent (f) categories.

approximately 0.66M images. For grounding, we use the GoldG [14] dataset, which has been refined by excluding images from the COCO to avoid overlap with the LVIS [9] dataset, ensuring a fair zero-shot evaluation. In terms of image-text pairs, we use a combined dataset from Conceptual Captions: CC12M [3], which includes 12M pairs, and CC3M [29], consisting of 3M pairs.

Implementation details. We use the Swin-Transformer [19] as the visual encoder and a pretrained FILIP model [38] for the text encoder, with a maximum token length of 16 to ensure efficient training and inference. Our training process varies depending on the type of data. For example, for detection and grounding, we use high-resolution inputs 1280×800 and a smaller batch size, setting to 128 for Swin-T and 256 for Swin-L models. In contrast, for image-text pairs, we use lower-resolution inputs (320×320) but increase the batch size significantly to 6144 (192 images per GPU for Swin-T and 96 for Swin-L), which also helps reduce the computational cost associated with processing large-scale image-text pairs. The hyperparameters in the training objective are set as follows: $\alpha = 2$, $\beta = 0.8$, and $\lambda = 0.2$ in the loss equation. By default, all models are trained for 12 epochs to ensure adequate learning without overfitting. The masking ratios set to 50%. In line with previous work [5, 20, 37, 41], we evaluate the zero-shot performance of our model on the LVIS [9] dataset, which consists of 1203 categories. We use the Fixed AP [4] on the LVIS minival5k throughout our experiments. We also conduct evaluations on the ODinW13 dataset [17] with 13 diverse downstream detection tasks. This evaluation on ODinW allows us to test our model's ability to generalize to new domains and distributions. Unlike the ViLD protocol [8], which splits LVIS into seen and unseen categories and partially relies on LVIS data for training, we adopt the GLIP [41] procedure, which does not assume any prior knowledge about downstream tasks, providing a realistic open-world evaluation setting.

Ablation Study. We conduct ablation experiments on the LVIS minival5k dataset to evaluate the effects of different strategies and settings on zero-shot object detection performance. Table 1a determines the best strategy for selecting region proposals for contrastive learning. We compare the following: classification and IoU scores [13, 35] are generated by adding an additional head after the regression branch, while, the centerness score [31] that is used in the ATSS detector [43] favors the proposals near the object center. Among these scores, centerness and IoU outperform classification, achieving 30.5 and 29.8 AP, respectively, showing a 7-8% improvement over classification (27.9). We also consider three different settings of our model (in section 3.2): i) our model with $b = 0$ (row5) serves as a neutral baseline, focusing solely on the of visual and textual features, yielding the highest AP of 31.6. ii) Our model (with $b = 0.1$) (row6), defining a slight positive bias results in a competitive AP of 31.2, which is still higher than traditional methods but indicates a 2% decrease compared to the optimal setting of $b = 0$. iii) Our model with $b = 0.4$ (row4), shows a slight drop

Data	AP / APr / APc / APf
Obj [28]	29.3 / 25.1 / 27.5 / 30.4
Obj + CC3M [29]	31.6 / 29.6 / 31.9 / 31.7
Obj + GoldG [14] + CC3M	**37.9 / 36.5 / 38.1 / 39.4**

(a) Impact of additional data on detection performance, using Objects365 (Obj), GoldG and CC3M.

Model	LVIS minival				ODinW-13
	AP	APr	APc	APf	mAP
GLIP [17]	SW ×	×	×	×	65.2
GLIPv2 [41]	SW 50.1	×	×	×	66.5
MQ-GLIP [34]	SW 50.4	43.8	52.6	50.3	68.3
AlignDet (ours)	SW 50.6	44.1	53.2	50.8	69.1

(b) Fine-tuning results using Swin-T backbone (SW).

Table 2: (a) Improvement in detection performance using additional data. (b) Transfer learning evaluation, reporting Fixed AP (%) on LVIS and mAP on ODinW-13 datasets.

Method	Backbone	AP	APr	APc	APf	Avg AP
ViLD [8]	ResNet50	25.5	16.6	24.6	30.3	24.3
DetPro [5]	ResNet50	25.9	19.8	25.6	28.9	25.1
Detic [46]	ResNet50	30.9	19.5	–	–	–
BARON [33]	ResNet50	29.5	23.2	29.3	32.5	28.6
CoDet [20]	ResNet50	30.7	23.4	30.0	34.6	29.7
VLDet [18]	ResNet50	30.1	21.7	29.8	34.3	29.0
AlignDet (Ours)	ResNet50	30.2	23.7	30.1	34.3	29.5
Detic [46]	Swin-B	38.4	23.9	40.2	42.8	36.3
DetCLIP [37]	Swin-T	28.4	25.0	27.0	31.6	28.0
GLIP [17]	Swin-T	26.5	21.3	21.7	31.2	25.2
GLIPv2 [41]	Swin-T	29.7	–	–	–	–
DetCLIPv2 [36]	Swin-T	32.8	31.0	31.7	34.8	32.6
VLDet [18]	Swin-T	34.1	27.5	32.7	35.1	32.4
CoDet [20]	Swin-B	39.2	29.4	39.5	43.0	37.8
MQ-GLIP-T [34]	Swin-T	30.4	21.0	27.5	34.6	28.4
AlignDet (Ours)	Swin-T	34.1	30.5	32.1	36.2	33.2

Table 3: Open-vocabulary object detection results on LVIS dataset for different backbones: ResNet50 [10] and Swin [19]. The average AP (Avg AP) is calculated to provide an overall performance summary. Our model achieves comparative performance along other state-of-the0arts methods.

in performance compared to the other variations. Table 1b evaluates different region-text matching strategies. The first row matches each text concept to the region with the maximum bounding box overlap, which shows decent performance in frequent categories (28.9). The one-to-one strategy pairs each text concept with the region having the highest similarity, resulting in the best overall AP of 31.6. One-to-many, allows each text concept to match with multiple regions, aggregating similarities across all these regions, yielding an AP of 31.2, slightly lower than the one-to-one strategy. Table 1c explores the impact of the number of proposals (k). Using 100 proposals yields the highest AP of 31.6. Increasing k to 200 introduces more low-quality candidates, slightly decreasing performance. Reducing k to 25 limits the extracted regions, resulting in a significant drop in performance. Tables 1d and 1e evaluates the effect of contrastive loss λ and different temperature η values. The standard settings ($\lambda = 1$, $\eta = 0.07$) from classic contrastive learning [38] are not optimal for our model. Our experiments found that $\lambda = 0.2$ and $\eta = 0.5$, enhance the performance.

Exploiting Extra Data. Table 2a shows the impact of using different pretraining datasets on the model's performance. Training on Objects365 alone results in an overall AP of 29.3. However, by adding the conceptual captions (CC3M), the AP increases to 31.6, which is an 8% improvement. This enhancement is especially notable in rare categories, where AP rises from 25.1 to 29.6, representing a significant gain of 19%. Furthermore, incorporating the GoldG along with Objects365 and CC3M, dramatically boosts the overall AP to 37.9, marking a 31% improvement over Objects365 alone. The results show that the inclusion of extra data significantly enhances the model's effectiveness.

Cross-dataset Transfer. We evaluate the transferability of our model by fine-tuning it on downstream tasks: the LVIS dataset with 1203 categories, and ODinW-13, which contains 13 detection tasks. As shown in Table 2b, AlignDet achieves an AP of 50.6 on LVIS, outperforming MQ-GLIP-T and showing strong performance across rare, common, and frequent categories. On ODinW-13, AlignDet

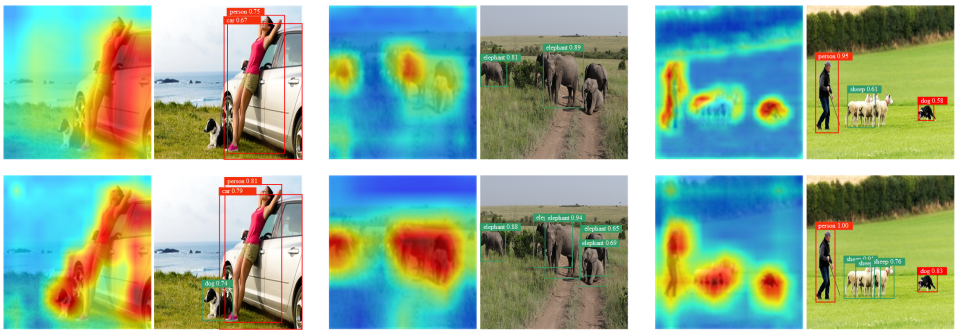

Figure 3: Comparison of detection results between AlignDet (bottom row) and GLIPv2 (top row) on COCO's validation set. The heatmaps (left) illustrate attention over both seen and novel categories, while the bounding boxes (right) highlight the detected objects. Green boxes indicate novel categories, and red boxes represent base categories. Our method demonstrates strong open-vocabulary capability and correctly detects challenging samples such as the occluded elephant and a small dog among the flock of sheep, even when these objects are not seen in training. In contrast, the heatmaps reveal that GLIPv2 struggles to effectively detect novel objects, where the model's focus is less precise.

attains an mAP of 69.1, surpassing GLIPv2 by +2.6. The superior performance on ODinW-13 demonstrates the model's ability to generalize across diverse detection tasks.

Result of LVIS OVDet Benchmark. We compare our model with state-of-the-art open-vocabulary detectors in Table 3, using both ResNet50 and Swin backbones. Our model uses a simple yet effective attentive masking strategy, yielding competitive results in open-vocabulary object detection. With a ResNet50 backbone, we achieve an average AP of 29.5, which is competitive with models like VLDet (29.0) and CoDet (29.7). When using a more advanced Swin-T backbone, AlignDet achieves an Avg AP of 33.2, which is higher than VLDet (32.4) and MQ-GLIP-T (28.4). While CoDet, with its heavier Swin-B architecture, achieves a higher Avg AP of 37.8, AlignDet remains highly competitive, especially in rare categories (APr), where it reaches 30.5, surpassing other models and tackling the critical challenge of recognizing rare objects in open-vocabulary detection.

Visualization . Fig. 1 demonstrates the effectiveness of AlignDet in aligning words with image regions across diverse contexts on the CC12M dataset [3]. For each textual concept, the model selects the best-matching region based on the highest similarity score (detailed in Section 3.2). For example, in example (b), our model accurately detects and aligns multiple entities– person, bicycle, and cow–highlighting interactions between humans and animals in an outdoor setting. In example (d), AlignDet correctly identifies specific food items on the plate, such as pasta, broccoli, and carrots, which shows the model's capacity for recognizing fine-grained categories that are not annotated in the detection datasets. These capabilities are critical for open-world detectors but are not fully captured by standard benchmarks such as LVIS [9]. We further visualize the detection results of AlignDet (bottom row) and GLIPv2 (top row), in Fig. 3. The images are taken from COCO's validation set. AlignDet generates focused and precise responses at locations corresponding to both seen and unseen object categories, while GLIPv2 tends to produce weaker or diffused responses. Notably, AlignDet excels at detecting multiple objects associated with a single textual concept. For example, the model accurately detects a person and a flock of sheep when prompted with the concept dog, even though the dog is much smaller and positioned far from the sheep. This illustrates the model's ability to capture fine-grained object relationships, a key requirement for open-vocabulary detection tasks.

5 Conclusion

We proposed AlignDet, a framework for open-vocabulary object detection that leverages extensive image-text data through a unified training approach. By employing a region-word alignment strategy and attentive masking, AlignDet effectively captures fine-grained relationships between image regions and textual concepts, significantly improving detection accuracy, particularly for novel or unseen categories. Our experiments demonstrate its superior performance and scalability, highlighting a promising direction for open-world detection by leveraging diverse, large-scale visual-textual pairs.

References

[1] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 384–400.

[2] L. Bianchi, F. Carrara, N. Messina, C. Gennaro, and F. Falchi, "The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 520–22 529.

[3] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.

[4] A. Dave, P. Dollár, D. Ramanan, A. Kirillov, and R. Girshick, "Evaluating large-vocabulary object detectors: The devil is in the details," *arXiv preprint arXiv:2102.01066*, 2021.

[5] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "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*, 2022, pp. 14 084–14 093.

[6] D. Fontanel, M. Tarantino, F. Cermelli, and B. Caputo, "Detecting the unknown in object detection," *arXiv preprint arXiv:2208.11641*, 2022.

[7] M. Gao, C. Xing, J. C. Niebles, J. Li, R. Xu, W. Liu, and C. Xiong, "Open vocabulary object detection with pseudo bounding-box labels," in *European Conference on Computer Vision*, 2022, pp. 266–282.

[8] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2022.

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

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] M. Inkawhich, N. Inkawhich, H. Li, and Y. Chen, "Tunable hybrid proposal networks for the open world," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[12] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*, 2021, pp. 4904–4916.

[13] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *European conference on computer vision*, 2018, pp. 784–799.

[14] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[15] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[16] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[17] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 965–10 975.

[18] C. Lin, P. Sun, Y. Jiang, P. Luo, L. Qu, G. Haffari, Z. Yuan, and J. Cai, "Learning object-language alignments for open-vocabulary object detection," *International Conference on Learning Representations*, 2023.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[20] C. Ma, Y. Jiang, X. Wen, Z. Yuan, and X. Qi, "Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection," *Advances in neural information processing systems*, vol. 36, 2024.

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

[22] S. Qiang, X. Li, Y. Liang, W. Liao, T. He, and P. Peng, "Open-vocabulary object detection via neighboring region attention alignment," *arXiv preprint arXiv:2405.08593*, 2024.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[25] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[26] P. Shamsolmoali, J. Chanussot, H. Zhou, and Y. Lu, "Efficient object detection in optical remote sensing imagery via attention-based feature distillation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023.

[27] P. Shamsolmoali, M. Zareapoor, E. Granger, and M. Felsberg, "Setformer is what you need for vision and language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4713–4721.

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

[29] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "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)*, 2018, pp. 2556–2565.

[30] Y. Shen, C. Fu, P. Chen, M. Zhang, K. Li, X. Sun, Y. Wu, S. Lin, and R. Ji, "Aligning and prompting everything all at once for universal visual perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[31] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection. arxiv 2019," *arXiv preprint arXiv:1904.01355*, 2019.

[32] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang *et al.*, "Towards open vocabulary learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[33] S. Wu, W. Zhang, S. Jin, W. Liu, and C. C. Loy, "Aligning bag of regions for open-vocabulary object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 254–15 264.

[34] Y. Xu, M. Zhang, C. Fu, P. Chen, X. Yang, K. Li, and C. Xu, "Multi-modal queried object detection in the wild," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[35] J. Yang, J. Lu, D. Batra, and D. Parikh, "A faster pytorch implementation of faster r-cnn," 2017.

[36] L. Yao, J. Han, X. Liang, D. Xu, W. Zhang, Z. Li, and H. Xu, "Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 497–23 506.

[37] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9125–9138, 2022.

[38] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," *International Conference on Learning Representations*, 2022.

[39] M. Zareapoor, P. Shamsolmoali, H. Zhou, Y. Lu, and S. García, "Fractional correspondence framework in detection transformer," in *ACM Multimedia*, 2024.

[40] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 393–14 402.

[41] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "Glipv2: Unifying localization and vision-language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 067–36 080, 2022.

[42] H. Zhang, Q. Zhao, L. Zheng, H. Zeng, Z. Ge, T. Li, and S. Xu, "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*, 2024, pp. 16 975–16 984.

[43] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[44] S. Zhao, L. Zhao, Y. Suh, D. N. Metaxas, M. Chandraker, S. Schulter *et al.*, "Generating enhanced negatives for training language-based object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 592–13 602.

[45] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[46] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*, 2022, pp. 350–368.