Exploring Open-Vocabulary Models for Category-Free Detection

Pablo Garcia-Fernandez[1 [0009−0009−9162−4073]], Daniel Cores[1 [0000−0002−5548−4837]], and Manuel Mucientes[1 [0000−0003−1735−3585]]

University of Santiago de Compostela, Spain
{pablogarcia.fernandez,daniel.cores,manuel.mucientes}@usc.es

Abstract. Object detection models typically rely on a predefined set of categories, limiting their applicability in real-world scenarios where object classes may be unknown. In this paper, we propose a novel, training-free framework that enables off-the-shelf open-vocabulary object detectors (OvOD) to perform category-free detection —localizing and classifying objects without any prior category knowledge. Our approach leverages image captioning to dynamically generate descriptive terms directly from the image content, followed by a WordNet-based filtering process to extract semantically meaningful category names. These discovered categories are then embedded and matched with visual region features using a frozen OvOD model to perform detection. We evaluate our method on the COCO dataset in a fully zero-shot setting and demonstrate that it significantly outperforms strong multimodal large language model baselines, achieving an improvement of over 30 AP points. This highlights our method as a promising direction for more adaptive solutions to real-world detection challenges.

Keywords: category-free · open-vocabulary object detection · captioning.

1 Introduction

Object detection has become a cornerstone of modern computer vision, with advances in deep learning enabling highly accurate detectors across a wide range of applications. However, the success of most object detectors hinges on a fundamental assumption: the complete set of object categories is known and fixed during training. This assumption is deeply embedded in the construction of popular benchmarks such as COCO and LVIS, where models are trained to detect a limited number of predefined categories and evaluated accordingly. While effective in controlled environments, this closed-world assumption breaks down in real-world settings, where objects of interest may not be part of the training vocabulary.

In many practical applications —such as autonomous driving, surveillance, or general-purpose scene understanding— the relevant object categories may be unknown, ambiguous, or context-dependent. Relying on a fixed set of known

categories in those cases severely limits the flexibility and generalization ability of traditional detectors. Ideally, we would like to build object detectors that do not rely on any prior knowledge of the category space, and can instead adaptively discover and detect objects based on the content of the image itself.

Open-vocabulary object detection (OvOD) represents a significant step toward this goal. These methods leverage vision-language alignment techniques to enable object detection for arbitrary textual categories at inference time. Typically, a detector is trained to align region-level features with textual embeddings (e.g., from CLIP [25]), allowing it to respond to category names outside of the training set. However, despite their impressive flexibility, OvOD methods still implicitly assume access to a relevant set of object names at test time —whether provided manually, sampled from a predefined vocabulary, or selected from a prompt. In other words, while the vocabulary is technically open, its selection is still guided by prior knowledge. This raises an important question: **can we detect objects in an image without assuming any prior knowledge of the possible categories?**

In this work, we propose a training-free framework that enables *off-the-shelf* open-vocabulary object detectors to operate without relying on predefined category priors. Our central idea is to discover a relevant vocabulary directly from the image using an image captioning model. Captioning provides a natural way to surface human-interpretable terms associated with objects in an image. We extract candidate object names from generated captions and use them to guide an open-vocabulary detector. This allows our system to perform object detection in a zero-prior setting, where no fixed label set, prompt, or external vocabulary is assumed.

One consequence of working without a fixed category set is that traditional baselines are no longer applicable. Most supervised detectors are not designed to operate without a predefined vocabulary. As such, we compare our method against multimodal large language models (MLLMs) trained for grounding and spatial reasoning, such as KOSMOS-2 [24], as well as general-purpose MLLMs like GPT-4o [1] and LLaVA-Video [17], which are not explicitly designed for object localization. These approaches represent reasonable references for detection guided purely by image understanding and free-form language. We show that our method outperforms state-of-the-art methods by over ↑30 AP points in this challenging zero-prior setting. To summarize, our key contributions are:

- We demonstrate that, although multimodal large language models (MLLMs) are currently the only models capable of operating in a zero-category prior setting, they exhibit significant limitations in reliably detecting objects.
- We propose a training-free approach that adapts *off-the-shelf* OvOD detectors to operate without any predefined or prompted category priors, by automatically deriving a relevant vocabulary from image captions.
- We show that our approach achieves state-of-the-art performance, surpassing the strongest baseline by over ↑30 AP points.

2 Related work

Open vocabulary object detection (OvOD) [38,26] has made significant progress driven by the emergence of vision-language models (VLMs), which enable detectors to move beyond fixed category sets and handle novel concepts at inference time. Unlike traditional zero-shot object detection, which relies solely on textual semantics to recognize unseen classes, OvOD methods incorporate various forms of weak supervision to improve both classification and localization performance. These methods can generally be grouped into four main strategies: (i) region-aware training, (ii) pseudo-labeling, (iii) knowledge distillation, and (iv) transfer learning.

Region-aware approaches aim to improve the alignment between image regions and textual descriptions during training, enhancing the detector's ability to generalize to unseen categories. Works like DetCLIP [33] , DetCLIPv2 [32], CORA [29], and VLDet [15] refine region-level feature-text correspondence to improve localization and semantic alignment. **Pseudo-labeling** methods expand the training data by using large-scale VLMs to generate object labels automatically. RegionCLIP [36], PromptDet [6], CoDET [21], GLIP [14], Detic [37], and Grounding DINO [18] follow this paradigm to build richer supervision from unlabeled images. **Knowledge distillation** techniques transfer the representational power of VLMs into detection models by using them as teachers. Approaches like BARON [27], DK-DETR [13], CLIPSelf [28], and SIC-CADS [5] distill knowledge from the VLM into the detector to improve its generalization ability. Finally, **transfer learning-based** methods directly incorporate pretrained vision-language encoders into the detection pipeline, either through fine-tuning (OWL-ViT [23]) or by freezing the encoder and learning lightweight heads (F-VLM [12]).

While classical OvOD approaches have significantly expanded the flexibility of object detectors, they still rely on externally supplied vocabularies —in the form of text prompts, or dataset-specific category lists. In contrast, our method removes this dependency entirely by discovering object categories dynamically through image captioning. This enables the detector to operate without any prior assumptions about which object classes might appear at inference time, representing a more flexible and realistic scenario.

Open-set and open world detection. The problem of detecting objects without prior knowledge of all possible categories has been studied under several paradigms. **Open-Set Object Detection** (OSOD) addresses scenarios in which a detector must correctly classify instances belonging to known categories while also identifying and localizing objects from unknown categories —without assigning them specific semantic labels. These unknown instances are typically grouped under a generic unknown class. Thus, the main objective of OSOD is to enable robustness against out-of-distribution (OOD) categories, focusing on their rejection rather than discovery. Dhamija et al.[4] were the first to formally define the OSOD setting, showing that the performance of conventional object detectors is often significantly overestimated under open-set conditions. Subsequent works

have explored methods to improve OOD robustness through various mechanisms such as background expansion[9], adaptive classification thresholds [19], or uncertainty modeling using Bayesian dropout [22], yet they still treat unknown objects as undifferentiated outliers and do not aim to recover their category semantics.

Open World Object Detection (OWOD) extends the open-set setting by introducing a continual learning framework in which novel categories are progressively encountered and incorporated over time. The open-world paradigm was first introduced in image classification by Bendale et al.[2], who proposed a model capable of rejecting unknown classes at test time and incrementally integrating them once labeled. Joseph et al.[11] later adapted this paradigm to object detection, formalizing the OWOD task and proposing a method based on exemplar replay to enable the model to learn new object categories while mitigating catastrophic forgetting. However, despite these contributions, OWOD models [39,31,30,8] remain limited in their ability to autonomously explore or infer the semantics of unknown categories; they still require explicit human annotation to incorporate new classes. This reliance on supervision poses a major bottleneck for scalability in realistic open-world scenarios, where novel objects frequently appear and manual labeling is impractical.

To alleviate this issue, Zheng et al. [35] propose a method to automatically discover categories of unknown objects based on their visual appearance. Positioned between the open-set and open-world paradigms, their approach clusters unknown instances into a fixed number of generic categories—each potentially corresponding to a novel class, without requiring labeled data. While this method enables unsupervised discovery, a fundamental limitation remains: it does not capture the semantics of these categories. The discovered groups lack meaningful and interpretable labels. Our approach addresses this gap by introducing a language-driven mechanism for both the discovery and semantic grounding of unknown categories. Rather than relying on manual supervision (as in OWOD) or unsupervised clustering without semantic interpretation (as in [35]), we leverage image captioning models to extract rich, contextual object-level descriptions directly from images. As a result, our system can not only localize previously unseen objects but also assign them human-interpretable labels.

3 Method

We propose a *training-free* method that enables *off-the-shelf* open-vocabulary object detectors to operate without predefined category priors, by discovering a relevant vocabulary directly from image captions. As illustrated in Fig. 1, our approach comprises two main components: Vocabulary Discovery and Category-Free Object Detection. In the **Vocabulary Discovery** stage, a captioning model generates textual descriptions for the input images. Using an external corpus, we filter these terms to retain only those that are valid object category candidates. This step allows the model to construct its vocabulary directly from the data, without relying on external supervision. In the **Category-Free Object De-**

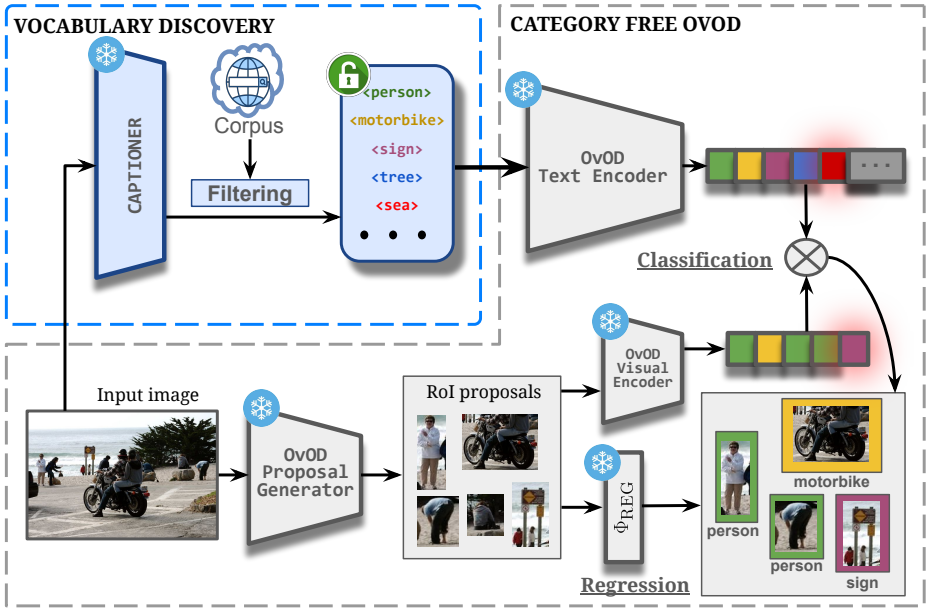

Fig. 1. Our framework consists of two main stages. In **Vocabulary Discovery** (3.1), we prompt a captioning model to produce object-centric descriptions, from which we extract candidate categories using WordNet-based noun filtering. In **Category-Free OVOD** (3.2), we use a frozen open-vocabulary detector to match visual region features with the discovered categories via cosine similarity in a shared embedding space.

tection stage, the discovered vocabulary is embedded using a text encoder, and the resulting text embeddings are matched with region-level visual embeddings to detect objects corresponding to the discovered categories.

3.1 Vocabulary Discovery

The vocabulary discovery stage aims to extract a set of candidate object categories $\mathcal{V}$ from a collection of unlabeled images $\{\mathbf{I_i}\}_{i=1}^{N}$, without relying on predefined labels or external supervision. This is achieved by generating object-centric textual descriptions using a vision-language model, followed by a filtering step to retain only valid nouns.

Caption Generation. For each image $\mathbf{I_i}$, we prompt a vision-language model to produce a structured caption consisting of visible object names, constrained to a semicolon-separated format. Specifically, the model is guided via prompting to output:

$$T_i = \mathcal{C}(\mathbf{I_i}), \tag{1}$$

where T_i is a string of candidate object terms (e.g., `"car; tree; road; sign"`). These strings are parsed into token lists $W_i = \texttt{Split}(T_i, \text{";"})$, and aggregated across the dataset to form a candidate term set:

$$\mathcal{W} = \bigcup_{i=1}^{N} W_i \tag{2}$$

WordNet-Based Noun Filtering. To ensure semantic validity, we filter the candidate terms using WordNet. A word $w \in \mathcal{W}$ is retained if it appears in the WordNet lexical corpus as a noun:

$$\texttt{IsNoun}(w) = (\exists s \in \texttt{WordNet}(w) \text{ s.t. } \texttt{POS}(s) = \texttt{'n'}), \tag{3}$$

where $\texttt{POS}(w)$ denotes the part of speech of $\texttt{w}$, and $\texttt{'n'}$ indicates a noun. The final vocabulary is defined as:

$$\mathcal{V} = \{w \in \mathcal{W} \mid \texttt{IsNoun}(w)\}. \tag{4}$$

This simple yet effective strategy allows us to construct a domain-relevant, semantically grounded vocabulary of object categories directly from the dataset, **without external priors**.

3.2 Category-Free Object Detection

Once the vocabulary $\mathcal{V} = \{v_1, v_2, \ldots, v_K\}$ has been discovered, we leverage an *off-the-shelf* open-vocabulary object detector to localize and classify instances of these categories. Our approach does not require **any retraining or fine-tuning**, relying entirely on the alignment between visual and textual embeddings in a shared feature space.

Text Embedding. Each category name $v_k \in \mathcal{V}$ is embedded into a shared semantic space using the OvOD pretrained text encoder ϕ_{text}:

$$\mathbf{z_k}^{\text{text}} = \phi_{\text{text}}(v_k), \quad \forall k \in \{1, \ldots, K\}. \tag{5}$$

This yields a matrix of category embeddings $\mathbf{Z}_{\text{text}} \in \mathbb{R}^{K \times d}$, where d is the dimensionality of the joint embedding space.

Visual Embedding and Region Proposal. Given an input image $\mathbf{I}$, the detector extracts a set of region proposals $\{z_j\}_{j=1}^{M}$, each associated with a visual embedding computed via a visual encoder ϕ_{vis}:

$$\mathbf{z_j}^{\text{vis}} = \phi_{\text{vis}}(z_j), \quad \forall j \in \{1, \ldots, M\}. \tag{6}$$

These embeddings $\mathbf{Z}_{\text{vis}} \in \mathbb{R}^{M \times d}$ are aligned with the text embeddings.

Similarity-Based Classification. To assign a category label to each region proposal, we compute the similarity between visual and textual embeddings using cosine similarity, $\langle \cdot, \cdot \rangle$:

$$\hat{y}_j = \arg\max_k \langle \mathbf{z_j}^{\text{vis}}, \mathbf{z_k}^{\text{text}} \rangle. \tag{7}$$

Each region is thus classified as the most similar discovered category, and detections are scored according to the similarity values. This approach allows the detector to operate without category priors, using only the vocabulary discovered from captions and a frozen *off-the-shelf* open-vocabulary detector.

4 Experiments

4.1 Experimental Setup

We evaluate our method on the MS COCO dataset [16], a widely used benchmark for object detection that contains more than 120,000 images, and over 860,000 annotated object instances spanning 80 object categories. The images in COCO are collected from complex everyday scenes. Since our method is training-free, we do not make use of the training split. All evaluations are performed on the COCO-2017 validation set, which contains 5,000 images and approximately 36,000 annotated object instances.

We assess the performance of our method in a fully zero-shot setting, without access to predefined category priors. This poses a challenge for evaluation, as our discovered vocabulary —obtained through the image captioning process— may not align directly with the canonical COCO category names. To address this, we employ a large language model, specifically LLaMA 3.3 [7], to map each discovered term to its most semantically similar COCO class. This mapping is performed automatically and is used solely for evaluation, without influencing the detection process. For performance assessment, we use the standard COCO evaluation protocol and report the average precision (AP), as well as AP at Intersection-over-Union (IoU) thresholds of 0.5 (AP_{50}) and 0.75 (AP_{75}).

4.2 Implementation details

For the vocabulary discovery stage, we use LLaVA-Video-7B-Qwen2 [34,17] as our captioning model $\mathcal{C}$. To guide the model towards producing object-centric outputs, we employ the following prompt:

```
Analyze the image and provide a list of all object categories
present. Then, based on your understanding of the scene, extend
this list by including other categories of objects that might
normally appear in similar contexts. Return all the category
names, separated by semicolons (;).
```

This prompt not only encourages precise object name extraction but also implicitly introduces a form of vocabulary augmentation by leveraging the model's contextual understanding to suggest additional plausible categories beyond those explicitly visible.

In the detection stage, we use Grounding DINO-SwinT [18] as our open-vocabulary object detector. It employs a dual-encoder architecture, where the visual encoder, ϕ_{vis}, is Swin-Tiny [20], and the textual encoder, ϕ_{text}, is BERT-Base [3].

4.3 Results

The results in Table 1 highlight a clear performance gap between our method and existing MLLM-based baselines. Both LLaVA-Video and Idefics3, when paired

	Method	AP	AP_{50}	AP_{75}
Free vocab.	LLaVA-Video-7B-Qwen2 [34] + CLIP [25] scoring	0.4	1.8	0.1
	Idefics3-8B-Llama3 [10] + CLIP [25] scoring	0.2	0.3	0.1
	KOSMOS-2 [24]	9.6	15.0	10.2
	OURS	**40.5**	**51.9**	**44.6**
Known vocab.	G-DINO (upper limit)	55.7	72.8	61.3

Table 1. SOTA comparison. Our method significantly outperforms general-purpose and localization-grounded Multimodal Large Language Models (MLLMs) under vocabulary-free conditions. G-DINO with access to the category vocabulary is reported as an upper bound.

with CLIP-based scoring, achieve very low AP scores (0.4 and 0.2, respectively). Since these models do not produce native confidence scores —which are essential for ranking detections in AP computation— we calculated them by computing the cosine similarity between the predicted region and its label embedding using CLIP, normalized between 0 and 1. While this allows for rough comparability, it does not overcome the fundamental limitations of these models in precise object localization.

KOSMOS-2, which is explicitly trained for localization-grounding, performs notably better with an AP of 9.6. However, it still falls far short of our method, which combines caption-based vocabulary discovery with G-DINO and achieves an AP of 40.5. This constitutes over a 4× improvement over KOSMOS-2, demonstrating the strength of dynamically adapting the detection vocabulary to the image content via captioning. The improvement is consistent across AP_{50} and AP_{75}. We also report the performance of G-DINO with access to the full ground-truth vocabulary, serving as an upper bound. This oracle setting achieves the highest AP (55.7), but our method recovers a substantial portion of this performance —despite having no prior knowledge of the object categories.

Finally, Fig. 2 provides a qualitative analysis showcasing the vocabulary discovery process and the final OvOD detections. In overall, these results validate our method as an effective way to equip open-vocabulary object detectors with the ability to perform detection without predefined categories in a training-free manner.

5 Conclusions

In this work, we have presented a novel, training-free framework that adapts open-vocabulary object detectors to operate without relying on predefined category priors. By leveraging image captioning to dynamically derive a relevant vocabulary for each image, our approach enables object detection in a fully category-free, zero-prior setting. We demonstrate that this strategy not only

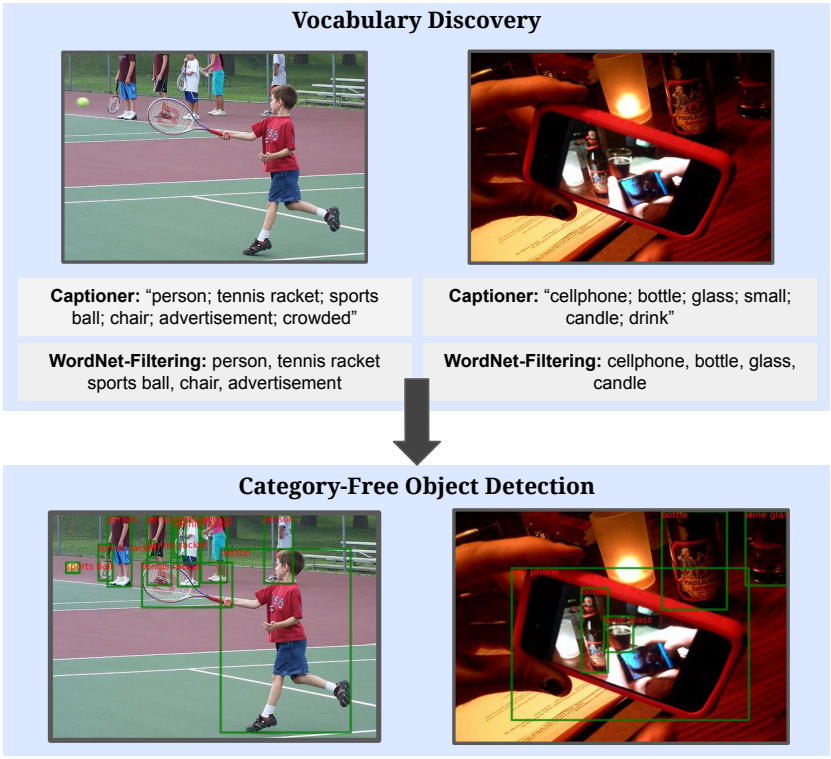

Fig. 2. Qualitative analysis showing both the vocabulary discovery process and the final OvOD detections with a confidence greater than 0.5.

removes the need for external supervision but also significantly outperforms strong multimodal baselines in zero-shot detection tasks, recovering a substantial portion of the upper-bound performance achieved by models with access to ground-truth vocabularies. Our results highlight the potential of using language-driven methods for both discovering and semantically grounding unknown object categories, pointing toward more scalable and adaptive solutions for real-world detection challenges.

Acknowledgements

This work was partially supported by the Spanish Ministerio de Ciencia e Innovación (grant numbers PID2020-112623GB-I00, PID2023-149549NB-I00), and the Galician Consellería de Cultura, Educación e Universidade (2024-2027 ED431G-2023/04). These grants are co-funded by the European Regional Development Fund (ERDF). Pablo Garcia-Fernandez is supported by the Spanish Ministerio de Universidades under the FPU national plan (grant number FPU21/05581).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023)
2. Bendale, A., Boult, T.: Towards open world recognition. In: CVPR (2015)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
4. Dhamija, A., Gunther, M., Ventura, J., Boult, T.: The overlooked elephant of object detection: Open set. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1021–1030 (2020)
5. Fang, R., Pang, G., Bai, X.: Simple image-level classification improves open-vocabulary object detection. In: AAAI (2024)
6. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: ECCV (2022)
7. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
8. Gupta, A., Narayan, S., Joseph, K., Khan, S., Khan, F.S., Shah, M.: Ow-detr: Open-world detection transformer. In: CVPR (2022)
9. Han, J., Ren, Y., Ding, J., Pan, X., Yan, K., Xia, G.S.: Expanding low-density latent regions for open-set object detection. In: CVPR (2022)
10. Hugging Face: Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics (2023), accessed: 2025-04-30
11. Joseph, K., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: CVPR (2021)
12. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv:2209.15639 (2022)
13. 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: ICCV (2023)
14. 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: CVPR (2022)
15. 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. In: ICLR (2023)
16. 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: ECCV (2014)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
18. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In: ECCV (2024)
19. Liu, Y.C., Ma, C.Y., Dai, X., Tian, J., Vajda, P., He, Z., Kira, Z.: Open-set semi-supervised object detection. In: ECCV. Springer (2022)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)

21. Ma, C., Jiang, Y., Wen, X., Yuan, Z., Qi, X.: CoDet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In: NeurIPS (2023)
22. Miller, D., Nicholson, L., Dayoub, F., Sünderhauf, N.: Dropout sampling for robust object detection in open-set conditions. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 3243–3249. IEEE (2018)
23. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: ECCV (2022)
24. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
25. 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: ICML (2021)
26. Wu, J., Li, X., Yuan, S.X.H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., Ghanem, B., Tao, D.: Towards Open Vocabulary Learning: A Survey. PAMI (2024)
27. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: CVPR (2023)
28. Wu, S., Zhang, W., Xu, L., Jin, S., Li, X., Liu, W., Loy, C.C.: Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In: ICLR (2024)
29. Wu, X., Zhu, F., Zhao, R., Li, H.: Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In: CVPR (2023)
30. Wu, Y., Zhao, X., Ma, Y., Wang, D., Liu, X.: Two-branch objectness-centric open world detection. In: Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis. pp. 35–40 (2022)
31. Wu, Z., Lu, Y., Chen, X., Wu, Z., Kang, L., Yu, J.: Uc-owod: Unknown-classified open world object detection. In: ECCV. Springer (2022)
32. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: CVPR (2023)
33. Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C., Xu, H.: Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In: NeurIPS (2022)
34. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data, 2024g. URL https://arxiv. org/abs/2410.02713
35. Zheng, J., Li, W., Hong, J., Petersson, L., Barnes, N.: Towards open-set object detection and discovery. In: CVPR (2022)
36. 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: CVPR (2022)
37. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022)
38. Zhu, C., Chen, L.: A survey on open-vocabulary detection and segmentation: Past, present, and future. IEEE TPAMI (2024)
39. Zohar, O., Wang, K.C., Yeung, S.: Prob: Probabilistic objectness for open world object detection. In: CVPR (2023)