ISRG Journal of Engineering and Technology (ISRGJET)

ISRG PUBLISHERS

Abbreviated Key Title: ISRG J Eng Technol

ISSN: 3107-5894 (Online)

Journal homepage: https://isrgpublishers.com/isrgjet/

Volume – II Issue- IV (July – August) 2026

Frequency: Bimonthly

Reliability Failure Modes of VLM-Based Open-Vocabulary Object Detection: A Narrative Review of Region-Text Alignment and Prompt Sensitivity

Sibtain Abbas[1], Saqlain Abbas[1*], Imtiaz Hussain[1], Mehdi Ali[2], Musawer Hussain[3]

[1] School of Mathematics, Hohai University, Nanjing, 210098, China

[2] Faculty of Computing, Shifa Tameer E Millat University, Islamabad, 44000, Pakistan

[3] College of Computer Science and Software Engineering, Hohai University, Nanjing, 210098, P.R. China

| **Received:** 12-08-2026 | **Accepted:** 16-08-2026 | **Published:** 25-08-2026

***Corresponding author: Saqlain Abbas**

Abstract

Open-vocabulary object detection (OVD) extends object detection beyond fixed category labels by allowing models to detect objects specified through class names, phrases, or natural-language descriptions. Vision-language models (VLMs) have enabled this transition by aligning visual and textual representations and supporting text-conditioned detection at inference time. However, strong benchmark performance does not guarantee reliable behavior. VLM-based open-vocabulary detectors remain vulnerable to failures that arise when image-level language supervision is transferred to region-level localization and when small changes in prompt wording alter detector output. This narrative review synthesizes representative work on CLIP, RegionCLIP, Detic, OWL-ViT, GLIP, Grounding DINO, OWLv2, YOLO-World, region-alignment methods, and recent prompt-sensitivity studies. Complementing broad task- and method-centered OVD surveys, the review uses reliability failure mechanisms as the organizing lens and jointly examines region-text misalignment and prompt-conditioned instability. The main failure patterns include incorrect region-text association, missed detections, false positives, weak localization, synonym instability, and inconsistent responses to semantically related prompts. Conventional metrics such as AP and mAP remain necessary but insufficient because they do not directly measure prompt robustness, alignment stability, or unsupported-detection risk. To operationalize reliability analysis, the review synthesizes complementary evaluation of selection consistency, box consistency, confidence stability, prompt-performance dispersion, generic-to-specific behavior, and absent-object false positives. Future progress should emphasize region-level alignment, transparent prompt reporting, uncertainty estimation, and qualitative failure analysis so that OVD systems become not only capable but also stable, interpretable, and practically reliable.

Keywords: *open-vocabulary object detection; vision-language models; region-text alignment; prompt sensitivity; prompt robustness; reliability; failure modes*

DOI: 10.5281/zenodo.22090248

1. Introduction

Object detection is a fundamental task in computer vision that requires both localization and semantic labeling of objects in an image. Closed-set detection has evolved from region-proposal-based two-stage detectors like Faster R-CNN to end-to-end transformer-based detectors such as DETR and DINO [1-3]. Most of these systems are trained and evaluated on detection datasets like COCO [4], LVIS [5], and Objects365 [6]. The closed-set paradigm has achieved strong results, but when actual users need to recognize new, unseen, domain-specific, and rare objects, these models may not generalize effectively. Open-vocabulary object detection (OVD) has been introduced to solve this problem. OVD does not try to limit the detector to a pre-defined set of labels, but instead, to localize the object specified in a short phrase or using a natural language description. This transition is tightly linked to large-scale Vision-Language Pretraining. The results of CLIP further illustrated the importance of transferable zero-shot recognition from natural-language supervision, and the study of multi-grained vision-language pretraining on scaling noisy image-text data and finer visual-language correspondences exhibited the benefits of training [7-9]. Later, OVD methods adopted these representations for, respectively, linking text embeddings to proposals, region features to proposals, detection queries to proposals, or cross-modal decoder outputs to proposals.

However, the reliability of these VLM-based open-vocabulary detectors is still not clear, even after significant advances. The difficulty of transferring image-level semantic knowledge into localized prediction has been detected in early OVD work, where different approaches have been utilized to link language with candidate regions for localized prediction (caption-supervised OVR-CNN, ViLD knowledge distillation, and RegionCLIP) [10-12]. Recently, there has been a growing number of approaches to this deficiency, such as frozen VLMs, region prompting, region-aware pretraining, localized contrastive learning, and key-feature alignment [13-17]. The alignment of regions and texts is a key reliability problem because of the mismatch between the image-level VLM pretraining and the region-level detection. Prompt sensitivity is a second core source of failure. In closed set detection, the set of categories is fixed by the model; in OVD, the prompt given by the user is part of the inference pipeline. Prompt learning has been demonstrated to have substantial impacts on vision-language transfer [18-22], and OVD-specific prompt methods further leverage the textual or multimodal prompt in region-level detection [23-25]. In an effort to directly test semantically overlapping grounding prompts, [26] found the mean instability using six such prompts was a 2.11 difference in selection, and the text-embedding proximity explained only 34% of the grounding disagreement. A 2026 peer-reviewed study also considers prompt sensitivity as a model-specific recognition task and demonstrates that prompt reformulation can boost open-vocabulary detection/segmentation accuracy when it is learned [27]. As a whole, these results suggest that prompt sensitivity is more than just background noise; it is a reliability problem in language-conditioned localization.

This review focuses on the failure modes of VLM-based open-vocabulary object detection, particularly region-text alignment and prompt sensitivity. There are existing surveys covering open-vocabulary (OVD) learning for detection and segmentation [28, 29], open-vocabulary (OVS) learning by task and supervision strategy [30], open-vocabulary learning (OVD) methods and visual-textual knowledge transfer [31], and multimodal large vision-language models (LMVLMs) for object detection [32]. Therefore, the novelty of this review does not lie in being the first review of OVD or VLM-based detection. It is a reliability-centered synthesis analysis, wherein the unit of analysis is the failure mechanisms, linked to the region-text misalignment, and then mapped to the observable detection failures, and the complementary evaluation checks, beyond single-template AP/mAP, are operationalized. This review is structured as follows: A method of narrative review, along with a literature identification approach, is described in Section 2. The conceptual and analytical framework for the analysis is based on the concepts and analysis provided in Section 3. The literature pertinent to the topic is summarized in Section 4, which has six thematic clusters: foundations for VLM, early architectures for OVD, grounding-based approaches, scaling strategies, real-time detection, and prompt sensitivity as a reliability problem. Section 5 brings together the key failure modes and implications for evaluations. Section 6 considers the implications, limitations of the evidence reviewed, and directions for reliability-driven research. Specific recommendations for future work are provided in the last part of Section 7.

2. Review Methodology

2.1 Review Design

The study design of this article is a narrative review. The goal is not to perform a meta-analysis or to list all published OVD studies but to summarize representative and influential studies that help to elucidate the common trustworthiness issues in open-vocabulary object detection. The review is organized around conceptual themes: VLM foundations, open-vocabulary detection architectures, visual-textual alignment failures, prompt sensitivity, evaluation gaps, and reliability-oriented mitigation strategies.

2.2 Literature Identification Strategy

Likewise, literature was collected by using specific search terms in the following databases: Google Scholar, arXiv, IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, CVF Open Access, and OpenReview. The search focused on publications released since 2021 due to CLIP's breakthrough of offering scalable image-text pretraining for zero-shot and open-vocabulary image recognition [8]. Earlier closed-set detection concepts are mentioned only when they help to explain the change from fixed-vocabulary detection to language-conditioned detection. The literature scope was brought up to date as of 10 August 2026, with specific reviews of the latest literature about OVD/VLM reviews and prompt-robustness studies to reinforce the scope and novelty argument.

2.3 Search Terms

Words used to search included “open-vocabulary object detection”, “vision-language object detection”, “open-set object detection”, “visual-textual alignment”, “region-text alignment”, “prompt sensitivity”, “text-conditioned object detection”, “CLIP object detection”, “Grounding DINO”, “OWL-ViT”, “OWLv2”, “YOLO-World”, “GLIP”, “RegionCLIP”, and “Detic”. Foundational papers were identified, then traced backwards and forwards to identify additional studies.

DOI: 10.5281/zenodo.22090248

2.4 Selection Logic

Studies were included only if they proposed an open-vocabulary or language-conditioned detection method, relied on VLM representations for detection or grounding tasks, demonstrated failure modes related to alignment or prompts, or provided review-level context for VLM-based detection. Studies in which image classification, captioning, pure segmentation, medical diagnosis, or domain adaptation were the main focus, or that were not clearly related to open-vocabulary object detection, were excluded. Foundational vision-language prompt-learning studies were retained when directly relevant to understanding prompt sensitivity in OVD.

2.5 Narrative Synthesis Procedure

The literature that was identified was summarized based on five cross-cutting themes for analysis: foundational image-text representation learning, OVD architectures, region-level visual-textual alignment, prompt-induced instability, and reliability-oriented evaluation and mitigation. The methodological contribution was investigated for each theme as well as the implications of failure, as the objective of this review is to gain insight into reliability, not simply to report on improvement in performance.

2.6 Methodological Limitations

Due to its narrative nature, the selection of the studies was carried out taking into account the analysis themes rather than the use of a screening methodology. It is not the intent of this present review to provide an extensive coverage of all studies on OVD. Furthermore, the review was focused on studies published from 2021 to 2026 to emphasize recent developments in VLM-based OVD. These design decisions facilitate conceptual synthesis but limit the reproducibility and comprehensiveness expected of a systematic review.

3. Conceptual Background and Analytical Framework

3.1 Closed-Set Detection and the Open-Vocabulary Shift

Typically, object detection is considered a supervised closed-set problem. A detector is trained with images labeled with bounding boxes and category names, and at inference, it only predicts from the pre-trained categories. Two-stage detectors and transformer detectors are examples of detectors that are built upon this fixed-vocabulary assumption, as shown in the traditional recognition pipelines [1-3]. Open-vocabulary detection is an extension of this, where the categories of objects are provided in text at test time.

3.2 From Zero-Shot Recognition to Open-Vocabulary Detection

Zero-shot recognition enables a model to recognize categories not explicitly seen during task-specific training by using semantic information such as text labels or language embeddings. OVD further takes the concept of classification and applies it to localization. The model has to not only determine the presence of a concept, but also the spatial region to which a concept applies. This is more challenging than zero-shot classification, as it involves semantic matching as well as accurate visual grounding.

3.3 Vision-Language Models as the Foundation of OVD

VLMs enable OVD by learning shared representations across visual and textual modalities. Large-scale image-text pairs are leveraged for CLIP [7-9], which is used to learn transferable visual concepts; ALIGN shows the effect of scaling noisy web supervision; multi-grained vision-language pretraining explicitly introduces finer visual-concept alignment. OVD methods build on this foundation by connecting text embeddings to region proposals, detection heads, object queries, or cross-modal attention modules. The main design challenge is to provide image-level semantic representations that are useful for object-level localization.

3.4 The Region-Level Alignment Problem

A central conceptual issue is that many VLMs are pretrained with image-level supervision, whereas object detection requires region-level alignment. To this end, RegionCLIP directly tackled the problem of region-text representation learning [12]. Subsequent methods addressed this gap through region prompting and anchor pre-matching [17], bag-of-regions alignment [33], region-aware pretraining [15], object-aware distillation [34], detector-internal region-word alignment [35], localized contrastive pretraining [13], and key-feature alignment [14]. The challenge is not only to detect the word "dog" or “vehicle”, but to correlate the concept with the proper visual area in a plausible ambiguous context.

3.5 Prompt-Conditioned Detection

In OVD, the prompt is not merely a label but an explicit input to the inference process. These cues can cause different localization and confidence behavior, with the same visual target being referred to by two or more different terms. Prompt learning approaches like CoOp, CoCoOp, MaPLe, PromptSRC, and knowledge-guided context optimization have shown that the behavior of VLM is highly dependent on how linguistic context is parameterized [18-22]. OVD-specific approaches including DetPro, PromptDet, and scene-adaptive region-aware prompting extend this insight to detection [23-25]. Prompt robustness should therefore be treated as part of detector reliability, not as a secondary implementation detail.

3.6 Analytical Framework

Figure 1 summarizes the analytical framework used in this review. VLM-based OVD is broken down into five functional components for analysis, instead of being taken as given to be a single sequential architecture. First, a visual encoder (usually a Vision Transformer or ResNet backbone) extracts spatial features from the input image. Second, candidate regions or object queries represent potential targets for localization. Third, a text encoder takes the input prompt, which is either a class name, a phrase, or a referring expression, and outputs a text representation. Fourth, region text matching/cross-modal fusion rates the similarity between the text meanings and the localized visual representations. Fifth, the detector produces bounding boxes, category labels, and confidence scores.

The functional components can fail. Small and obscured objects could have weak or spatially imprecise visual features. The proposal generator or learned query might not be able to describe the target region. Text encoders may be different for synonyms, or may overemphasize fine-grained attributes. Several factors can cause cross-modal matching to select the wrong region: background contamination, context bias, semantic ambiguity, and domain mismatch. Unstable or overconfident detections can then be achieved in the final scoring. A conceptual decomposition is used to direct the failure-mode analysis in Sections 4 and 5.

DOI: 10.5281/zenodo.22090248

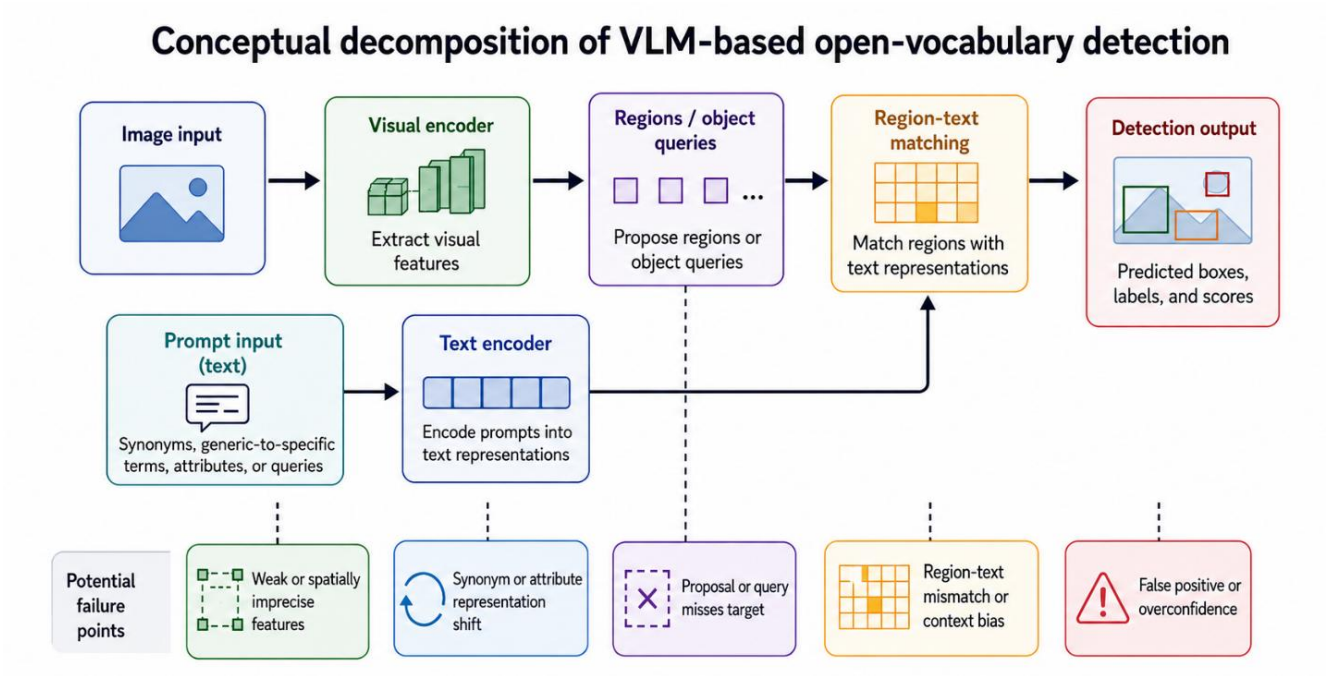


Figure 1. Conceptual decomposition of VLM-based open-vocabulary object detection, showing the interaction among visual encoding, region or object-query generation, text encoding, region-text matching, and detection output. Representative reliability failure points include weak or spatially imprecise features, synonym or attribute representation shifts, missed proposals or queries, region-text mismatch or context bias, and false positives or overconfidence.

4. Review of Related Literature

The literature is reviewed in six thematic clusters that are designed to follow the evolution of VLM-based OVD methods and the unique aspects of reliability concerns introduced by each thematic cluster. The subsections are organized by fundamental image-text representation learning (4.1), early detection architectures (4.2), grounding-based methods (4.3), scaling strategies (4.4), and real-time detection (4.5), and finally addressing the growing problem of prompt sensitivity as a reliability concern (4.6). This review is concluded with a comparative discussion of this review to previous surveys (4.7). Table 1 (in Section 4.1 at the end of the section) lists representative methods discussed.

4.1 Vision-Language Foundations for Open-Vocabulary Detection

Large-scale image-text pretraining plays a key role in the development of OVD. However, CLIP broke out of the fixed-label paradigm by learning visual representations from natural-language supervision, and multi-grained vision-language pretraining showed complementary scaling and alignment strategies [7-9]. Semantic image-level alignment alone, however, does not give object-level localization. Similarly, distributional-robustness analysis on CLIP further reveals that the underlying pretraining data significantly impacts downstream robustness, highlighting the importance of separating representation breadth from reliable localized behaviour [36].

This distinction raised the critical research question for OVD: What adaptations are needed for image-text representations so that they could be used at the regional level for localized detection? The literature has answered through caption-supervised detection, VLM knowledge distillation, region-level pretraining, prompt learning, conditional matching, pseudo-box supervision, dictionary-enriched pretraining, self-training, grounding-based fusion, and efficient one-stage detection [10-12, 23, 24, 37-43]. Although these routes are significant advances, they can still fail in various ways when the evidence and the text's prompts are not consistently aligned.

Table 1: Summary of representative VLM-based open-vocabulary object detection methods and their reliability implications.

Method	Main Idea	Training	Prompt Mechanism	Contribution	Reliability Concern
CLIP	Image-text contrastive pretraining	Large-scale image-text pairs	Text labels/templates	Foundation for zero-shot recognition	Image-level alignment does not localize objects directly
RegionCLIP	Region-level CLIP-style representation learning	Region-text pairs generated with CLIP guidance	Class names/text concepts	Addresses region-level alignment	Region proposals and captions can still be noisy
Detic	Detector vocabulary expansion using image-level labels	Detection boxes plus classification data	Large label vocabulary	Extends detectors to many categories	Vocabulary expansion does not guarantee correct region-text grounding

DOI: 10.5281/zenodo.22090248

OWL-ViT	Vision Transformer adapted to text-conditioned detection	Contrastive pretraining plus detection fine-tuning	Free-text queries	Simple and influential OVD recipe	Prompt wording and fine-grained localization remain challenging
GLIP	Unifies object detection and phrase grounding	Detection, grounding, and pseudo-grounding data	Phrases/categories	Learns object-level language-aware features	Quality of grounding supervision affects reliability
Grounding DINO	DINO detector with grounded pretraining and cross-modal fusion	Grounded pretraining and detection data	Category names and referring expressions	Strong open-set and grounding capability	Language-conditioned stages can amplify prompt ambiguity
OWLv2 / OWL-ST	Scales OVD through self-training and pseudo boxes	Large-scale pseudo-annotated image-text data	Text-conditioned detection	Improves rare-category performance at scale	Pseudo-label noise and prompt instability can persist
YOLO-World	YOLO-style efficient open-vocabulary detection	Detection, grounding, and image-text pretraining	Prompt-then-detect workflow	Practical real-time OVD direction	Speed does not solve alignment or prompt robustness

4.2 Region-Level Alignment and Early OVD Methods

One of the significant solutions to the gap between whole-image CLIP pretraining and object-level localization was RegionCLIP. It generalized the CLIP-style image region representation learning to the image region and connected the local visual features with text-based concepts [12]. Some of the previous works on early OVD include caption supervision [11], vision-language knowledge distillation [10], conditional transformer matching [43], hierarchical one-stage distillation [40], and pseudo bounding-box labels [38]. Their shared lesson is that semantic generalization must be spatially grounded: a detector that recognizes a concept at the image level can still fail if the corresponding region is weak, incomplete, or confused with context.

From a complementary perspective, Detic extended the vocabulary of the detectors with the image-level classification data. Detic leverages image-level classification data containing many more categories than standard detection datasets and demonstrated improvements on open-vocabulary and long-tail benchmarks, including gains of 2.4 mAP on all classes and 8.3 mAP on novel classes in open-vocabulary LVIS [44]. However, expanding the detector vocabulary does not necessarily guarantee correct localization or accurate region-text matching.

OWL-ViT also took text-conditioned object detection one step further by using a Vision Transformer pretrained with contrastive image-text learning and fine-tuned specifically for detection [45]. Open-vocabulary detection is also addressed in OV-DETR by conditioning the language and matching the conditioned target with the detector output [43]. These methods, in combination with RegionCLIP and Detic, demonstrate the possibility of implementing OVD, albeit with the assumption of sufficient semantic coverage and stable grounding at the region level.

The second generation of methods specifically targeted VLM representations to regions. The vision-language backbone is kept frozen in F-VLM [16], region prompting and anchor pre-matching are added with CORA [17], bag of regions is matched with language in BARON [33], and region-aware pretraining increases the locality of ViT representations [15]. [34] and [46] propose Object-Aware Distillation Pyramid and DK-DETR, respectively, that transfer visual-linguistic knowledge into detector-specific features. Architecturally, these approaches vary, but they all share the same reliability goal: that the semantic transfer is done at the semantic level and that the spatial and contextual information is preserved such that the intended region can be distinguished from objects nearby and from the background.

4.3 Grounded Language-Image Pretraining and Language-Conditioned Detection

GLIP is a significant grounding-based method as it combines the two aspects of object detection and phrase grounding in the pretraining stage. Rather than a fixed category classification problem, GLIP views object detection as a grounding problem wherein text concepts or text phrases should be matched to regions in the image. Grounding supervision is scaled based on detection and grounding data and pseudo-grounding data on image-text data [47]. The concepts in DetCLIP and DetCLIPv2 are also enriched by dictionary-level supervision, and word-region alignment is increased with a scalable approach, demonstrating that the quality and diversity of region-language supervision are crucial to open-vocabulary transfer [42, 48].

Building upon the previous step, Grounding DINO further extended the detection method based on language conditioning by integrating the DINO transformer detector with a grounding pretraining strategy. It enhances visual features, selects queries using language guidance, performs cross-modal decoding, and can detect objects using category names or referring expressions [39]. Grounded pretraining is strong, as evidenced by the reported zero-shot results of 52.5 AP on COCO and 26.1 mean AP on ODinW. Similar to this, the Language-conditioned Detection Transformer (DECOLA) investigates language-conditioned detection from a dedicated transformer formulation [49]. These systems are particularly sensitive to the prompt since language is directly involved in region/query selection and multimodal fusion.

Grounding-based approaches therefore offer both benefits and reliability risks in open-vocabulary detection. Recent works extend the design space by integrating detector-internal region-word alignment, background-prompt learning, retrieval-augmented visual and language cues, semantic hierarchy information,

DOI: 10.5281/zenodo.22090248

language-model instruction, more versatile generative OVD, etc. [35, 50-55]. More recent systems include large-language-model-based supervision [56], localized CLIP pretraining [13], decoupled local representation learning [57], and visual-textual co-occurrence modeling [58]. The advances made in increasing capability can be compromised by the use of ambiguous, overly broad, overly specific, and poorly evidenced language. Thus, the stability of phrase grounding and OVD should be assessed in addition to average accuracy.

4.4 Scaling Open-Vocabulary Detection through Self-Training

OWLv2 addressed the limitation of a lack of box-annotated detection data by using self-training on large-scale image-text data to scale the detection training. In this framework, self-training is performed with OWL-ST as the recipe instead of the model name, and a new detector is trained with large-scale supervision (pseudo-boxes) generated by the existing detector [41]. Weak or web-scale supervision is also used to extend vocabulary coverage in earlier pseudo-box OVD and PromptDet [24, 38], while the study in SAS-Det explicitly analyzes how pseudo-box OVD self-training is destabilized by noisy pseudo-box OVD labels and changing pseudo-box OVD distributions [59].

The direction of scaling is significant as it affects the performance of OVDs, not just the architecture, but the amount and variety of region-level supervision. [41] showed that for an L/14 model, OWL-ST achieved an improvement of 43% relative to the original LVIS baseline model, resulting in a rare-class AP of 44.6, compared with 31.2 for the baseline model. But scaling does not remove failure modes. If pseudo-labels are noisy, incomplete, or biased, the model may learn unstable region-text associations. In addition to using more positive pseudo-labels, several approaches have been proposed to handle the negative/background information, such as retrieval-augmented losses/features and background-prompt learning [51, 52].

4.5 Real-Time Open-Vocabulary Detection

Recent research activities have been directed towards developing OVDs that are practically usable. YOLO-World introduces YOLO-style detection for open-vocabulary scenarios using vision-language modeling, a Re-Parameterizable Vision-Language Path Aggregation Network, and region-text contrastive learning [37]. The 35.4 AP it reported with a V100 GPU on LVIS at 52.0 FPS shows the direction of the ongoing research of OVD: accuracy-efficiency. Previous studies of hierarchical visual-language distillation also showed how to embed open-vocabulary transfer within a one-stage detector design [40].

Efficiency does not eliminate key reliability issues, though. Prompt and region-text matching are still required for a real-time detector. The model can make false positive or false negative predictions, or suffer from fluctuating confidence numbers when the prompt is imprecise, the object is small or hard to see, or the embedding of the text does not match the region feature. YOLO-World hence not only shows a step towards practical progress but also highlights the importance of assessing reliability when using prompt variation and visual-textual ambiguity.

4.6 Prompt Sensitivity as a Reliability Problem

One of the most significant issues that has arisen in OVD and visual grounding is prompt sensitivity. An ideal open-vocabulary detector should generate consistent detections for semantically similar prompts, given the same concept in the image. However, earlier prompt-learning research shows that linguistic context, conditioning, placement of the multimodal prompt, and regularization can greatly influence the behavior of VLM transfer [18-22]. Further examples include prompt methods that are specific to the OVD (DetPro, scene-adaptive region-aware prompting), which show that prompt construction is linked to region classification and alignment [23, 25]. [26] tested prompts like "a person", "a human", and "a pedestrian" and found that semantically overlapping prompts often elicited different instances of objects, with a mean of 2.11 distinct selected instances across six prompts. [27] reinforce prompt sensitivity as a performance issue for the model and demonstrate the capability of RL prompt reformulation to boost open-vocabulary recognition of detection and segmentation models.

It is this fact that confirms the main conclusion of the review that reliable language-conditioned detection is not an implementation detail but a fundamental reliability issue that requires explicit robustness evaluation. The observation of fine-grained OVD evaluation shows that existing detectors can perform poorly in the presence of subtle attribute discrimination requirements, whereas the observation of semantic-hierarchy methods demonstrates that behaviour can also vary with the granularity of the vocabulary [54, 60]. Where there are multiple boxes or confidence rankings corresponding to synonyms, broad category names, and descriptive phrases, it is hard to reproduce and compare the OVD results. Settings should thus be reported objectively and validated against semantically varied, controlled settings rather than optimized for a preferred setting.

4.7 Positioning and Novelty Relative to Existing Surveys

This topic has been addressed in several recent surveys, each using a different organizing lens. In this paper, [28] provide a survey of open-vocabulary learning in the context of detection, segmentation, other settings, benchmarks, and shared design themes. [29] offer a wide classification of open-vocabulary detection and segmentation tasks and supervision methods. [30] tie their work specifically to OVD and highlight the concepts of knowledge transfer, candidate-region design, visual-textual alignment, semantic enrichment, localization, robustness, and bias. [31] survey and assess VLM-based detection and segmentation in various downstream applications and fine-tuning levels, such as text-prompt adaptation. [32] offer a more comprehensive survey of multimodal large vision-language models in object detection, focusing on their architectures, adaptability, contextual reasoning, and their ability to generalize.

Table 2. Positioning of the present reliability-centered review relative to representative recent surveys.

Review	Scope	Primary organizing lens	Reliability emphasis	Relation to present review
[28]	Open-vocabulary learning across detection +	Cross-setting concepts, shared design themes, methods,	Broad challenges and future directions across	Broader field-level survey; not primarily

DOI: 10.5281/zenodo.22090248

	segmentation	benchmarks	open-vocabulary learning	organized around OVD reliability failure mechanisms or prompt-consistency testing
[29]	OVD + OVS across 2D/3D/video tasks	Task and supervision taxonomy; method families	Broad challenges, strengths, weaknesses, and benchmarks	Broader multi-task taxonomy; prompt consistency is not the primary organizing axis
[30]	OVD	Knowledge transfer, proposals, visual-textual alignment, semantic enrichment	Robustness, localization accuracy, generalization, and bias	Method-centered OVD survey; does not primarily organize the synthesis around paired alignment and prompt-instability failure mechanisms
[31]	VLM-based detection + segmentation	Cross-scenario evaluation and fine-tuning strategies	Evaluates robustness-related scenarios and text-prompt adaptation	Broader empirical evaluation; not primarily a failure-mode taxonomy for prompt-consistent OVD
[32]	Multimodal LVLMs for object detection	Architectures, fusion, adaptability, performance, applications	Discusses limitations and future robustness needs	Broader LVLM/object-detection scope; not primarily centered on region-text alignment plus prompt consistency
Present review	VLM-based OVD	Reliability failure mechanisms	Region-text misalignment, prompt instability, unsupported detections, metric gaps	Integrates the two reliability axes and synthesizes complementary prompt-consistency and failure-oriented checks

In this review, the authors make no claim to being the first to review the detection, visual-textual alignment, or prompt effects of OVD and VLM. A key difference is that this review treats region-text misalignment and prompt-conditioned instability as interacting failure mechanisms, links them to observable outputs such as wrong-region selection, unsupported detections, synonym instability, and confidence variation, and connects them to reliability-oriented evaluation checks. This failure-mechanism framing is meant to be used in addition to method taxonomies and general benchmark surveys, not in place of them.

5. Synthesis of Failure Modes and Evaluation Implications

5.1 Representative Cross-Study Empirical Findings

Table 3 consolidates representative empirical findings already discussed in the reviewed literature. The values are reported in their original study contexts and are not treated as a direct ranking because datasets, vocabulary splits, prompt protocols, hardware, and evaluation settings differ across papers.

Table 3. Representative cross-study empirical findings and their reliability implications.

Study / Method	Reported empirical finding	Evaluation context	Reliability implication
Detic [44]	+2.4 mAP for all classes and +8.3 mAP for novel classes.	Open-vocabulary LVIS.	Vocabulary expansion improves novel-category performance but does not itself guarantee correct region-text grounding.
Grounding DINO [39]	52.5 AP on COCO and 26.1 mean AP on ODinW.	Zero-shot/open-set detection benchmarks.	Strong aggregate capability can coexist with prompt sensitivity because language participates in query selection and cross-modal fusion.
OWL-ST / OWLv2 [41]	Rare-class AP for an L/14 model increased from 31.2 to 44.6 (about 43% relative improvement).	LVIS rare-category evaluation.	Large-scale pseudo-annotation improves rare-category localization, but pseudo-label noise and prompt instability can

DOI: 10.5281/zenodo.22090248

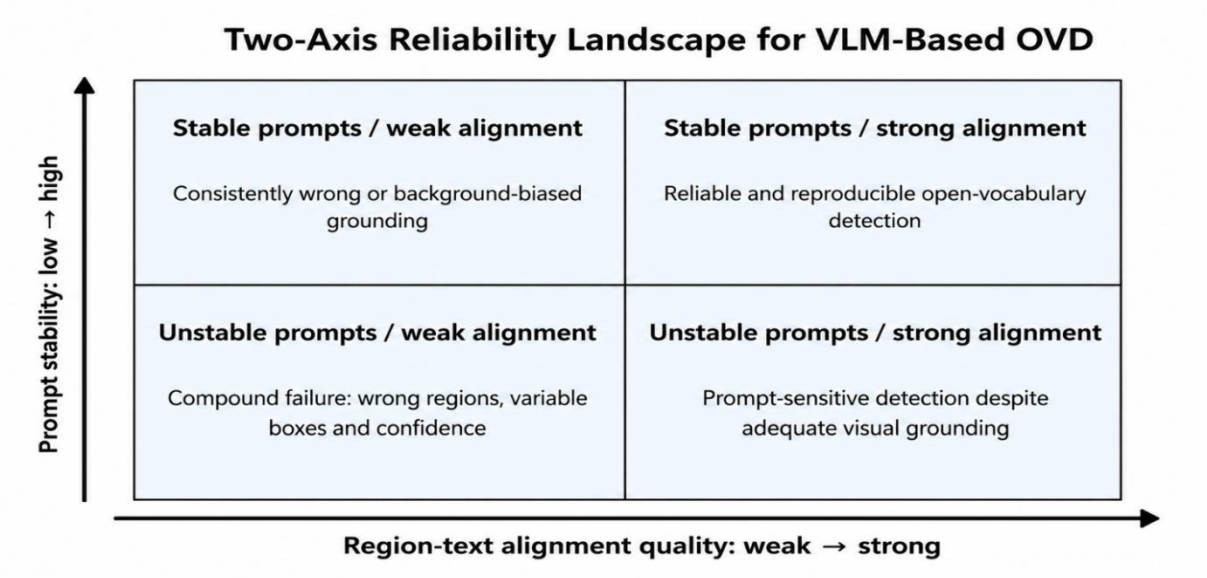
			remain.
YOLO-World [37]	35.4 AP at 52.0 FPS.	LVIS using a V100 GPU.	Efficiency improves deployment practicality, but speed does not resolve alignment or prompt-robustness failures.
Prompt sensitivity [26]	Mean instability of 2.11 distinct selections across six prompts; text-embedding proximity explained 34% of grounding disagreement.	Controlled prompt-variation grounding analysis.	Direct evidence that semantically related wording can change selected instances and that embedding similarity alone does not explain instability.
Fine-grained OVD [60]	Hard-negative and attribute-level vocabularies exposed weaknesses that standard OVD evaluation can mask.	Fine-grained OVD stress testing.	Aggregate AP should be complemented by fine-grained and hard-negative evaluation.
Prompt reformulation [27]	Learned prompt reformulation improved open-vocabulary recognition across detection and segmentation models.	Model-aware prompt adaptation.	Prompt wording is an actionable performance variable, reinforcing transparent prompt reporting and robustness testing.
CLIP robustness [36]	Distributional robustness was strongly shaped by the underlying pretraining data.	Contrastive language-image pretraining robustness analysis.	Broad semantic transfer should not be assumed to imply stable downstream or localized reliability.

5.2 Alignment Failures

Failure to align a textual concept to the intended visual region is an example of alignment failure. The detector may align the text with only part of an object, select a visually similar object, match a background region, or assign a semantically related label to the wrong region. This gap is mitigated by various methods such as region-level pretraining, prompting, bag-of-regions alignment, detector-specific distillation, and region-word alignments built into the model [12, 15, 17, 33, 35, 46]. More recent methods continue to address this problem through localized pretraining, decoupled dense features, and key-feature alignment [13, 14, 57]. None of these design families ensures 100% grounding in clutter, domain shift, or unclear language.

5.3 Prompt-Induced Instability

Prompt-induced instability is when various prompts or similar prompts result in different detections. Long prompts can lead to more recall but to false positives; short prompts can lead to less ambiguity but to loss of objects when fine-grained attributes are not encoded in the visual representation. The sensitivity of VLM transfer to text context is shown through prompt-learning results on CoOp, CoCoOp, MaPLe, PromptSRC, and knowledge-guided context optimization [18-22]. This effect is intertwined with region/query ranking and fine-grained vocabulary structure in OVD [26, 54, 60]. Reliable OVD therefore requires stable prompt-conditioned matching in addition to strong visual detection. Figure 2 shows the summary of this 2-axis reliability perspective, where prompt stability and region-text alignment are considered to be distinct but correlated properties.



DOI: 10.5281/zenodo.22090248

Figure 2. Two-axis reliability landscape for VLM-based open-vocabulary object detection. Region-text alignment quality and prompt stability are treated as distinct but interacting dimensions of detector reliability.

5.4 False Positives, Hallucination, and Overconfidence

The danger of false positives in VLM-based detection is exacerbated due to the potential for language priors to overwhelm weak visual information. For this review, detection hallucination refers to an unsupported category prediction that is made for a queried concept that is not in the image, or there is only partial support for the concept. The methods that explicitly model background prompts or retrieve semantically related negative classes emphasize the significance of negative evidence in OVD [51, 52]. Similarly, fine-grained hard-negative evaluation reveals that semantically plausible alternatives can reveal recognition errors not apparent under easier vocabularies [60]. Evaluation should therefore be conducted in a reliability-oriented way, implying the use of absent-object queries, hard negatives, uncertainty-aware reporting, and qualitative inspection.

5.5 Metric Gaps

Traditional metrics like AP, mAP, AP50, AP@.50:.95, AP-base, and AP-novel are still required to quantify the detection quality. They do not directly address the robustness of a model to prompt variation or whether semantically equivalent prompts align with the same object regions. Strong standard OVD performance is compatible with weaknesses under hard negative and attribute-level vocabularies [60], and controlled grounding experiments have directly revealed synonym-induced selection instability [26]. In general, the results of CLIP robustness studies indicate that the pre-training data can heavily influence the distributional behavior [36]. Prompt-consistency evidence needs to be reported alongside standard detection measurements as a part of OVD evaluation.

This review builds on and formalizes the concept of observing instability in observations like those reported by [26] and complements fine-grained stress testing like [60], and summarizes five operational measures of instability for reliability-oriented prompt evaluation: (1) selection consistency, defined as the proportion of semantically equivalent prompts that select the same object instance; (2) box consistency, summarized by the IoU of matched boxes across prompt variants; (3) confidence stability, measured by the dispersion of confidence scores across semantically related prompts; (4) prompt-performance dispersion, measured by the variance or range of AP across a predefined prompt set; and (5) generic-to-specific retention, measured by the fraction of valid detections that are maintained as prompts become more specific. In the case of absent-object queries, an explicit false-positive rate should also be reported. They are not community standards or entirely new metrics but rather are proposed as a synthesis and operationalization of reliability checks motivated by previous prompt-sensitivity and fine-grained evaluation findings. This synthesis is then implemented as an operationalized controlled evaluation workflow in Figure 3 with semantic prompt variants and object-absent queries.

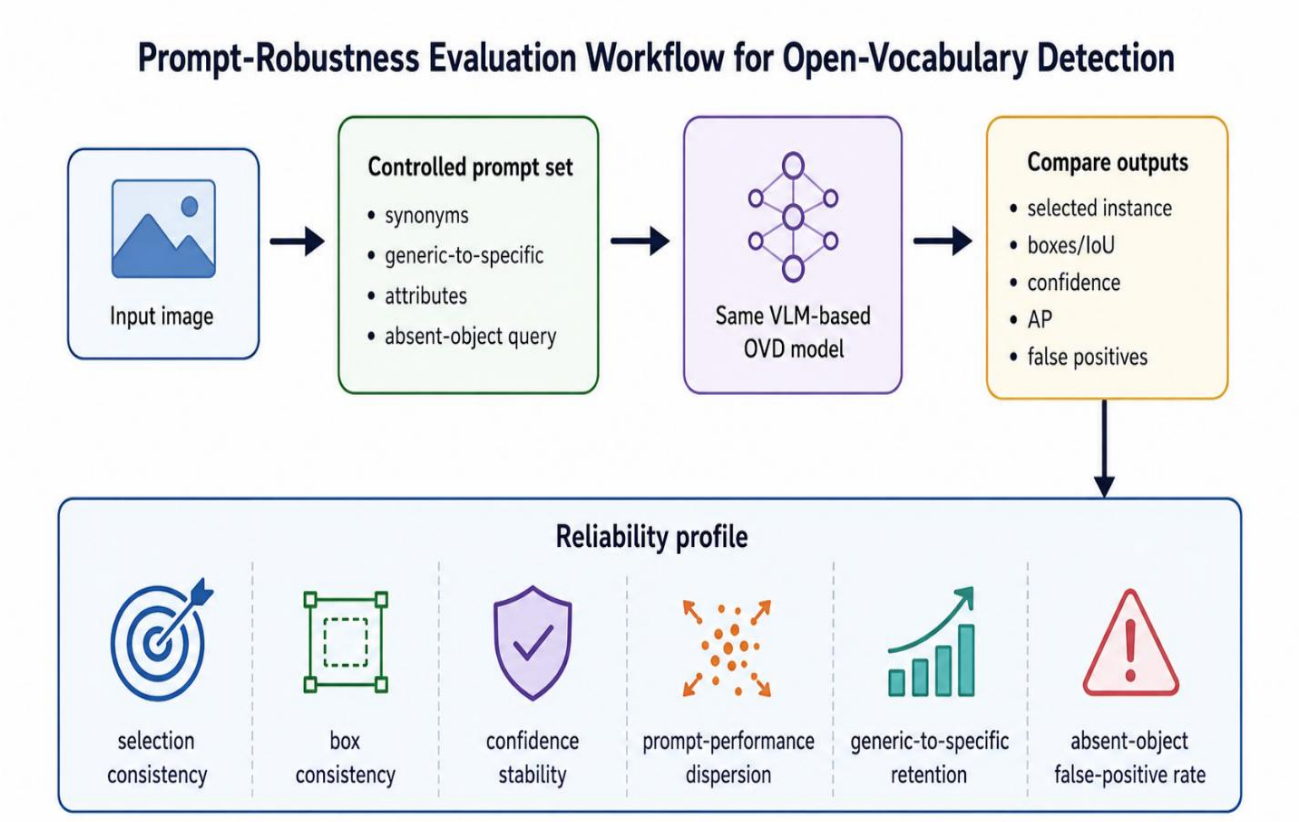


Figure 3. Prompt-robustness evaluation workflow for VLM-based open-vocabulary object detection. Controlled prompt variants are evaluated using the same model and summarized through selection consistency, box consistency, confidence stability, prompt-performance dispersion, generic-to-specific retention, and absent-object false-positive rate.

DOI: 10.5281/zenodo.22090248

Table 4. Reliability-centered failure taxonomy and suggested evaluation checks for VLM-based open-vocabulary object detection.

Failure Mode	Typical Cause	Observable Output	Suggested Evaluation Check
Region-text mismatch	Insufficient or noisy region-text supervision; context/background bias; cross-modal mismatch	Wrong object or background box is selected	Region-level grounding tests and qualitative box inspection
Missed detection	Small, occluded, rare, or poorly represented object	Object is present but no valid box is returned	Rare-class AP, small-object AP, recall by prompt group
Prompt synonym instability	Prompt-dependent text representations and changes in region or query ranking	Synonyms produce different boxes or confidence scores	Prompt sets with semantically equivalent terms
Over-specific prompt failure	Fine-grained attributes not encoded in region features	Specific phrase fails although generic category works	Generic-to-specific prompt ladder
False positive / hallucination	Language prior dominates weak visual evidence	Detector returns unsupported object box	Absent-object queries and object-absent evaluation
Metric masking	Fixed prompt template and aggregate AP reporting	High mAP but poor prompt robustness	Prompt-consistency metrics plus AP/mAP

6. Discussion

Reviewed literature demonstrates that the OVD transforms the problem of object detection from a fixed-label classification and localization problem into a language-conditioned grounding problem. This transformation is significant because users can define categories flexibly, but it also changes the nature of reliability. Closed set detection is typically evaluated on the basis of classification errors, localization errors, or missed detections. Errors can also occur in VLM-based OVD because of prompt interpretation and region-text matching. Thus, reliability is influenced by the interplay between visual features, verbal cues, cross-modal matching, and detection thresholds.

One key message is that the semantic generalization at the image level is insufficient to achieve reliable detection. While visual representations can be transferred, CLIP-style pretraining does need localized grounding for object detection. RegionCLIP improves region-sensitive transfer through region-level representation learning, while CORA introduces region prompting and anchor pre-matching; BARON aligns bags of regions with language, region-aware pretraining improves the locality of visual representations, and object-aware distillation transfers vision-language knowledge into detector-specific features [12, 15, 17, 33, 34]. To make the mismatch between image and region levels even more explicit, localized contrastive learning and key-feature alignment are used in CLOC and OV-KFA, respectively [13, 14]. Future models should therefore take region-aware alignment as a primary design goal, and not as an additional adaptation step.

6.1 Reliability-Oriented Evaluation and Reporting

There are also methodological implications for reporting research that relate to prompt sensitivity. The results reported in the studies vary due to differences in prompt templates, category descriptions, semantic hierarchies, or prompt-learning strategies. Prompt learning has been shown to improve vision-language transfer through prompt-learning literature, but this does not necessarily mean that the optimized prompts are insensitive to the natural wording changes [18-22]. OVD studies should, therefore, report variation in boxes, labels, confidence scores, and AP across prompt variants and include multiple semantically related prompt formulations [26].

Care should be taken in assessing mitigation strategies. While average performance can be improved by prompt learning, prompt ensembling, background prompting, retrieval augmentation, semantic hierarchy, self-training, language-model-guided concept enrichment, and learned prompt reformulation, each of these methods may involve reliability trade-offs [23, 27, 41, 51, 52, 54, 56, 58, 59]. A larger training scale can lead to better performance on rare categories, but may not necessarily yield natural prompt consistency, and richer language priors can lead to higher semantic coverage but also require good negative evidence. The question of reducing failure diversity, rather than just increasing mAP, should thus be used to evaluate a method in a reliability-driven manner.

This discussion leads to a practical evaluation agenda for OVD. In addition to AP and mAP, future studies should consider selection and box consistency per synonym prompt, confidence stability, prompt performance dispersion, generic-to-specific prompt behavior, absence of object false positives, uncertainty calibration, and representative qualitative failures. This evidence could help to differentiate between models that simply can and models that can and do withstand application to real-world tasks like robotics, surveillance, remote sensing, inspection of industrial systems, and interactive search.

6.2 Limitations of the Evidence Base

The evidence synthesized in this review has several limitations. Firstly, the studies employ different data sets, different word definitions, different prompt formulations, different levels of confidence, and different types of evaluation measures, and the results of quantitative evaluations cannot be directly compared across articles. Second, the scope of existing reviews covers significantly wider families of tasks and organizing taxonomies than the narrower focus on reliability examined here, as found in [28-32]. Third, the OVD field is in a state of constant change: the outputs of recent research are both preprints and peer-reviewed conference/journal articles; the former should be read with caution. Finally, alignment and prompt sensitivity are explicitly covered, whereas other failure sources like the domain shift, long-tail

DOI: 10.5281/zenodo.22090248

imbalance, prompt noise, prompt calibration, and data leakage are not covered as extensively as in a comprehensive OVD survey.

7. Conclusion and Future Directions

In this narrative review, we examined reliability failure modes in VLM-based open-vocabulary object detection and the connection between region-text alignment and prompt sensitivity. The evidence reviewed demonstrates that the utility of OVD has been greatly enhanced in terms of detection, allowing models to detect objects whose descriptions are not fixed in a list of categories but in free-form language. This flexibility, however, introduces reliability costs that are not fully captured by standard benchmark metrics. The goal of this review is not, in essence, to exhaustively describe the taxonomy of methods to achieve alignment, but to synthesize the relationship between alignment and prompt formulation in order to observe the failures and how they can be assessed more directly.

The general message from the literature is that stable region-level grounding and prompt-conditioned matching are two essential properties of reliable OVD. Localized supervision has consistently been shown to be important [12, 13, 15, 17, 33], and controlled prompt-sensitivity, model-aware prompt adaptation, and fine-grained evaluations have demonstrated that quality and language robustness are distinct properties [26, 27, 60]. Benchmark accuracy is therefore a necessary but insufficient measure of deployment readiness, because a model can perform well on the benchmark but still fail on real-world tasks.

The review pinpoints a number of specific areas for future studies. First, it is essential to include region-level alignment as a first-class design goal to assess the grounding consistency of architectures and pretraining objectives beyond image-level transfer [13, 14]. Second, reporting requirements need to be in place: transparent reporting of prompt templates, semantically related sets of prompts, and differences between the prompt variants should be reported [26]. Third, evaluation suites should include synonym robustness, generic-to-specific prompt ladders, tests that include absent objects, hard negatives, and fine-grained vocabulary stress tests [60]. Finally, uncertainty estimation and control of unsupported detections need to be given more attention in safety-critical or high-consequence applications.

To sum up, the open-vocabulary paradigm has turned detection from a closed, label-based problem to an open, language-conditioned one. To make that transition reliable, evaluation must extend beyond single-template mAP to include prompt robustness, region-text alignment quality, uncertainty calibration, and qualitative failure analysis. The next generation of open-vocabulary detectors should be region-aware, prompt-robust, uncertainty-aware, and transparent on how to use language prompts. These properties are the conditions under which OVD can transition from a compelling research direction to a dependable technology for real-world deployment.

References

1. Carion, N., et al. *End-to-end object detection with transformers*. in *European conference on computer vision*. 2020. Springer.
2. Ren, S., et al., *Faster r-cnn: Towards real-time object detection with region proposal networks.* Advances in neural information processing systems, 2015. **28**.
3. Zhang, H., et al., *Dino: Detr with improved denoising anchor boxes for end-to-end object detection.* arXiv preprint arXiv:2203.03605, 2022.
4. Lin, T.-Y., et al. *Microsoft coco: Common objects in context*. in *European conference on computer vision*. 2014. Springer.
5. Gupta, A., P. Dollar, and R. Girshick. *Lvis: A dataset for large vocabulary instance segmentation*. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. IEEE.
6. Shao, S., et al. *Objects365: A large-scale, high-quality dataset for object detection*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. IEEE.
7. Jia, C., et al. *Scaling up visual and vision-language representation learning with noisy text supervision*. in *International conference on machine learning*. 2021. PMLR.
8. Radford, A., et al. *Learning transferable visual models from natural language supervision*. in *International conference on machine learning*. 2021. PmLR.
9. Zeng, Y., X. Zhang, and H. Li, *Multi-grained vision language pre-training: Aligning texts with visual concepts.* arXiv preprint arXiv:2111.08276, 2021.
10. Gu, X., et al., *Open-vocabulary object detection via vision and language knowledge distillation.* arXiv preprint arXiv:2104.13921, 2021.
11. Zareian, A., et al. *Open-vocabulary object detection using captions*. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. IEEE.
12. Zhong, Y., et al. *Regionclip: Region-based language-image pretraining*. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. IEEE.
13. Chen, H.-Y., et al., *Contrastive localized language-image pre-training.* arXiv preprint arXiv:2410.02746, 2024.
14. Jiang, Y., et al., *OV-KFA: Open-vocabulary object detection via key feature alignment.* Neurocomputing, 2025: p. 131790.
15. Kim, D., A. Angelova, and W. Kuo. *Region-aware pretraining for open-vocabulary object detection with vision transformers*. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. IEEE.
16. Kuo, W., et al., *F-vlm: Open-vocabulary object detection upon frozen vision and language models.* arXiv preprint arXiv:2209.15639, 2022.
17. Wu, X., et al. *Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching*. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. IEEE.
18. Khattak, M.U., et al. *Maple: Multi-modal prompt*

DOI: 10.5281/zenodo.22090248

learning. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. IEEE.

19. Khattak, M.U., et al. *Self-regulating prompts: Foundational model adaptation without forgetting*. in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. IEEE.
20. Yao, H., R. Zhang, and C. Xu. *Visual-language prompt tuning with knowledge-guided context optimization*. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. IEEE.
21. Zhou, K., et al. *Conditional prompt learning for vision-language models*. in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 2022. IEEE.
22. Zhou, K., et al., *Learning to prompt for vision-language models.* International journal of computer vision, 2022. **130**(9): p. 2337–2348.
23. Du, Y., et al. *Learning to prompt for open-vocabulary object detection with vision-language model*. in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 2022. IEEE.
24. Feng, C., et al. *Promptdet: Towards open-vocabulary detection using uncurated images*. in *European conference on computer vision*. 2022. Springer.
25. Zhao, X., et al. *Scene-adaptive and region-aware multi-modal prompt for open vocabulary object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
26. Deka, D.J., A. Sethi, and S.M. Ali, *Prompt Sensitivity in Vision-Language Grounding: How Small Changes in Wording Affect Object Detection.* arXiv preprint arXiv:2604.17126, 2026.
27. Avshalumov, M., et al., *Say it better: RL-based prompt tuning for enhancing open-vocabulary recognition.* Neurocomputing, 2026: p. 133353.
28. Wu, J., et al., *Towards open vocabulary learning: A survey.* IEEE transactions on pattern analysis and machine intelligence, 2024. **46**(7): p. 5092–5113.
29. Zhu, C. and L. Chen, *A survey on open-vocabulary detection and segmentation: Past, present, and future.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. **46**(12): p. 8954–8975.
30. Yu, X., Y. Wang, and H. Sun, *Open-vocabulary object detection survey*. Procedia Computer Science, 2025. **266**: p. 62–71.
31. Feng, Y., et al., *Vision-language model for object detection and segmentation: A review and evaluation.* arXiv preprint arXiv:2504.09480, 2025.
32. Sapkota, R. and M. Karkee, *Object detection with multimodal large vision-language models: An in-depth review*. Information Fusion, 2026. **126**: p. 103575.
33. Wu, S., et al. *Aligning bag of regions for open-vocabulary object detection*. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. IEEE.
34. Wang, L., et al. *Object-aware distillation pyramid for open-vocabulary object detection*. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. IEEE.
35. Zhang, H., et al. *Exploring region-word alignment in built-in detector for open-vocabulary object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
36. Fang, A., et al. *Data determines distributional robustness in contrastive language image pre-training (clip)*. in *International conference on machine learning*. 2022. PMLR.
37. Cheng, T., et al. *Yolo-world: Real-time open-vocabulary object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
38. Gao, M., et al. *Open vocabulary object detection with pseudo bounding-box labels*. in *European Conference on Computer Vision*. 2022. Springer.
39. Liu, S., et al. *Grounding dino: Marrying dino with grounded pre-training for open-set object detection*. in *European conference on computer vision*. 2024. Springer.
40. Ma, Z., et al. *Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
41. Minderer, M., A. Gritsenko, and N. Houlsby, *Scaling open-vocabulary object detection.* Advances in Neural Information Processing Systems, 2023. **36**: p. 72983–73007.
42. Yao, L., et al., *Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection.* Advances in neural information processing systems, 2022. **35**: p. 9125–9138.
43. Zang, Y., et al. *Open-vocabulary detr with conditional matching*. in *European conference on computer vision*. 2022. Springer.
44. Zhou, X., et al. *Detecting twenty-thousand classes using image-level supervision*. in *European conference on computer vision*. 2022. Springer.
45. Minderer, M., et al. *Simple open-vocabulary object detection*. in *European conference on computer vision*. 2022. Springer.
46. Li, L., et al. *Distilling detr with visual-linguistic knowledge for open-vocabulary object detection*. in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. IEEE.
47. Li, L.H., et al. *Grounded language-image pre-training*. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. IEEE.
48. Yao, L., et al. *Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment.*

DOI: 10.5281/zenodo.22090248

in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. IEEE.

49. Cho, J.H. and P. Krähenbühl. *Language-conditioned detection transformer*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
50. Du, P., et al. *Lami-detr: Open-vocabulary detection with language model instruction*. in *European Conference on Computer Vision*. 2024. Springer.
51. Kim, J., et al. *Retrieval-augmented open-vocabulary object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
52. Li, J., et al. *Learning background prompts to discover implicit knowledge for open vocabulary object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
53. Lin, C., et al. *Generative region-language pretraining for open-ended object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
54. Liu, M., et al. *Shine: Semantic hierarchy nexus for open-vocabulary object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
55. Yao, L., et al. *Detclipv3: Towards versatile generative open-vocabulary object detection*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.
56. Fu, S., et al. *Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models*. in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025. IEEE.
57. Wang, J., et al. *Declip: Decoupled learning for open-vocabulary dense perception*. in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025. IEEE.
58. Li, Y., J. Niu, and T. Ren. *Benefit from seen: Enhancing open-vocabulary object detection by bridging visual and textual co-occurrence knowledge*. in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025. IEEE.
59. Zhao, S., et al. *Taming self-training for open-vocabulary object detection*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
60. Bianchi, L., et al. *The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding*. in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. IEEE.

DOI: 10.5281/zenodo.22090248