MSPL: Multi-Step Pseudo-Labeling for Open-Vocabulary Object Detection

Hojun Choi[1], Youngsun Lim[2], Jaeyo Shin[1], and Hyunjung Shim[1†]

[1] KAIST AI, South Korea
{hchoi256,jaeyo_shin,kateshim}@kaist.ac.kr
[2] Boston University, USA
youngsun@bu.edu

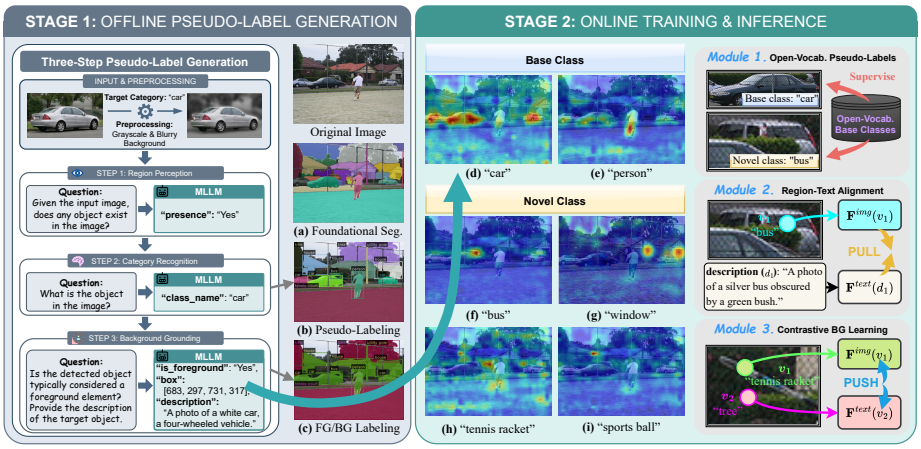


Fig. 1: Offline: Our method generates robust pseudo-labels via a 3-step prompting process: (a) object localization, (b) category recognition, and (c) background grounding. **Online:** The open-vocabulary detector leverages these additional supervisory signals to detect both base (d–e) and novel (f) classes, and even unlabeled objects (g–i).

Abstract. Open-vocabulary object detection (OVD) aims to recognize and localize object categories beyond the training set. Recent approaches leverage vision-language models to generate pseudo-labels using image-text alignment, allowing detectors to generalize to unseen classes without explicit supervision. However, these methods depend heavily on single-step image-text matching, neglecting the intermediate reasoning steps crucial for interpreting semantically complex visual contexts, such as crowding or occlusion. In this paper, we introduce MSPL, a framework that incorporates multi-step visual reasoning into the pseudo-labeling process for OVD. It decomposes complex scene understanding into three interpretable steps—object localization, category recognition, and background grounding—where these intermediate reasoning states serve as rich supervision sources. Extensive experiments on standard OVD evaluation protocols demonstrate that MSPL achieves state-of-the-art performance with superior pseudo-labeling efficiency, outperforming the strong

†Corresponding author

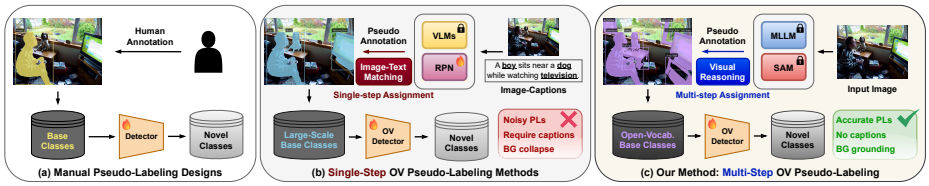


Fig. 2: (a) Manual pseudo-labels for novel classes is costly and does not scale. (b) Recent approaches automate this process via single-step semantic assignment with vision-language models and image captions, struggling in complex scenes. (c) Our caption-free method leverages multi-step reasoning to interpret semantically complex scenes.

baseline by 9.4 AP_{50} for novel classes on OV-COCO and improving box and mask AP_r by 3.2 and 2.2, respectively, on OV-LVIS.

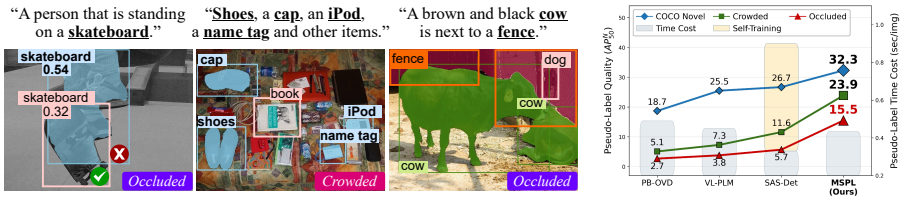
Keywords: Open Vocabulary · Object Detection · Pseudo-Labeling

1 Introduction

Open-vocabulary object detection (OVD) aims to localize both seen (base) and unseen (novel) categories at test time, using only base-class annotations during training. To bridge this supervision gap between seen and unseen categories, recent approaches leverage vision-language models (VLMs) pre-trained on large-scale image-text pairs [36]. These VLMs map textual descriptions to visual representations, allowing OVD methods to recognize novel classes.

Among such efforts, pseudo-labeling has emerged as a state-of-the-art approach for OVD by augmenting the base set with automatically generated annotations that partially cover novel classes [9,33]. In Fig. 2-a, early pseudo-labeling methods for OVD relied on manual annotation of novel classes, which was costly and lacked scalability. In Fig. 2-b, more recent approaches [65,66] leverage VLMs to automate the generation of pseudo-annotations for novel classes based on the similarity between visual features and text embeddings of potential object categories—including some novel classes—typically derived from image captions [3].

Despite their strong performances in general scenes, state-of-the-art OVD approaches still struggle in challenging scenarios involving crowding or occlusion. We identify the root cause as a reliance on single-step image–text matching via CLIP [56]. Because complex scenes require disentangling overlapping visual elements, this direct mapping collapses, leading to three critical failures in pseudo-labeling. **(L1) Noisy pseudo boxes**: Single-step alignment assigns labels based on surrounding context rather than region-specific content. Since VLMs trained with image-level supervision encode co-occurrence statistics rather than object-level semantics [67], a region inherits the label of a contextually dominant neighbor. In Fig. 3-a, the crop of partially occluded feet is incorrectly labeled "skateboard" due to its strong co-occurrence with the skateboard in the scene. **(L2) Caption dependency**: Single-step alignment requires a predefined candidate



(a) Noisy Pseudo Boxes **(b)** Caption Dependency **(c)** Background Collapse **(d)** Comparisons of Pseudo-Labeling Methods

Fig. 3: Challenges in pseudo-labeling for complex scenes. (a) Errors in single-step VLM-based semantic assignment, (b) coarse caption semantics, and (c) unlabeled objects treated as background. (d) MSPL generates robust pseudo-labels on ground-truth novel classes in complex scenes at the lowest per-image computational cost.

set, making it structurally bound to image captions as the sole category source. Any object absent from the caption or under-described remains undiscovered by design. In Fig. 3-b, "book" goes entirely unlabeled simply because it is omitted from the caption, while "iPod" can be misclassified as a visually similar object (*e.g.*, "cell phone") due to its coarse description as a simple class name—a failure inherent to single-step alignment's inability to provide fine-grained discovery beyond the provided candidate set. **(L3) Background collapse**: Recognizing an occluded object requires sequential reasoning—first identifying the occluder, then inferring the hidden instance. Single-step alignment bypasses this decomposition, causing unmatched region to be erroneously absorbed into the background during training [21]. In Fig. 3-c, "the dog occluded by a fence" is never assigned any label, and is instead learned as background. This is a direct consequence of single-step reasoning's inability to decompose the scene.

We argue that these limitations stem from a common bottleneck: single-step alignment compresses the entire scene into a single reasoning unit, leaving no room to disentangle co-occurring objects, discover underspecified categories, or reason through occlusion. To overcome this, we propose reformulating pseudo-labeling as an interpretable multi-step visual prompting process [52]. Our three-step prompting framework directly addresses all three limitations within a single structured reasoning pass: (1) *object localization* grounds each region in object-level visual evidence via SAM [19], bypassing co-occurrence bias (L1); (2) *category recognition* assigns zero-shot labels and their region description via MLLM reasoning without caption-derived candidates, enabling fine-grained discovery of any object in the scene (L2); and (3) *background grounding* explicitly identifies background concepts to disentangle them from occluded foreground instances (L3). In the online phase, a detector is trained under a contrastive objective using the resulting pseudo-labels along with their intermediate reasoning outputs. By shifting complex reasoning offline, this decoupled strategy minimizes training overhead and structurally isolates noisy supervision from the training gradient, ensuring only high-fidelity annotations propagate into online learning.

Importantly, this design is not a naive combination of existing models. In fact, directly integrating SAM and MLLMs without structured reasoning fails on two

fronts: SAM's class-agnostic masks span inconsistent semantic granularities, incurring redundant MLLM inference to hallucinate labels for partial regions; and single-pass MLLM queries not only degrade label accuracy but eliminate the intermediate reasoning states providing essential supervision for online training. Instead, our principled multi-step design addresses the root causes of single-step failures (L1–L3), enabling SAM and MLLM to work synergistically to generate high-fidelity, exclusively object-level annotations where a direct integration would otherwise be computationally prohibitive and prone to collapse.

We conduct extensive experiments on two OVD benchmarks, OV-COCO [25] and OV-LVIS [11]. In Fig. 3-d, under two challenging conditions such as crowding and occlusion, our method demonstrates superior pseudo-label quality compared to previous pseudo-labeling methods [9, 65, 66], with the most competitive runtime. Furthermore, our method sets a new state-of-the-art, improving box AP_{50} for novel classes on OV-COCO by 9.4, and further enhancing both box and mask AP_r on OV-LVIS by 3.2 and 2.2 respectively, compared to prior work [53].

2 Related Work

2.1 Multi-Step Reasoning in Vision-Language Models

Multi-step prompting addresses complex tasks by decomposing them into intermediate, interpretable queries. In language models, such intermediate reasoning improves performance on complex reasoning problems [51]. This idea has been extended to vision and vision-language tasks through structured prompts or predictions, including bounding boxes [41], future states for autonomous driving [43], image infillments [39], and synthesized CLIP embeddings [12]. It has also been used in embodied tasks such as textual planning [32], reward guidance [63], and robotic sub-goal generation [64]. In this work, we apply multi-step prompting to open-vocabulary object detection by decomposing pseudo-label generation into interpretable stages for robust labeling in complex scenes.

2.2 Open-Vocabulary Object Detection

Open-vocabulary object detection (OVD) aims to detect novel objects unseen during training by leveraging vision-language models (VLMs) [36] trained on large-scale image-text pairs. Recent OVD methods [6, 7, 55] use prompt modeling to transfer knowledge through learned prompts, while others align detectors with VLM features via knowledge distillation [10, 53]. Some approaches enhance the text modality with large language models [14, 26], improve novel class prediction using synthetic images [8], or facilitate cross-modal information exchange for prompt-based detection [27]. Another line of work [16, 20] fine-tunes VLMs with learnable parameters for feature extraction, which is often computationally costly. Recently, pseudo-labeling methods [9, 65, 66] have used caption-derived supervision [3] to expand beyond restricted base classes. However, their vocabulary coverage depends on caption quality and content, while requiring additional

annotations. In contrast, our caption-free approach leverages the zero-shot capabilities of powerful MLLMs to expand the vocabulary without costly caption annotations. Moreover, to overcome the limited reasoning ability of direct, single-step CLIP matching in complex scenes [56], we reformulate OVD pseudo-labeling into multiple interpretable steps for robust pseudo-label generation.

3 MSPL: Multi-Step Pseudo-Labeling

We introduce MSPL, an offline-to-online framework for robust pseudo-labeling in OVD, tailored for challenging scenarios such as crowding and occlusion. Unlike conventional single-step VLM alignment methods that rely on coarse caption-driven vocabulary and struggle with complex visual entanglement, MSPL performs structured multi-step reasoning to explicitly disentangle scene components. In the offline phase (Sec. 3.1), this reasoning is decomposed into three interpretable steps—object verification, label assignment, and background grounding—yielding robust pseudo-labels together with semantically rich intermediate representations. By decoupling structured reasoning from online, these representations serve as denoised supervision for online OVD training while substantially reducing online computational overhead. Building upon this supervision, the online phase (Sec. 3.2) optimizes a contrastive objective that promotes generalization beyond base classes to potential unseen objects, grounds image-level captions at the region level, and alleviates background collapse in complex scenes.

3.1 Three-Step Pseudo-Label (PL) Generation

This section presents a three-step visual prompting framework for offline pseudo-labeling in OVD. To circumvent the aforementioned limitations of single-step VLM alignment, our approach leverages the synergy between SAM's foundational segmentation [19] and MLLM zero-shot reasoning. However, such direct integration is inherently prone to instability. Since SAM generates masks across diverse semantic granularities (*e.g.*, whole objects, parts, or sub-parts), partial or non-object regions often receive erroneous semantic labels during MLLM reasoning. Furthermore, even for object-level regions, single-pass MLLM inference struggles in complex scenes where attention diffuses across overlapping objects or contextual distractors. Consequently, like single-step VLM alignment, this naive coupling inherits both proposal-level ambiguity and reasoning-level entanglement, often yielding inconsistent or hallucinated pseudo-labels.

To address these dual challenges, we impose structural constraints at both the region and reasoning levels. First, we restrict SAM outputs to whole-instance masks via hierarchical grouping [35], ensuring that subsequent reasoning operates on semantically coherent, object-level entities. Second, we regulate the interaction between each target region and its surrounding context by preserving the global scene structure while selectively attenuating non-target areas through controlled desaturation and blurring. As demonstrated by the visualizations of

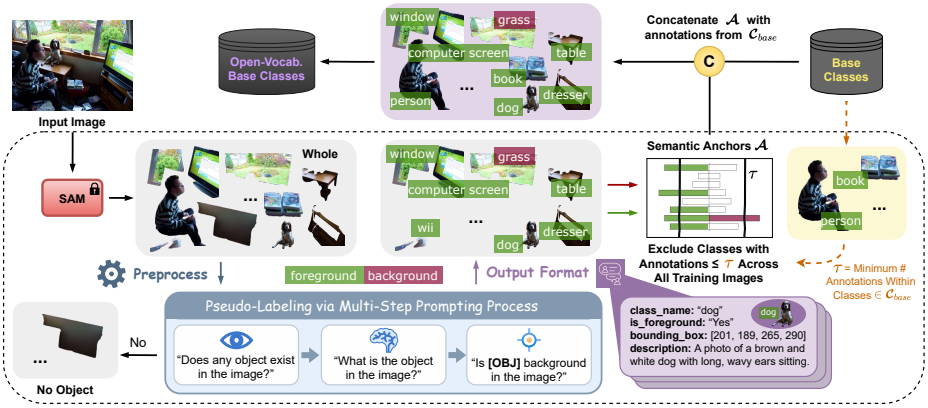


Fig. 4: Our offline multi-step pseudo-labeling. We leverage foundational segmentation and MLLM-based zero-shot reasoning to implement a three-step framework (localization, recognition, grounding) with explicit intermediate reasoning goals. The resulting pseudo-labels are semantically refined and consolidated into the base dataset.

various modulations in the supplementary material, this visual context modulation stabilizes the visual evidence available for MLLM reasoning, mitigating distractions while retaining essential environmental cues. Although these measures enhance robustness relative to single-step VLM alignment, single-pass MLLM inference may still fall short in scenes with complex, overlapping, or heavily occluded objects. To address these residual ambiguities, we introduce a three-step reasoning process that sequentially verifies region validity, performs fine-grained category discovery, and disambiguates background elements.

Step 1: Pseudo-Box Verification Given the object-level region proposals, the first PL step verifies whether each region contains a valid object. Because SAM generates class-agnostic proposals, some regions may correspond to partial structures or non-object areas. Confirming object existence prior to semantic assignment restricts subsequent reasoning to visually grounded candidates, ensuring that later stages operate on reliable object regions. To this end, a robust MLLM [1] evaluates each region using the query "`Does any object exist in the image?`" and returns "`Yes`", "`No`", or "`Unsure`", forming a ternary decision filter. Regions confirmed as "`Yes`" proceed to the next reasoning step, whereas those labeled "`No`" or "`Unsure`" are discarded or recorded for explainability. As illustrated in Fig. 4, regions lacking discernible objects (*e.g.*, plain dark areas) are removed. By narrowing the candidate set to visually valid objects, this step establishes a stable foundation for subsequent pseudo-label assignment.

Step 2: Pseudo-Label Assignment. Building upon the verified object regions from the previous step, we depart from conventional OVD paradigms that rely on single-step CLIP alignment against predefined class lexicons. Such approaches

typically depend on vocabularies distilled from coarse image captions [3], which may omit or underspecify object details. To eliminate this vocabulary dependency, the second PL stage replaces rigid vocabulary-constrained alignment with zero-shot MLLM reasoning. By querying the model with category recognition, we explicitly elicit a category name along with a textual description for each validated region. In Fig. 4, this caption-agnostic formulation produces flexible pseudo-labels (*e.g.*, "dog") while simultaneously generating region-level descriptions (*e.g.*, "A photo of a brown and white dog with long, wavy ears sitting."). Importantly, any potential semantic ambiguities among the predicted labels are naturally resolved within the CLIP embedding space, where semantically related concepts occupy closely aligned representations, as detailed in Sec. 3.2. Such ambiguities include synonym distinctions, such as "table" vs. "dining table", and superclass variations, such as "bird" vs. "parrot". By leveraging the strong zero-shot recognition capabilities of MLLMs [44], this step enables fine-grained object discovery beyond the static caption vocabulary.

Step 3: Background Grounding. Despite the structural reasoning in the previous steps, residual ambiguities may persist, particularly when severely occluded objects yield "`Unsure`" responses in the first PL step. Such objects remain unlabeled and are consequently assimilated into the learnable background class embedding during training [2,21]. In addition, background regions (*e.g.*, "tree" or "sky") can be inadvertently labeled, generating noisy pseudo-labels that deviate from the objective of detecting foreground objects. This mislabeling biases the model toward background regions, which can interfere with learning discriminative features for semantically similar object classes. To address these issues, the third PL stage explicitly separates each category prediction into foreground and background decisions. Specifically, the model performs a binary verification to determine whether it corresponds to a true background concept ("`Yes`" or "`No`"). For example, in Fig. 4, the model identifies "grass" as background while retaining "drawer" as foreground. While denoising pseudo-labels, these reasoning decisions function as intermediate supervision that facilitates feature disentanglement during training, thereby enabling the recovery of object features that might otherwise be absorbed into background embeddings.

Pseudo-label Refinement. Although the proposed three-step PL reasoning improves overall pseudo-label quality, spurious or hallucinated predictions may still occur due to inherent MLLM limitations. To enhance reliability, we introduce an MLLM-agnostic refinement strategy that filters pseudo-labels based on per-class prediction frequency. As demonstrated in Tab. 7, consistently predicted categories—such as clearly recognizable objects—accumulate high-frequency assignments, whereas ambiguous regions yield scattered predictions across classes, resulting in low frequencies. This observation motivates the use of per-class prediction frequency as a proxy for pseudo-label reliability. As illustrated in Fig. 4, pseudo-labels falling below a minimum threshold, derived from the reliable base-class distribution, are discarded. This ensures that the retained labels, termed *se-*

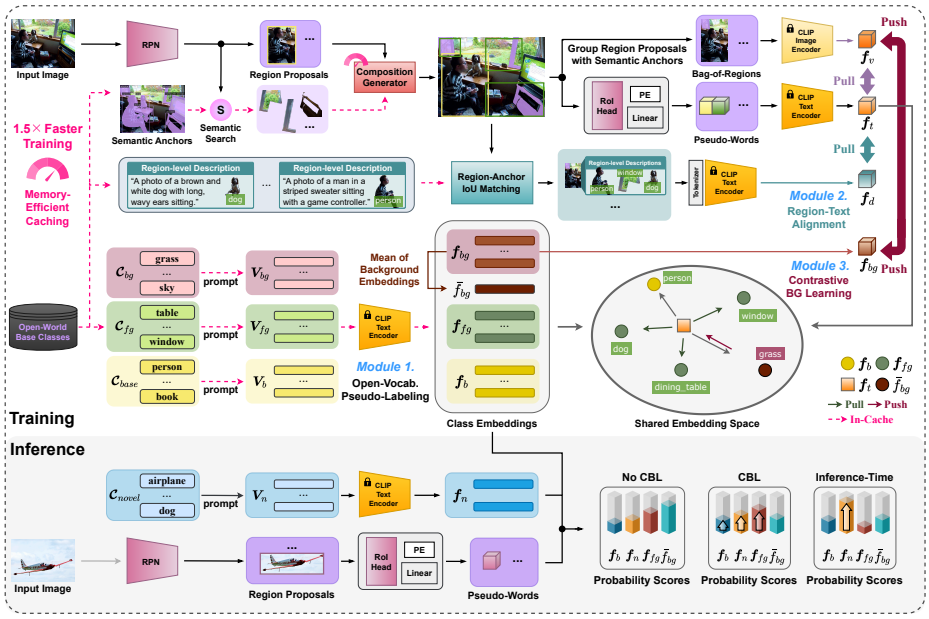


Fig. 5: Our online MSPL framework. Module 1 performs pseudo-label-driven OVD; (2) Module 2 leverages description annotations; and (3) Module 3 utilizes background annotations to promote feature disentanglement. Memory caching accelerates composition augmentation, with pseudo-labels employed exclusively during training.

mantic anchors, exhibit support comparable to trusted categories. Finally, these anchors are integrated with the base classes to construct an open-vocabulary base set, providing reliable supervision for subsequent OVD training.

3.2 Contrastive Learning with Multi-Step Supervision

This section details the online framework of MSPL, leveraging pre-computed offline supervision via a contrastive objective. As illustrated in Fig. 5, our architecture builds upon BARON [53], an effective OVD approach that captures rich contextual signals through the sampling of co-occurring objects. However, despite its strong performance, BARON suffers from computationally expensive online sampling, which creates a significant training bottleneck. To overcome this, we introduce an efficient composition generator that replaces online sampling with a *bag-of-regions* grouping of cached semantic anchors, thereby accelerating training by a factor of 1.5. Within each bag, sampled region features are mapped into a joint word embedding space via a linear projection layer to form *pseudo-word* embeddings. The text encoder $\mathcal{T}$ then processes these pseudo-words to generate a bag-of-regions text embedding $f_t^i = \mathcal{T}(w_0^i + p_0^i, w_1^i + p_1^i, \ldots, w_{N^i-1}^i + p_{N^i-1}^i)$, where N^i is the number of regions in the i-th bag, and p_j^i is the learnable positional embedding for the j-th region. Finally, this text embedding is aligned

with the corresponding visual embedding $f_v^i = \mathcal{V}(b_0^i, b_1^i, \ldots, b_{N^i-1}^i)$ derived from the image encoder $\mathcal{V}$, with b_j^i representing the visual feature of the j-th region.

For standard OV classification, each predicted region feature is assigned to the category yielding the highest CLIP cosine similarity. The candidate category set comprises both the original base classes $\boldsymbol{\mathcal{C}}_{base}$ and the foreground pseudo-labels $\boldsymbol{\mathcal{C}}_{fg}$, with their text embeddings pre-computed via the CLIP text encoder using prompt templates [10,68]. Given the continuous nature of the CLIP embedding space, aligning a region with a specific pseudo-label during training projects its feature into a shared semantic neighborhood. This naturally resolves synonym or superclass ambiguities (*e.g.*, the novel class "couch" aligns closely with "sofa"). Leveraging this property, our method employs pseudo-labels exclusively during training and discards them at inference to prevent misclassification.

Region-Text Alignment (RTA). While basic visual features align with simple class names, such labels lack the descriptive richness required for fine-grained distinctions in complex scenes. For instance, detailed text can easily differentiate visually similar objects (*e.g.*, a small red "apple" and a red "ball"). To leverage this linguistic granularity, we propose RTA, which aligns pseudo-words with their corresponding region descriptions. Concretely, we define a matched set $\mathcal{S}$ of proposals with an IoU > 0.7 against pseudo-boxes. For each proposal in $\mathcal{S}$, its pseudo-word embedding f_t^k and description embedding f_d^k are aligned via the following objective [40], with cosine similarity $\langle \cdot, \cdot \rangle$ and temperature τ:

$$\mathcal{L}_{\text{RTA}} = -\frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} \log \frac{\exp(\tau \cdot \langle f_t^k, f_d^k \rangle)}{\sum_{l \in \mathcal{S}} \exp(\tau \cdot \langle f_t^k, f_{d'}^l \rangle)}. \tag{1}$$

Contrastive Background Learning (CBL). To mitigate background collapse, we propose a CBL strategy that explicitly disentangles all objects, including unlabeled ones, from background representations (*e.g.*, "sky") in the feature space. In Fig. 5, the B background concepts $\boldsymbol{\mathcal{C}}_{bg}$ identified during the third PL phase are encoded using the CLIP text encoder. These embeddings are treated as negative samples and averaged to initialize a learnable background prior $\bar{f}_{bg}$, which serves as a convergence target for true background features [40]:

$$\mathcal{L}_{\text{CBL}} = \frac{1}{2} \sum_{k=0}^{G-1} \left(\log p_{t,v}^k + \log p_{v,t}^k \right), \tag{2}$$

where G is the number of bags, and the $p_{t,v}^k$ and $p_{v,t}^k$ can be calculated as:

$$p_{t,v}^k = \frac{\exp(\tau' \cdot \langle f_t^k, f_v^k \rangle)}{\sum_{l=0}^{G-1} \exp(\tau' \cdot \langle f_t^k, f_v^l \rangle) + \sum_{j=0}^{B-1} \exp(\tau'' \cdot \langle f_t^k, f_{bg}^j \rangle)}, \tag{3}$$

$$p_{v,t}^k = \frac{\exp(\tau' \cdot \langle f_v^k, f_t^k \rangle)}{\sum_{l=0}^{G-1} \exp(\tau' \cdot \langle f_v^k, f_t^l \rangle) + \sum_{j=0}^{B-1} \exp(\tau'' \cdot \langle f_v^k, f_{bg}^j \rangle)}, \tag{4}$$

where τ' and τ'' are scaling factors. The loss promotes alignment of matched pairs and separation of background features, aiding foreground feature recovery.

Table 1: Results on OV-COCO [25]. Methods are grouped by additional supervision (*e.g.*, weak or pseudo labels) beyond $\mathcal{C}_B$ instance labels. $\mathcal{C}_N$ denotes novel classes.

Methods	Backbone	AP_{50}^{N}	AP_{50}^{B}
Instance labels in $\mathcal{C}_B$ (CLIP Supervision)			
ViLD-ens [10]	RN50 (24M)	27.6	51.3
BARON [53]	RN50 (24M)	34.0	60.4
CORA [55]	RN50 (24M)	35.1	35.4
BIND [62]	ViT-B/16 (86M)	36.3	50.2
CLIP-Self [54]	ViT-B/16 (86M)	37.6	-
LBP [21]	RN50 (24M)	37.8	58.7
CCKT-Det [60]	RN50 (24M)	38.0	35.0
CAKE [29]	RN50 (24M)	38.2	-
OV-DQUO [45]	RN50 (24M)	39.2	-
DeCo-DETR [48]	RN50 (24M)	41.3	-
BIND [62]	ViT-L/16 (307M)	41.5	54.8
CCKT-Det [60]	SwinB (88M)	41.9	40.9
CORA+ [55]	RN50×4 (87M)	43.4	43.8
CLIP-Self [54]	ViT-L/14 (307M)	44.3	-
OV-DQUO [45]	RN50×4 (87M)	45.6	-

Methods	Supervision	Backbone	AP_{50}^{N}	AP_{50}^{B}
Extra caption datasets, Weak/Pseudo Labels in $\mathcal{C}_B \cup \mathcal{C}_N$				
Detic [68]	IN21K & CC3M	RN50 (24M)	27.8	42.0
OV-DETR [57]	Pseudo annotations	RN50 (24M)	29.4	52.7
CoDet [28]	CC3M & COCO Caption	RN50 (24M)	30.6	46.4
PB-OVD [9]	COCO Caption	RN50 (24M)	30.8	46.4
VL-PLM [66]	Pseudo annotations	RN50 (24M)	34.4	60.2
RegionCLIP [67]	CC3M	RN50 (24M)	35.2	57.6
OC-OVD [37]	COCO Caption	RN50 (24M)	36.6	49.4
SAS-Det [65]	COCO Caption	RN50 (24M)	37.4	58.5
DITO [17]	LAION-2B	ViT-B/16(86M)	36.6	48.8
LP-OVOD [33]	Pseudo annotations	RN50 (24M)	40.5	60.5
MSPL (Ours)	Pseudo annotations	RN50 (24M)	**43.4**	58.9
CFM-ViT [15]	LAION-2B	ViT-L/16 (307M)	34.3	46.4
RegionCLIP [67]	CC3M	RN50×4 (87M)	39.3	61.6
DITO [17]	DataComp-1B	ViT-L/16(307M)	40.2	54.6
CORA+ [55]	COCO Caption	RN50×4 (87M)	43.1	56.2
MSPL (Ours)	Pseudo annotations	RN50×4 (87M)	**47.8**	60.9

4 Experiments

Datasets and evaluation metrics. We evaluate MSPL on two widely used OVD benchmarks: OV-COCO [25] and OV-LVIS [11]. Note that we generate pseudo annotations exclusively from base-category training images. For OV-COCO, we adopt the category split from OVR-CNN [58], dividing categories into 48 base and 17 novel classes. For OV-LVIS, following ViLD [10], we treat the 337 rare categories as novel and the common/frequent categories as base. Evaluation follows the OVR-CNN protocol: we report box AP at IoU 0.5 (AP_{50}^{N}) for novel categories on OV-COCO, and mask mAP (AP_r) for rare categories on OV-LVIS.

Implementation details. MSPL is implemented on Faster R-CNN [38] with a ResNet50-FPN backbone. Following recent works [6, 53], the backbone is initialized with SOCO [50] pre-trained weights and fine-tuned using synchronized batch normalization [61] for a fair comparison. We employ 1× and 2× training schedules for OV-COCO and OV-LVIS, respectively. For mask generation in the offline phase, we adopt the SAM-B (ViT-B) model with default settings. As for the VLM, we utilize ViT-B-16 [5] with hand-crafted prompts from ViLD [10] by default, adopting learned prompts [6] only for comparisons on OV-LVIS. The temperature scaling factors τ, τ', and τ'' are fixed at 0.2, 1.0, and 0.1. Following standard benchmarks [25,34], we define *Crowded* images as those with over eight objects and *Occluded* instances as having over 50% ground-truth box overlap. All other hyperparameters follow the settings of our baseline [53].

4.1 Main Results

Comparison with state-of-the-art methods. We compare MSPL against state-of-the-art OVD methods on the OV-COCO and OV-LVIS benchmarks.

Table 2: Statistics of pseudo-labels.

Metric	BLIP2	InstructBLIP	Qwen2
Total			
# Classes	6.0K	3.1K	3.9K
# Annotations	395K	567K	637K
# "Unsure"	1.5M	1.1M	563K
OV-COCO (17 novel classes only)			
# Classes	31	30	65
# Annotations	197K	294K	202K
Hard Hit (%)	41.2	**47.1**	**47.1**
Soft Hit (%)	85.0	81.8	**86.0**
OV-LVIS (337 rare classes only)			
# Classes	5.3K	2.5K	3.3K
# Annotations	137K	315K	232K
Hard Hit (%)	31.1	27.6	**34.1**
Soft Hit (%)	77.9	**85.7**	85.3

Table 3: Comparison on OV-LVIS.

Method	Detection				Segmentation			
	AP_r	AP_c	AP_f	AP	AP_r	AP_c	AP_f	AP
ViLD [10]	16.7	26.5	34.2	27.8	16.6	24.6	30.3	25.5
RegionCLIP [67]	17.1	27.4	34.0	28.2	-	-	-	-
CCKT-Det++ [60]	18.2	-	-	27.1	-	-	-	-
OV-DETR [57]	-	-	-	-	17.4	25.0	32.5	26.6
VLDet [24]	-	-	-	-	21.7	29.8	34.3	30.1
Detic [68]	-	-	-	-	17.8	26.3	31.6	26.8
MIC [49]	22.9	34.0	39.9	34.4	20.8	**30.5**	**35.4**	**30.7**
DetPro [6]	20.8	27.8	32.4	28.4	19.8	25.6	28.9	25.9
OC-OVD [37]	21.1	25.0	29.1	25.9	-	-	-	-
OADP [47]	21.9	28.4	32.0	28.7	21.7	26.3	29.0	26.6
DK-DETR [23]	22.2	32.0	**40.2**	33.5	20.5	28.9	**35.4**	30.0
BARON [53]	23.2	29.3	32.5	29.5	22.6	27.6	29.8	27.6
CoDet [28]	23.4	30.0	34.6	30.7	-	-	-	-
LBP [21]	24.1	29.5	32.8	29.9	23.7	27.7	30.1	28.0
CAKE [29]	25.0	**34.8**	38.4	**34.9**	23.9	29.1	33.6	28.7
BIRDet [59]	26.0	21.7	29.5	25.5	-	-	-	-
RALF [18]	21.9	26.2	29.1	26.6	-	-	-	-
MSPL (Ours)	**26.4**	**34.8**	38.2	**34.9**	**24.8**	28.5	33.0	28.6

In Tab. 1, MSPL establishes a new state-of-the-art on OV-COCO among recent methods leveraging auxiliary datasets for pseudo-annotations [28,33,57,67]. Specifically, it achieves 43.4 and 47.8 AP_{50}^{N} for ResNet-50 and ResNet-50×4 backbones, respectively. Notably, our approach consistently outperforms distillation-based methods relying strictly on CLIP knowledge and base-class labels, surpassing the recent leading DeCo-DETR [48] by 2.1 AP_{50}^{N}. On the challenging OV-LVIS benchmark in Tab. 3, MSPL sets a new record for rare categories with a detection AP_r of 26.4 and a segmentation AP_r of 24.8. For detection, MSPL outperforms BIRDet [59] by 0.4 AP_r and surpasses CAKE [29] by 1.4 AP_r. In instance segmentation, MSPL exceeds strong baselines [21, 29, 53] by significant margins. This consistent superiority across benchmarks confirms that our proposed strategies scale exceptionally well to large-vocabulary datasets.

Statistics. Tab. 2 compares pseudo-label statistics across MLLM variants. Notably, Qwen2 [1] produces the densest and most confident annotations—637K across 3.9K categories—while yielding minimal "Unsure" responses. To evaluate coverage of unseen domains, we measure both *Hard Hit* (exact string match, *e.g.*, "cup" to "cup") and *Soft Hit* (CLIP cosine similarity > 0.8 following [46], *e.g.*, "vehicle" to "bus"). Although Hard Hit provides direct supervision for OVD, its maximum coverage is inherently limited by synonym and superclass ambiguities. Since open-vocabulary detectors operate within a continuous CLIP embedding space, exact textual matches are not strictly necessary. Pseudo-labels comprising synonyms or semantically related terms remain closely aligned with the corresponding novel classes, thereby providing effective supervision. Under this broader semantic criterion, Qwen2 achieves strong Soft Hit rates of 86.0% and 85.3% on OV-COCO and OV-LVIS, respectively. These results demonstrate

Table 4: Component ablation.

Multi-Step Pseudo-Label		RTA	CBL	AP_{50}^{N}
1-Step	3-Step			
-	-	-	-	34.0
✓	-	-	-	37.6
-	✓	-	-	41.6
-	✓	✓	-	42.5
-	✓	✓	✓	**43.4**

Table 5: Per-image pseudo-labeling time and novel-class quality (AP_{50}^{N}) on the OV-COCO validation set using a single A6000 GPU.

Method	OV-COCO	*Crowded*	*Occluded*	Time (s)
PB-OVD [9]	18.7	5.1	2.7	0.49
VL-PLM [66]	25.5	7.3	3.8	0.45
SAS-Det [65]	26.7	11.6	5.7	$\gg 1.0$
MSPL (Ours)	**32.3**	**23.9**	**15.5**	**0.43**

the large scale, vocabulary diversity, and broad semantic coverage of our pseudo-labels, supporting their suitability as supervision for OVD.

Pseudo-label analysis. Tab. 5 evaluates the trade-off between pseudo-labeling efficiency and quality under varying scene complexities, specifically in *Crowded* and *Occluded* settings. Prior methods [9, 65, 66] rely on single-step VLM alignment and are susceptible to visual interference, resulting in suboptimal pseudo-labels. On a single A6000 GPU, SAS-Det [65] achieves a per-image generation time of 0.13s after training; however, its online self-training increases the total cost to over one second per image. In contrast, our method follows the offline paradigm [9, 66], performing intensive VLM or MLLM inference only once during pseudo-label generation while avoiding iterative and expensive online self-training. This design improves overall throughput for large-scale labeling. By combining a segmentation model [19] with asynchronous MLLM inference via batch parallelism, our method achieves a favorable accuracy–efficiency balance, averaging 0.43s per image on a single GPU and processing the COCO training set in approximately 5 hours using multi-threaded execution across 8 GPUs.

4.2 Ablation Analysis

Impact of individual modules. Tab. 4 evaluates the incremental contribution of each component in the MSPL framework. The one-step variant directly predicts pseudo-labels without intermediate reasoning supervision, yet it still improves over the baseline. This gain arises from our structured integration of SAM and MLLM—specifically, object-level SAM masks and visual context modulation—which alleviates the limitations of naive model coupling. Building on this foundation, transitioning to three-step PL reasoning yields larger gains, indicating that sequential decomposition with intermediate supervision is more robust for parsing complex scenes than single-pass prompting. RTA further refines label quality by incorporating fine-grained linguistic attributes, which are essential for disambiguating semantically similar categories. Finally, CBL effectively mitigates background collapse, resulting in additional performance gains.

Table 6: Zero-shot transfer on MS-COCO and Objects365 [42] with models trained on OV-LVIS without fine-tuning.

Methods	MS-COCO [25]						Objects365 [42]					
	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	*AP*	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Supervised	46.5	67.6	50.9	27.1	67.6	77.7	25.6	38.6	28.0	16.0	28.1	36.7
ViLD [10]	34.1	52.3	36.5	21.6	38.9	46.1	11.5	17.8	12.3	4.2	11.1	17.8
DetPro [6]	34.9	53.8	37.4	22.5	39.6	46.3	12.1	18.8	12.9	4.5	11.5	18.6
BARON [53]	36.2	55.7	39.1	24.8	40.2	47.3	13.6	21.0	14.5	5.0	13.1	20.7
LBP [21]	36.8	56.5	39.8	25.6	40.6	48.1	14.3	21.8	15.1	5.5	13.7	21.6
MSPL (Ours)	**37.5**	**57.4**	**40.8**	**26.4**	**41.4**	**49.1**	**15.1**	**22.7**	**15.9**	**6.1**	**14.4**	**22.5**

Table 7: Thresholds of semantic anchors.

Threshold	AP_{50}^N
`RANDOM`	41.7
`ALL`	42.1
`MEDIUM`	42.5
`AVERAGE`	42.6
`MIN`	**43.4**

Table 8: Impact of generators.

Proposal generator	AP_{50}^N
Mask R-CNN [13]	40.9
MAVL [31]	42.2
SAM [19]	**43.4**

Table 9: MLLM variants.

Model	Size	AP_{50}^N
BLIP2 [22]	2.7B	39.6
InstructBLIP [4]	7B	42.6
Qwen2 [1]	7B	**43.4**

Table 10: Visual context.

Strategy	AP_{50}^N
Bounding box	33.2
Black mask	38.7
Blur & gray	**43.4**

Collectively, these results highlight the complementary roles of structured reasoning and semantically rich supervision in achieving high-fidelity pseudo-labeling.

Transfer to the other datasets. Following standard protocols [10,53], we evaluate the zero-shot transfer capability of MSPL by evaluating a model trained on OV-LVIS across two benchmarks: MS-COCO [25] and Objects365 [42]. In Tab. 6, MSPL achieves superior performance across all datasets, outperforming existing state-of-the-art methods. This robust generalization across diverse data distributions underscores the broad applicability of our approach in real-world scenarios.

Impact of semantic anchors. Tab. 7 evaluates semantic anchor policies based on base-class statistics. The `RANDOM` (70% sampling) and `ALL` policies yield the lowest performance, as they disregard per-class annotation frequency derived from the base-class statistics. The improved results of `AVERAGE` and `MEDIAN` indicate that lower annotation counts are associated with less reliable anchors. Accordingly, `MIN` achieves the best performance by using the minimum base-class frequency as a strict cutoff to filter noisy or hallucinated anchors.

Impact of proposal generators. In Tab. 8, we evaluate different class-agnostic proposal generators. We observe that SAM [19] provides higher-quality pseudo-box candidates by localizing arbitrary objects beyond closed-set vocabularies [13, 31]. The consistent gains across generators indicate that our multi-step reasoning is robust and benefits from stronger open-world localization.

Impact of MLLM variants. Tab. 9 analyzes the impact of MLLM scale on pseudo-label quality. While performance remains comparable among models

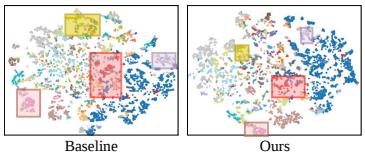


Fig. 6: t-SNE results of features.

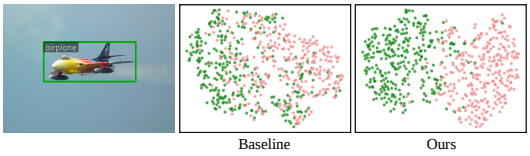


Fig. 7: t-SNE results of background separation.

within the same parameter tier (*e.g.*, 7B variants), we observe an approximately linear improvement as the scale increases from 2.7B to 7B. Notably, even the compact 2.7B model [22] yields substantial gains over the baseline, demonstrating effectiveness in resource-constrained settings. These results indicate that our framework is robust to architectural differences at a fixed scale, while reliably translating increased model capacity into improved label fidelity.

Impact of context modulation. We examine preprocessing strategies to reduce MLLM sensitivity to visual context. Raw proposals often degrade reasoning accuracy, while masking non-target regions improves focus but can cause hallucination from context loss and silhouette artifacts. In contrast, our blurred-and-grayscale strategy suppresses background interference while preserving essential context, achieving the best performance. This suggests that balancing target emphasis and contextual preservation is crucial for high-fidelity pseudo-labeling.

Further analysis. We use t-SNE [30] to visualize feature distributions and analyze background interpretation. As shown in Fig. 6, MSPL learns more compact and discriminative representations for novel categories than the baseline [53]. We further examine background–foreground disentanglement for the "airplane" class in Fig. 7, where MSPL forms a distinct cluster well separated from learnable background embeddings, while the baseline shows substantial overlap.

5 Conclusion

In this paper, we introduce MSPL, a multi-step pseudo-labeling framework for open-vocabulary object detection (OVD) that reformulates complex scene understanding through multi-step visual reasoning. By decomposing labeling into three interpretable stages—object localization, category recognition, and background grounding—MSPL addresses the limitations of single-step vision–language alignment, particularly in crowded or occluded scenes. The resulting intermediate representations provide semantically enriched supervision for contrastive learning: region–text alignment incorporates fine-grained linguistic attributes, while contrastive background learning alleviates background collapse. Extensive evaluations on OVD benchmarks demonstrate that MSPL achieves state-of-the-art performance, scales effectively with multimodal large language model capacity, and maintains strong pseudo-label quality and efficiency. We hope our work inspires further exploration of multi-step visual reasoning in perception tasks.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)). IITP grant funded by the Korea government(MSIT) and KEIT grant funded by the Korea government(MOTIE) (No. 2022-0-00680). IITP grant funded by the Korea government(MSIT) and KEIT grant funded by the Korea government(MOTIE) (No. 2022-0-01045). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, National AI Research Lab Project). This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (No. RS-2025-00520207). This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) and Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No.RS-2025-02217259, Development of self-evolving AI bias detection-correction-explain platform based on international multidisciplinary governance, Contribution Rate : 11.1%). This research was supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT). This research was supported by the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (RQT-25-120217). This research was supported by next-generation CultureTech technology development (R&D) project for cultural space AX transformation through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026. (Project Name : Development of experiential library technology that allows users to speak out and experience with their whole body. , Project Number : RS-2026-25508598, Contribution Rate : 11.1%)

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report. CoRR (2023)
2. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I (2018)
3. Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. CoRR (2015)

4. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
6. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022 (2022)
7. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images (2022)
8. Feng, C., Zhong, Y., Jie, Z., Xie, W., Ma, L.: Instagen: Enhancing object detection by training on synthetic dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024 (2024)
9. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X (2022)
10. Gu, X., Lin, T., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022 (2022)
11. Gupta, A., Dollár, P., Girshick, R.B.: LVIS: A dataset for large vocabulary instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (2019)
12. Harvey, W., Wood, F.: Visual chain-of-thought diffusion models. CoRR (2023)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017 (2017)
14. Jin, S., Jiang, X., Huang, J., Lu, L., Lu, S.: Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024 (2024)
15. Kim, D., Angelova, A., Kuo, W.: Contrastive feature masking open-vocabulary vision transformer. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023 (2023)
16. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. IEEE (2023)
17. Kim, D., Angelova, A., Kuo, W.: Region-centric image-language pretraining for open-vocabulary detection. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIII (2024)
18. Kim, J., Cho, E., Kim, S., Kim, H.J.: Retrieval-augmented open-vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024 (2024)

19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023 (2023)
20. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A.J., Angelova, A.: F-VLM: open-vocabulary object detection upon frozen vision and language models. CoRR (2022)
21. Li, J., Zhang, J., Li, J., Li, G., Liu, S., Lin, L., Li, G.: Learning background prompts to discover implicit knowledge for open vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024 (2024)
22. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (2023)
23. 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: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE (2023)
24. 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: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 (2023)
25. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (2014)
26. Liu, M., Hayes, T.L., Ricci, E., Csurka, G., Volpi, R.: Shine: Semantic hierarchy nexus for open-vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024 (2024)
27. 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: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII (2024)
28. Ma, C., Jiang, Y., Wen, X., Yuan, Z., Qi, X.: Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
29. Ma, S., Qian, D., Ye, K., Zhang, S.: CAKE: category aware knowledge extraction for open-vocabulary object detection. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. AAAI Press (2025)
30. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9** (2008)
31. Maaz, M., Rasheed, H.A., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.: Class-agnostic object detection with multi-modal transformer. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X (2022)

32. Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: Vision-language pre-training via embodied chain of thought. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
33. Pham, C., Vu, T., Nguyen, K.: LP-OVOD: open-vocabulary object detection by linear probing. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024 (2024)
34. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S.J., Yuille, A.L., Torr, P.H.S., Bai, S.: Occluded video instance segmentation: A benchmark. Int. J. Comput. Vis. **130**, 2022–2039 (2022)
35. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024 (2024)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (2021)
37. Rasheed, H.A., Maaz, M., Khattak, M.U., Khan, S.H., Khan, F.S.: Bridging the gap between object and image-level representations for open-vocabulary detection. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
38. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada (2015)
39. Rose, D., Himakunthala, V., Ouyang, A., He, R., Mei, A., Lu, Y., Saxon, M., Sonar, C., Mirza, D., Wang, W.Y.: Visual chain of thought: Bridging logical gaps with multimodal infillings. CoRR (2023)
40. Rusak, E., Reizinger, P., Juhos, A., Bringmann, O., Zimmermann, R.S., Brendel, W.: Infonce: Identifying the gap between theory and practice. CoRR (2024)
41. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024)
42. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019 (2019)
43. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. In: Conference on Robot Learning, 6-9 November 2024, Munich, Germany (2024)
44. Wang, G., Ge, Y., Ding, X., Kankanhalli, M.S., Shan, Y.: What makes for good visual tokenizers for large language models? CoRR (2023)

45. Wang, J., Chen, B., Kang, B., Li, Y., Xian, W., Chen, Y., Xu, Y.: OV-DQUO: open-vocabulary DETR with denoising text query training and open-world unknown objects supervision. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA (2025)
46. Wang, K., Cheng, L., Chen, W., Zhang, P., Lin, L., Zhou, F., Li, G.: Marvelovd: Marrying object recognition and vision-language models for robust open-vocabulary object detection. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII (2024)
47. Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., Liu, S.: Object-aware distillation pyramid for open-vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 11186–11196 (2023)
48. Wang, S., Li, Y., Hu, B., Yao, Z., Li, Z., Li, L., HaiboZhan, Liu, W., Dong, J., Qian, R., Wu, G., Zhang, Shen, J., Koniusz, P., Sun, Q.: Deco-DETR: Decoupled cognition DETR for efficient open-vocabulary object detection. In: The Fourteenth International Conference on Learning Representations (2026)
49. Wang, Z., Li, A., Zhou, F., Li, Z., Dou, Q.: Open-vocabulary object detection with meta prompt representation and instance contrastive optimization. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. p. 93 (2023)
50. Wei, F., Gao, Y., Wu, Z., Hu, H., Lin, S.: Aligning pretraining for detection via object-level contrastive learning. In: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual (2021)
51. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
52. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. CoRR (2023)
53. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023 (2023)
54. 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: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024 (2024)
55. Wu, X., Zhu, F., Zhao, R., Li, H.: CORA: adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023 (2023)
56. Yüksekgönül, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 (2023)

57. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary DETR with conditional matching. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX (2022)
58. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.: Open-vocabulary object detection using captions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021 (2021)
59. Zeng, R., Zhang, L., Yang, X., Liu, Z.: Boosting open-vocabulary object detection by handling background samples. CoRR (2024)
60. Zhang, C., Zhu, C., Dong, P., Chen, L., Zhang, D.: Cyclic contrastive knowledge transfer for open-vocabulary object detection. CoRR (2025)
61. Zhang, H., Dana, K.J., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (2018)
62. Zhang, H., Zhao, Q., Zheng, L., Zeng, H., Ge, Z., Li, T., Xu, S.: Exploring region-word alignment in built-in detector for open-vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024 (2024)
63. Zhang, K., Yin, Z.H., Ye, W., Gao, Y.: Learning manipulation skills through robot chain-of-thought with sparse failure guidance (2025)
64. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., Handa, A., Lin, T., Wetzstein, G., Liu, M., Xiang, D.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025 (2025)
65. Zhao, S., Schulter, S., Zhao, L., Zhang, Z., Kumar, B.G.V., Suh, Y., Chandraker, M., Metaxas, D.N.: Taming self-training for open-vocabulary object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024 (2024)
66. Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Kumar, B.G.V., Stathopoulos, A., Chandraker, M., Metaxas, D.N.: Exploiting unlabeled data with vision and language models for object detection. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX (2022)
67. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022 (2022)
68. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX (2022)