CoT-PL: Chain-of-Thought Pseudo-Labeling for Open-Vocabulary Object Detection

Hojun Choi[1] Youngsun Lim[2*] Jaeyo Shin[1] Hyunjung Shim[1†]
[1]Kim Jaechul Graduate School of AI, KAIST [2]Boston University
{hchoi256, jaeyo_shin, kateshim}@kaist.ac.kr, youngsun@bu.edu

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 CoT-PL, a framework that incorporates visual chain-of-thought 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 CoT-PL achieves state-of-the-art performance with superior pseudo-labeling efficiency, outperforming the strong 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.

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 [35]. 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, 32]. In Fig. 1-a, early pseudo-labeling methods for OVD relied on manual annotation of novel classes, which was costly and lacked scalability. In Fig. 1-b, more recent approaches [63, 64] 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—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 [54]. 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 [65], a region inherits the label of a contextually dominant neighbor. In Fig. 2-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 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. 2-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 [22]. In Fig. 2-c, "the dog occluded by a fence" is never assigned any label, and is in-

*Work was done while at KAIST AI.
†Corresponding author.

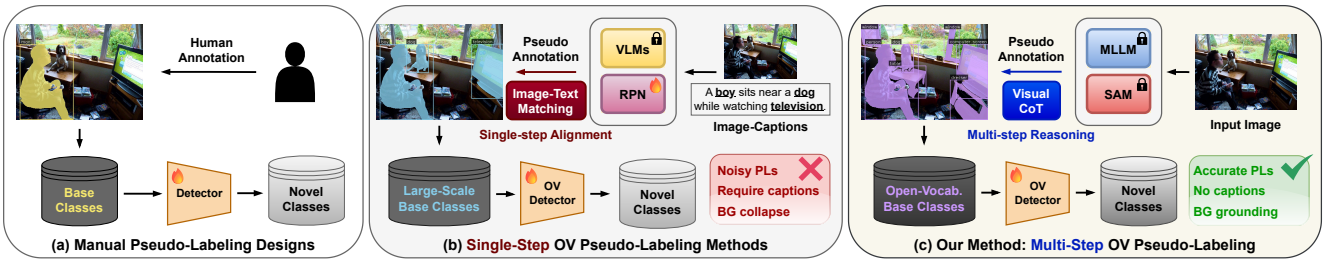

Figure 1. (a) Manual pseudo-labels for novel classes are costly and do not scale. (b) Recent approaches automate this process via single-step semantic assignment using vision-language models (VLMs), often limited in complex scenes. (c) Our method leverages a multi-step chain-of-thought reasoning process to interpret semantically complex scene.

stead 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 recasting pseudo-labeling as an interpretable visual chain-of-thought (CoT) process [50]. Our three-step CoT 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 image captions, 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 CoT 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]. As validated in Tab. 5, under two challenging conditions such as crowding and occlusion, our method demonstrates superior pseudo-label quality compared to previous pseudo-labeling methods [9, 63, 64], 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 [51].

(a) Noisy Pseudo Boxes **(b)** Caption Dependency **(c)** Background Collapse

Figure 2. **Challenges in pseudo-labeling for complex scenes.** (a) Errors in single-step VLM semantic assignment, (b) coarse caption semantics, and (c) unlabeled objects treated as background.

2. Related Work

Chain-of-Thought (CoT) Reasoning. CoT reasoning has emerged as a powerful approach in natural language processing, enabling models to tackle complex reasoning tasks by incrementally decomposing them into interpretable steps. Initial work [49] demonstrated that large language models produced more accurate outcomes by generating intermediate reasoning before arriving at a final answer. In the visual domain, multimodal CoT methods process visual inputs sequentially to reason about future states. These approaches have been applied to diverse tasks, including bounding box prediction [40], planning in autonomous driving [41], intermediate image infillments [38], and CLIP embedding synthesis [12]. Recently, CoT reasoning has been explored in embodied applications, including generating textual plans for multi-stage execution [31] and providing fine-grained reward guidance for reinforcement learning [61]. In the vision-language-action setting, CoT reasoning has recently gained traction for guiding closed-loop robotic manipulation through sub-goal images as interme-

diate reasoning steps [62]. In this work, we extend visual CoT reasoning to generate robust pseudo-labels for open-vocabulary object detection in complex scenes.

Open-Vocabulary Object Detection (OVD). OVD aims to detect novel objects not seen during training by leveraging vision-language models (VLMs) [35] trained on large-scale image-text pairs. Recent OVD methods [6, 7, 53] employ prompt modeling to transfer knowledge through learned prompts, enabling more precise contextual descriptions of each class. Several studies [10, 51] use knowledge distillation to align detectors with VLM features for recognizing unseen objects. Other approaches [14, 26] reinforce the text modality using large language models. In addition, InstaGen [8] focuses on the image modality, improving novel class prediction via synthetic images from an image generation model. Furthermore, Grounding DINO [27] introduces prompt-based object detection by facilitating cross-modal information exchange between VLMs and transformers. Another line of research [16, 20] fine-tunes VLMs with learnable parameters for feature extraction—a process that is often computationally prohibitive. Recently, pseudo-labeling methods [9, 63, 64] have addressed the limitations of restricted base classes by leveraging pseudo-annotations or weak supervision derived from image captions [3]. However, since direct CLIP matching lacks the reasoning needed for complex scenes [54], we reformulate OVD pseudo-labeling into interpretable visual CoT steps for robust performance in challenging environments.

3. CoT-PL: Visual CoT Pseudo-Labeling

We introduce CoT-PL, 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, CoT-PL performs structured multi-step CoT 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. Notably, our framework transcends the reasoning gaps of naive SAM–MLLM coupling.

3.1. Three-Step CoT Pseudo-Label Generation

This section presents a three-step visual chain-of-thought (CoT) 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 [34], 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. 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 CoT 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 CoT 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. 3, regions lacking discernible objects (*e.g.*, plain dark areas) are removed. By narrowing the candidate set

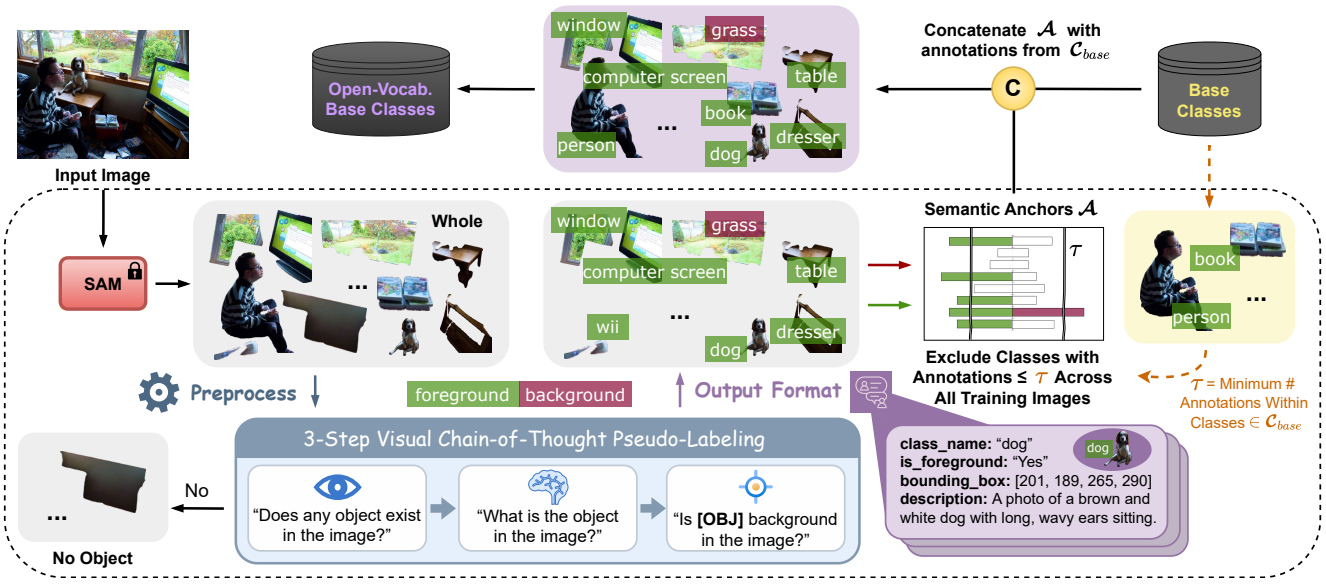

Figure 3. **Our offline CoT pseudo-labeling.** We leverage foundational segmentation and MLLM-based zero-shot reasoning to implement a three-step CoT framework (localization, recognition, grounding) with explicit intermediate reasoning goals. The resulting pseudo-labels are semantically refined and consolidated into the base dataset.

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 CoT 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. 3, 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 [42], 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 CoT step. Such objects remain unlabeled and are consequently assimilated into the learnable background class embedding during training [2, 22]. Moreover, 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 CoT 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, 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 CoT 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 shown in Tab. 6, 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. 3, pseudo-labels falling below a minimum

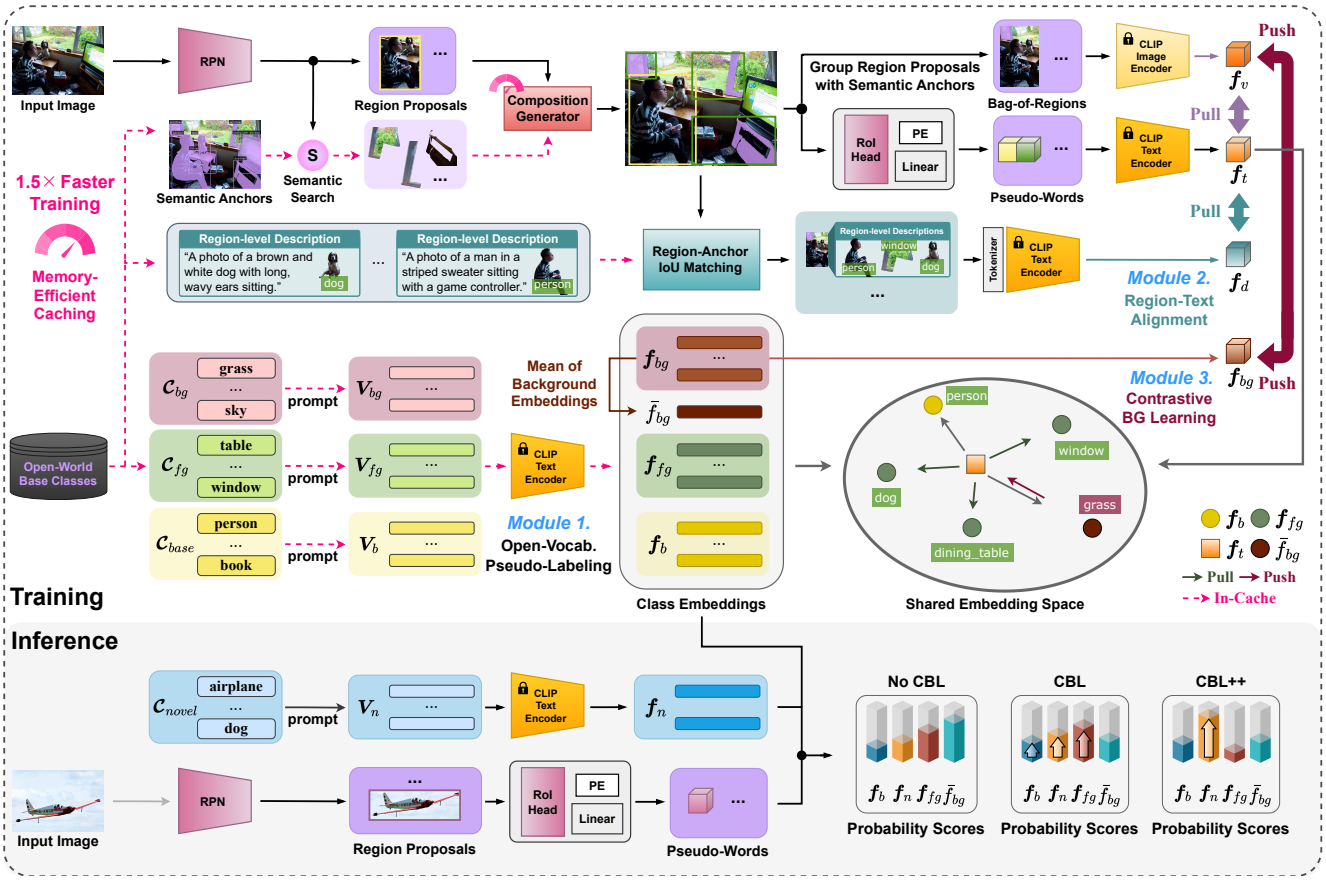

Figure 4. **Our online CoT-PL framework.** Module 1 performs pseudo-label-driven OVD; (2) Module 2 leverages description annotations; and (3) Module 3 utilizes background annotations to promote feature disentanglement. Memory caching accelerates composition augmentation, with pseudo-labels employed exclusively during training.

threshold, derived from the reliable base-class distribution, are discarded. This ensures that the retained labels, termed *semantic 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 CoT Supervision

This section details the online framework of CoT-PL, leveraging pre-computed offline supervision via a contrastive objective. In Fig. 4, our architecture builds upon BARON [51], 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 precomputed via the CLIP text encoder using prompt templates [10, 66]. 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, CBL++ employs pseudo-labels exclusively during training and dis-

cards 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 [39], with cosine similarity $\langle \cdot, \cdot \rangle$ and temperature τ:

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

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. 4, the B background concepts $\boldsymbol{\mathcal{C}}_{bg}$ identified during the third CoT 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 as follows [39]:

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

$$p_{a,b}^k = \frac{\exp(\tau' \langle f_a^k, f_b^k \rangle)}{\sum_{l=0}^{G-1} \exp(\tau' \langle f_a^k, f_b^l \rangle) + \sum_{j=0}^{M-1} \exp(\tau'' \langle f_a^k, f_{bg}^j \rangle)}, \quad (3)$$

where G is the number of bags, τ' and τ'' are scaling factors. The loss promotes alignment of matched foreground pairs and separation of background features.

4. Experiments

Datasets and evaluation metrics. We evaluate CoT-PL on two widely used OVD benchmarks: OV-COCO [25] and OV-LVIS [11]. For OV-COCO, we adopt the category split from OVR-CNN [56], 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 ($\text{AP}_{50}^{\text{N}}$) for novel categories on OV-COCO, and mask mAP (AP_r) for rare categories on OV-LVIS.

Implementation details. CoT-PL is implemented on Faster R-CNN [37] with a ResNet50-FPN backbone. Following recent works [6, 51], the backbone is initialized with SOCO [48] pre-trained weights and fine-tuned using synchronized batch normalization [59]. We employ $1\times$ and $2\times$ training schedules for OV-COCO and OV-LVIS, respectively. For mask generation in the offline phase, we adopt the SAM-Base model with default settings. We also utilize ViT-B-16 [5] with hand-crafted prompts from ViLD [10] by default, with learned prompts [6] only for comparisons on OV-LVIS. The temperature scaling factors τ, τ', and τ'' are fixed at 0.05, 1.0, and 0.1. Following standard benchmarks [25, 33], 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 [51].

4.1. Main Results

Comparison with state-of-the-art methods. We compare CoT-PL against state-of-the-art OVD methods on the OV-COCO and OV-LVIS benchmarks. In Tab. 1, CoT-PL establishes a new state-of-the-art on OV-COCO among recent methods leveraging auxiliary datasets for pseudo-annotations [28, 32, 55, 65]. 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 [46] by 2.1 AP_{50}^{N}. On the OV-LVIS benchmark in Tab. 2, CoT-PL 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, CoT-PL outperforms BIRDet [57] by 0.4 AP_r and surpasses CAKE [29] by 1.4 AP_r. In instance segmentation, CoT-PL exceeds strong baselines [22, 29, 51] by significant margins. This consistent superiority across benchmarks confirms that our proposed strategies scale exceptionally well to large-vocabulary datasets.

Statistics. Tab. 3 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"–"cup") and *Soft Hit* (CLIP similarity > 0.8 following [44], *e.g.*, "vehicle"–"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 the large scale, vocabulary diversity, and broad semantic coverage of our pseudo-

Table 1. **Results on OV-COCO** [25]. Methods are grouped by additional supervision 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 [51]	RN50 (24M)	34.0	60.4
CORA [53]	RN50 (24M)	35.1	35.4
BIND [60]	ViT-B/16 (86M)	36.3	50.2
CLIP-Self [52]	ViT-B/16 (86M)	37.6	-
LBP [22]	RN50 (24M)	37.8	58.7
CCKT-Det [58]	RN50 (24M)	38.0	35.0
CAKE [29]	RN50 (24M)	38.2	-
OV-DQUO [43]	RN50 (24M)	39.2	-
DeCo-DETR [46]	RN50 (24M)	41.3	-
BIND [60]	ViT-L/16 (307M)	41.5	54.8
CCKT-Det [58]	SwinB (88M)	41.9	40.9
CORA+ [53]	RN50×4 (87M)	43.4	43.8
CLIP-Self [52]	ViT-L/14 (307M)	44.3	-
OV-DQUO [43]	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 [66]	IN21K & CC3M	RN50 (24M)	27.8	42.0
OV-DETR [55]	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 [63]	Pseudo annotations	RN50 (24M)	34.4	60.2
RegionCLIP [65]	CC3M	RN50 (24M)	35.2	57.6
OC-OVD [36]	COCO Caption	RN50 (24M)	36.6	49.4
SAS-Det [64]	COCO Caption	RN50 (24M)	37.4	58.5
DITO [17]	LAION-2B	ViT-B/16(86M)	36.6	48.8
LP-OVOD [32]	Pseudo annotations	RN50 (24M)	40.5	60.5
CoT-PL (Ours)	Pseudo annotations	RN50 (24M)	**43.4**	58.9
CFM-ViT [15]	LAION-2B	ViT-L/16 (307M)	34.3	46.4
RegionCLIP [65]	CC3M	RN50×4 (87M)	39.3	61.6
DITO [17]	DataComp-1B	ViT-L/16(307M)	40.2	54.6
CORA+ [53]	COCO Caption	RN50×4 (87M)	43.1	56.2
CoT-PL (Ours)	Pseudo annotations	RN50×4 (87M)	**47.8**	60.9

Table 2. Detection and instance segmentation on OV-LVIS [11].

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 [65]	17.1	27.4	34.0	28.2	-	-	-	-
CCKT-Det++ [58]	18.2	-	-	27.1	-	-	-	-
OV-DETR [55]	-	-	-	-	17.4	25.0	32.5	26.6
VLDet [24]	-	-	-	-	21.7	29.8	34.3	30.1
Detic [66]	-	-	-	-	17.8	26.3	31.6	26.8
MIC [47]	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 [36]	21.1	25.0	29.1	25.9	-	-	-	-
OADP [45]	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 [51]	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 [22]	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 [57]	26.0	21.7	29.5	25.5	-	-	-	-
RALF [18]	21.9	26.2	29.1	26.6	-	-	-	-
CoT-PL (Ours)	**26.4**	**34.8**	38.2	**34.9**	**24.8**	28.5	33.0	28.6

Table 3. Statistics of pseudo-labels generated by our method.

Metric	BLIP2 [21]	InstructBLIP [4]	Qwen2 [1]
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

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, 63, 64] rely on single-step VLM alignment and are susceptible to visual interference, resulting in suboptimal pseudo-labels. On a single A6000 GPU, SAS-Det [64] 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, 63], 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 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 CoT-PL 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—object-level SAM masks and visual context modulation—which alleviates the limitations of naive model coupling. Building on this foundation, transitioning to three-step CoT 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 lin-

Table 4. Ablation of the key components of CoT-PL.

CoT 1-Step	CoT 3-Step	RTA	CBL++	AP_{50}^{N}
✓	-	-	-	37.6 (+3.6)
-	✓	-	-	41.6 (+7.6)
-	✓	✓	-	42.5 (+8.5)
-	✓	✓	✓	**43.4** (+9.4)

Table 5. Per-image pseudo-labeling time and novel-class quality on a GPU.

Method	OV-COCO	*Crowded*	*Occluded*	Time (s)
PB-OVD [9]	18.7	5.1	2.7	0.49
VL-PLM [63]	25.5	7.3	3.8	0.45
SAS-Det [64]	26.7	11.6	5.7	$\gg 1.0$
CoT-PL (Ours)	**32.3**	**23.9**	**15.5**	**0.43**

Table 6. Anchor ablation.

Threshold	AP_{50}^{N}
ALL	42.1
AVERAGE	42.6
MIN	**43.4**

Table 7. Impact of generators.

Proposal generator	AP_{50}^{N}
Mask R-CNN [13]	40.9
MAVL [30]	42.2
SAM [19]	**43.4**

Table 8. Ablation of MLLM variants.

Model	Size	AP_{50}^{N}
BLIP2 [21]	2.7B	39.6
InstructBLIP [4]	7B	42.6
Qwen2 [1]	7B	**43.4**

Table 9. Visual context.

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

guistic attributes, which are essential for disambiguating semantically similar categories. Finally, CBL effectively mitigates background collapse, resulting in additional performance gains. Collectively, these results highlight the complementary roles of structured reasoning and semantically rich supervision in achieving high-fidelity pseudo-labeling.

Impact of semantic anchors. We evaluate semantic anchor policies in Tab. 6, using the reliable base-class distribution as a reference. The ALL policy yield the lowest performance, as they disregard per-class annotation frequency derived from the base-class statistics. The improved result of AVERAGE indicate that lower annotation counts are associated with less reliable anchors. Motivated by this, the MIN policy—the minimum base-class frequency—achieves the best performance by filtering low-quality noise.

Impact of proposal generators. In Tab. 7, we evaluate various class-agnostic proposal generators. SAM [19] provides higher-quality pseudo-box candidates, benefiting from its foundation-level capability to localize arbitrary objects in open-world settings. This contrasts with the vocabulary constraints of closed-set models [13, 30]. The consistent improvements across diverse generators indicate that our CoT reasoning is robust and scales effectively with the open-world localization capacity of foundation models.

Impact of MLLM variants. Tab. 8 analyzes the impact of MLLM scale on pseudo-label quality. While performance remains comparable among models within the same parameter tier (*e.g.*, 7B variants), we observe a linear improvement as the scale increases. Notably, even the compact 2.7B model [21] yields substantial gains over the baseline, demonstrating its effectiveness in resource-constrained settings. These results indicate that our framework is robust to architectural differences at a fixed scale, while translating increased model capacity into improved label fidelity.

Impact of visual context modulation. We examine preprocessing strategies to reduce MLLM sensitivity to visual context. Raw proposals often degrade reasoning accuracy, whereas masking non-target regions improves focus but introduces hallucination due to contextual loss and silhouette artifacts. In contrast, our blurred-and-grayscale strategy suppresses background interference while preserving essential contextual cues. This approach achieves the best performance, balancing target emphasis with environmental context is critical for high-fidelity pseudo-labeling.

5. Conclusion

In this paper, we introduce CoT-PL, a pseudo-labeling framework for open-vocabulary object detection (OVD) that reformulates complex scene understanding via chain-of-thought (CoT) reasoning. CoT-PL addresses the limitations of single-step vision–language alignment, particularly in crowded or occluded scenes, by decomposing labeling into three interpretable stages—object localization, category recognition, and background grounding—where intermediate semantics provide enriched OVD supervision. Concretely, region–text alignment incorporates fine-grained linguistic attributes, while contrastive background learning alleviates background collapse. Extensive evaluations on OVD benchmarks demonstrate that CoT-PL achieves state-of-the-art performance, scales effectively with MLLM capacity, and maintains strong pseudo-labeling efficiency.

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 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).

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, 2023. 3, 6, 7, 8

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, 2018. 4

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, 2015. 1, 3, 4

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 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. 7, 8

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 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

[6] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 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. 3, 6, 7

[7] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. 2022. 3

[8] Chengjian Feng, Yujie Zhong, Zequn Jie, Weidi Xie, and Lin Ma. 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. 3

[9] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. 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. 1, 2, 3, 7, 8

[10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 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. 3, 5, 6, 7

[11] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. 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. 2, 6, 7

[12] William Harvey and Frank Wood. Visual chain-of-thought diffusion models. *CoRR*, 2023. 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017. 8

[14] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. 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. 3

[15] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 7

[16] Dahun Kim, Anelia Angelova, and Weicheng Kuo. 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. 3

[17] Dahun Kim, Anelia Angelova, and Weicheng Kuo. 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. 7

[18] Jooyeon Kim, Eulrang Cho, Sehyung Kim, and Hyunwoo J. Kim. 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. 7

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 2, 3, 7, 8

[20] Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. F-VLM: open-vocabulary object detection upon frozen vision and language models. *CoRR*, 2022. 3

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 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. 7, 8

[22] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. 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. 1, 4, 6, 7

[23] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. 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. 7

[24] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. 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. 7

[25] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 2014. 2, 6, 7

[26] Mingxuan Liu, Tyler L. Hayes, Elisa Ricci, Gabriela Csurka, and Riccardo Volpi. 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. 3

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 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. 3

[28] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. 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. 6, 7

[29] Shiyuan Ma, Donglin Qian, Kai Ye, and Shengchuan Zhang. 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. 6, 7

[30] Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. 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. 8

[31] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 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. 2

[32] Chau Pham, Truong Vu, and Khoi Nguyen. 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. 1, 6, 7

[33] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge J. Belongie, Alan L. Yuille, Philip H. S. Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *Int. J. Comput. Vis.*, 130:2022–2039, 2022. 6

[34] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. 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. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 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. 1, 3

[36] Hanoona Abdul Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman H. Khan, and Fahad Shahbaz Khan. 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. 7

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 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. 6

[38] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings. *CoRR*, 2023. 2

[39] Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. *CoRR*, 2024. 6

[40] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 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. 2

[41] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and

Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, 2024. 2

[42] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan S. Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models? *CoRR*, 2023. 4

[43] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, Weizhi Xian, Yichi Chen, and Yong Xu. 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. 7

[44] Kuo Wang, Lechao Cheng, Weikai Chen, Pingping Zhang, Liang Lin, Fan Zhou, and Guanbin Li. 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. 6

[45] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. 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*, pages 11186–11196, 2023. 7

[46] Siheng Wang, Yanshu Li, Bohan Hu, Zhengtao Yao, Zhengdao Li, Linshan Li, HaiboZhan, Weiming Liu, Junhao Dong, Ruizhi Qian, Guangxin Wu, Zhang, Jifeng Shen, Piotr Koniusz, and Qiang Sun. Deco-DETR: Decoupled cognition DETR for efficient open-vocabulary object detection. In *The Fourteenth International Conference on Learning Representations*, 2026. 6, 7

[47] Zhao Wang, Aoxue Li, Fengwei Zhou, Zhenguo Li, and Qi Dou. 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*, page 93, 2023. 7

[48] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. 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. 6

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 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. 2

[50] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, 2023. 2

[51] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. 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. 2, 3, 5, 6, 7

[52] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. 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. 7

[53] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. 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. 3, 7

[54] Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 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. 1, 3

[55] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. 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. 6, 7

[56] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021. 6

[57] Ruizhe Zeng, Lu Zhang, Xu Yang, and Zhiyong Liu. Boosting open-vocabulary object detection by handling background samples. *CoRR*, 2024. 6, 7

[58] Chuhan Zhang, Chaoyang Zhu, Pingcheng Dong, Long Chen, and Dong Zhang. Cyclic contrastive knowledge transfer for open-vocabulary object detection. *CoRR*, 2025. 7

[59] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. 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. 6

[60] Heng Zhang, Qiuyu Zhao, Linyu Zheng, Hao Zeng, Zhiwei Ge, Tianhao Li, and Sulong Xu. 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. 7

[61] Kaifeng Zhang, Zhao-Heng Yin, Weirui Ye, and Yang Gao. Learning manipulation skills through robot chain-of-thought with sparse failure guidance, 2025. 2

[62] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. 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. 3

[63] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, B. G. Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N. Metaxas. 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. 1, 2, 3, 7, 8

[64] Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, B. G. Vijay Kumar, Yumin Suh, Manmohan Chandraker, and Dimitris N. Metaxas. 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. 1, 2, 3, 7, 8

[65] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. 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. 1, 6, 7

[66] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 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. 5, 7