DiPEx: Dispersing Prompt Expansion for Class-Agnostic Object Detection

Jia Syuen Lim* **Zhuoxiao Chen*** **Mahsa Baktashmotlagh**
Zhi Chen **Xin Yu** **Zi Huang** **Yadan Luo**†
The University of Queensland
{jiasyuen.lim, zhuoxiao.chen, m.baktashmotlagh}@uq.edu.au
{zhi.chen, xin.yu, helen.huang, y.luo}@uq.edu.au

Abstract

Class-agnostic object detection (OD) can be a cornerstone or a bottleneck for many downstream vision tasks. Despite considerable advancements in bottom-up and multi-object discovery methods that leverage basic visual cues to identify salient objects, consistently achieving a high recall rate remains difficult due to the diversity of object types and their contextual complexity. In this work, we investigate using vision-language models (VLMs) to enhance object detection via a self-supervised prompt learning strategy. Our initial findings indicate that manually crafted text queries often result in undetected objects, primarily because detection confidence diminishes when the query words exhibit semantic overlap. To address this, we propose a Dispersing Prompt Expansion (**DiPEx**) approach. DiPEx progressively learns to expand a set of distinct, non-overlapping hyperspherical prompts to enhance recall rates, thereby improving performance in downstream tasks such as out-of-distribution OD. Specifically, DiPEx initiates the process by self-training generic parent prompts and selecting the one with the highest semantic uncertainty for further expansion. The resulting child prompts are expected to inherit semantics from their parent prompts while capturing more fine-grained semantics. We apply dispersion losses to ensure high inter-class discrepancy among child prompts while preserving semantic consistency between parent-child prompt pairs. To prevent excessive growth of the prompt sets, we utilize the maximum angular coverage (MAC) of the semantic space as a criterion for early termination. We demonstrate the effectiveness of DiPEx through extensive class-agnostic OD and OOD-OD experiments on MS-COCO and LVIS, surpassing other prompting methods by up to 20.1% in AR and achieving a 21.3% AP improvement over SAM. The code is available at https://github.com/jason-lim26/DiPEx.

1 Introduction

In real-world applications, the class of interest may constantly change, prompting the need for new tasks like out-of-distribution (OOD) detection [53, 11], open-world detection [60, 52, 22, 62, 55] and open-vocabulary [48, 54, 31, 28] object detection (OD) to ensure reliable operation of detectors. A significant bottleneck in these OD tasks is the ability to locate *all* objects in a scene - typically referred to as class-agnostic OD [36]. Ensuring a high *recall* rate is essential in this task as it lays the foundation for correctly classifying objects, thereby improving the average precision for classes of interest. Conversely, a low recall implies that some objects will be missed entirely, negatively impacting downstream recognition tasks.

---

*Equal Contribution
†Corresponding Author

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

(a) Class-agnostic OD and Downstream OOD-OD (b) UNIVERSAL and CLASS-WIDE Queries
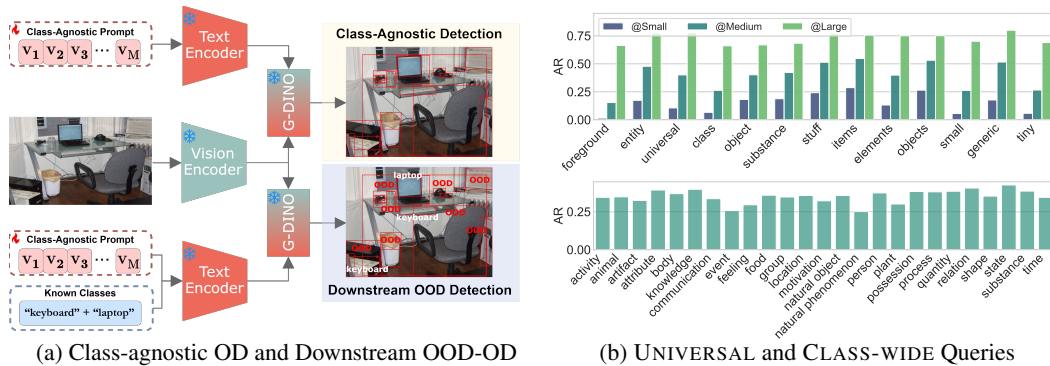
Figure 1: (a) An exemplar of the studied class-agnostic OD and downstream OOD-OD tasks. (B) Zero-shot class-agnostic OD performance of Grounding DINO [33] on MS-COCO [32], with the hand-crafted UNIVERSAL query from ChatGPT and CLASS-WIDE query from WordNet [14].

Conventional solutions to the under-explored class-agnostic OD task often rely on bottom-up strategies [47, 61, 40, 41] such as selective search [47] or EdgeBox [19], which generate a large ranked set of class-agnostic proposals based on low-level visual cues. To address the low precision and scalability issues of these approaches, another line of research has explored multi-object discovery by leveraging (self-)supervised features from vision transformers (ViT) (*e.g.*, DINO [39], MoCo-v2 [5], SwAV [3]), or external motion information to support region proposal regression. However, these methods still fall short, achieving only about 30% average recall (AR) on benchmark datasets like MS-COCO due to the lack of intrinsic knowledge about a wide range of objects. The newly released vision-language models (VLMs) such as Grounding DINO [33], GLIP [29], T-Rex2 [21], which are pretrained on large-scale grounding datasets, have opened up new opportunities for acquiring common knowledge for generic object localization. VLMs have demonstrated impressive zero-shot recognition capacities given the provided *textual prompt*. However, to effectively locate all objects, one would need to input all class names accurately, which is impractical in real-world applications.

To better understand the limitation of modern VLMs in generic object localization, we investigated the design of hand-crafted text queries (Section 2) to enhance detection recall through two approaches: (1) We employed a **UNIVERSAL query,** using ChatGPT to generate 13 types of broad nouns and adjectives (*e.g.*, "objects", "generic") as queries for the Grounding DINO model, aiming to detect a wide array of objects without focusing on specific categories; (2) We implemented a **CLASS-WIDE query**, selecting 25 high-level semantic words (*e.g.*, "plant", "animal") from the top layer of the WordNet hierarchy (also used for the ImageNet vocabulary) to cover extensive object categories. Our findings, depicted in Figure 1b and Table 1, reveal that while VLMs can generalize across universal object categories, combining all queries into *one string* significantly reduces detection performance (by up to 52% in AR) due to the "**semantic overlap**" among words. This suggests that optimal detection requires conducting multiple *separate* inferences, presenting substantial computational demands for large datasets.

To overcome the aforementioned limitations, we propose a novel self-supervised Dispersing Prompt Expansion (DiPEx) strategy. This approach progressively expands a set of non-overlapping hyperspherical prompts for capturing all objects in a given dataset, thereby benefiting downstream tasks such as out-of-distribution object detection. Specifically, we start with a generic parent prompt that is self-supervised using the UNIVERSAL and CLASS-WIDE text queries. To capture more fine-grained semantics, we split the parent prompts with high semantic uncertainty into a set of distinct child prompts. We initialize child prompts by diversifying the parent token embedding, randomly rotating it to different angles on the hypersphere to yield a range of unique prompts. Dispersion losses are employed to minimize semantic overlap among child prompts while maintaining semantic consistency across parent-child prompt pairs. To prevent excessive growth of the prompt sets, we estimate the maximum angular coverage (MAC) of the semantic space as a criterion to terminate the prompt expansion process, balancing semantic richness and computational overhead. Extensive experiments on the MS-COCO and LVIS datasets verify the effectiveness and versatility of the proposed DiPEx strategy. With a single pass of inference, DiPEx can achieve by up to 20.1% improvements in average recall (particularly 35.2% for small objects) and outperforms segment anything model (SAM) [26] by 21.3% in average precision.

2

Table 1: Zero-shot class-agnostic object detection performance of Grounding DINO [33] on MS-COCO [32], with *hand-crafted* prompts from various sources. We report average recall (AR) and precision (AP) limited to a maximum of 100 detections per image. ΔAR quantifies the percentage decrease in AR comparing "query-merging" to "prediction-merging" for forming *multi-word* queries.

Word Source	Merging Strategy	AR	ΔAR	AR@S	AR@M	AR@L	AP
ChatGPT [38]	query-merging	0.345	-52.46%	0.122	0.360	0.718	0.067
	prediction-merging	0.526		0.317	0.606	0.781	0.274
WordNet [14]	query-merging	0.461	-23.64%	0.234	0.522	0.774	0.229
	prediction-merging	0.570		0.382	0.646	0.796	0.344
ChatGPT [38]+WordNet [14]	query-merging	0.408	-44.36%	0.162	0.471	0.751	0.121
	prediction-merging	0.589		0.410	0.665	0.798	0.353

Related Study. The full discussions can be found in Section A.1. Traditional bottom-up approaches for region proposal generation, such as those by [47] and [27], often face precision constraints despite high recall rates, limiting their scalability. Recent advancements in Vision Transformers (ViTs) by [4] and [10] have enabled self-supervised learning on massive datasets, extracting semantically meaningful features. Methods like LOST [45] and TokenCut [51] use graph-based techniques but are limited to detecting a single object per image. MOST [43] addresses this with entropy-based box analysis but struggles with generalization. MAVL [36] uses a late fusion strategy with text queries, requiring full supervision and multiple inferences. Our approach eliminates the need for labels and achieves state-of-the-art performance with one-pass inference using non-overlapping prompts. Vision-Language Models (VLMs), like those by [42] and [20], have shown potential in learning generic concepts. HierKD [35] and OV-DETR [58] align image representations with captions and extend DETR to open-vocabulary settings. GLIP [29], Grounding DINO [33], and T-Rex2 [21] integrate object detection and visual grounding. However, VLMs' effectiveness depends on textual cues, and prompt tuning, as introduced by CoOp [24] and improved by CoCoOp and MaPLe [25], offers a solution by optimizing soft prompts while keeping the model's parameters frozen. ProDA [34] learns diverse prompts using a Gaussian model. DFKD-VLFM [56] and PromptStyler [7] attempted to diversify a fixed number of prompts through contrastive approach. Despite these advancements, full supervision is typically required. UPL [17] and POUF [46] introduced unsupervised prompt learning, but adaptation for object detection remains limited. DiPEx is the first to apply prompt learning to class-agnostic object detection through a progressive self-training approach.

2 Pilot Study

In this section, we detail our preliminary exploration of the zero-shot detection capabilities using state-of-the-art VLM, Grounding DINO [33], to detect all objects irrespective of the associated classes on the MS-COCO dataset [32] as illustrated in Figure 1b. We conduct experiments using two types of text queries: UNIVERSAL queries generated by ChatGPT for general object detection, and CLASS-WIDE queries derived from WordNet, representing broad object categories. Our experiments reveal that semantic overlap between text queries impacts detection performance. To support this hypothesis, we conduct a case study showing that similar concatenated prompts reduce the model's detection confidence.

2.1 Hand-crafted Queries for Class-agnostic Object Detection

UNIVERSAL **Query.** We employ ChatGPT to generate 13 synonyms of **universal** concepts, including nouns and adjectives, which are displayed as x-axis labels. The zero-shot object detection results, measured by average recall (AR) and precision (AP) across the top 100 confident boxes for each query text, are presented. The plot reveals that more general terms such as "generic" and "items" yield the highest AR. Surprisingly, more specific descriptors like "foreground", "small", or "tiny" tend to reduce AR and do not effectively aid in identifying foreground or small objects.

CLASS-WIDE **Query.** We utilize 25 semantically independent beginner words (listed as x-axis labels in the bottom figure) from the highest level of the WordNet hierarchy [14] as **class-wide** text queries. A variation in AR (0.26∼0.43) is observed with different textual queries from WordNet, with a mean AR of 0.35. Compared to the mean AR of 0.37 across class-agnostic queries generated by ChatGPT, the zero-shot detection ability remains similar, regardless of the types of queries used.

Discussion on Multi-Word Queries. The zero-shot results presented in Figure 1b were obtained using *single-word* prompts for the Grounding DINO. To explore whether combining *multiple words*

3
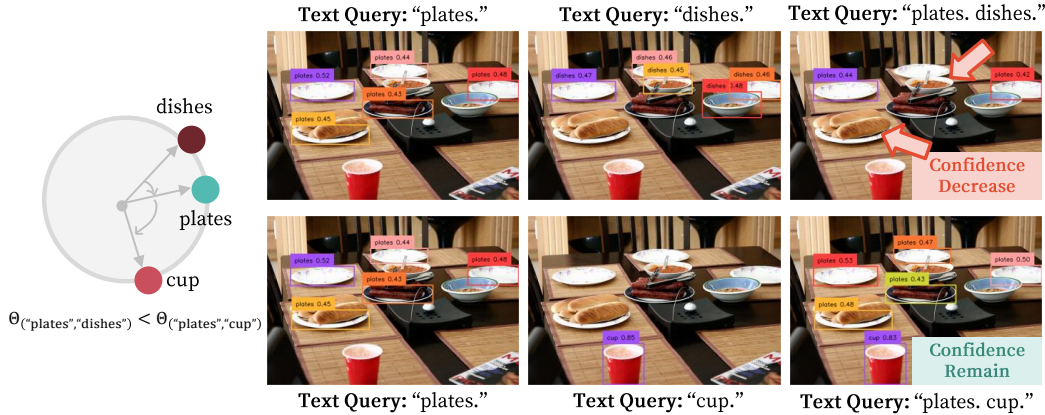
Figure 2: A case study investigating the impact of semantic overlap between text queries on the detection confidence of the pre-trained Grounding DINO [33]. Semantic overlaps are quantified by the angular distance, denoted as Θ, between tokenized embeddings of word pairs using BERT [9].

as prompts from a given source (*e.g.*, WordNet) could improve zero-shot detection performance, we developed strategies for merging at both the input stage (query-merging) and the output stage (prediction-merging) as shown in Table 1. The query-merging strategy concatenates all input text queries (*e.g.*, "foreground . elements . ⋯ tiny . objects .") and performs a single-pass inference to obtain detections. The prediction-merging strategy, on the other hand, uses each text query individually for separate inference and then combines all box predictions. Table 1 shows that applying query-merging to UNIVERSAL words results in a 52.46% reduction in AR compared to prediction-merging, whereas CLASS-WIDE queries (*e.g.*, from WordNet) achieve a smaller decrease in AR of only 23.64%. These findings suggest that large semantic overlaps in concatenated queries (*e.g.*, "stuff", "objects" and "item" from ChatGPT) may greatly contribute to diminished object detection performance. To further investigate this phenomenon, we conducted a case study analyzing the impact of semantic overlap on detection performance, which is presented in the following section.

2.2 Confidence Diminishing when Text Query Semantically Overlap

To verify our hypothesis, we conduct a case study to demonstrate how semantic overlap in multi-word query leads to diminished detection confidence. We quantify semantic overlap by calculating the angular distance between pairs of textual token embeddings generated by BERT [9]. As shown in Figure 2, a small angular distance θ of $53.73°$ between the text tokens ``plates`` and ``dishes`` diminishes the model's confidence. Consequently, some boxes that could be precisely localized with high confidence using the single token ``plates`` are omitted. In contrast, concatenating two text tokens with a larger angular distance (*e.g.*, $60.99°$ between ``plates`` and ``cup``) maintained high detection confidence. This combination resulted in bounding box predictions that encompassed all boxes predicted with each individual token (``plates`` or ``cup``). This case study supports our hypothesis that semantic overlap between concatenated text queries can interfere with the detection confidence of the model. Therefore, we propose that developing a method to learn a set of semantically *non-overlapping* prompts for the target dataset could enable efficient object localization with *one-pass* inference using VLMs.

3 Proposed Approach

In this section, we first mathematically formulate the task of class-agnostic detection using a general VLM and, without loss of generality, illustrate the process using Grounding DINO [33] as an exemplar model. We detail the steps of the proposed dispersing prompt expansion in Section 3.2, followed by the early termination strategy of the prompt set growth.

3.1 Problem Formulation

Class-agnostic OD. Let $\mathbf{I}$ denote the input image and $\mathbf{T}$ the associated text query. For the zero-shot object detection in a class-agnostic setting, we consider the text query $\mathbf{T}$ to be of the form of "`a photo of a {class}`", where the class token `{class}` is sampled from our predefined UNIVERSAL (*e.g.,* "objects") or CLASS-WIDE (*e.g.,* "plant") sets as described in Section 2. The text query is then tokenized and projected into word embeddings as $\mathbf{P} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M, \mathbf{c}\}$, where

4

L-Layer Prompt Expansion   Child Prompt Initialization   Maximum Angular Coverage
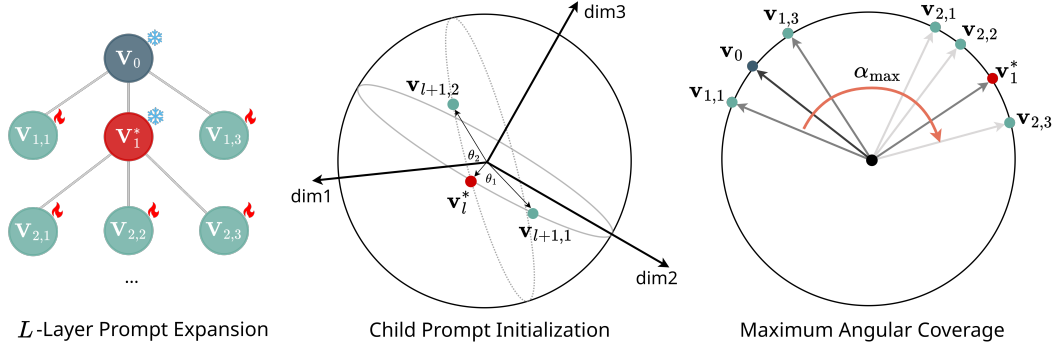
Figure 3: An illustration of the ① proposed prompt expansion strategy that selectively grows a set of child prompts for the highlighted parent prompt across L iterations; ② diversifying initialized embeddings of the child prompt on a hypersphere and ③ quantifying maximum angular coverage $\alpha_{\max}$ for early termination of the prompt growth.

$\mathbf{v} = \{\mathbf{v}_i\}_{i=1}^M \in \mathbb{R}^{M \times d}$ indicates a set of M contextual embeddings and $\mathbf{c}$ is the query text embedding. Here, d indicates the dimensions of learnable tokens. The visual embeddings $\mathbf{E}_v$ extracted from the visual encoder and prompt embeddings $\mathbf{P}$ are fused jointly to prompt the VLM and generate the final bounding box predictions $\mathbf{O} = f(\mathbf{E}_v, \mathbf{P}) \in \mathbb{R}^{N_B \times 4}$, with f being the VLM, and N_B being the number of predicted boxes. Formally, the objective of class-agnostic OD is to ensure that the generated bounding boxes can capture any objects as comprehensively as possible.

Adapt Prompt Tuning for Class-agnostic OD. Instead of relying on hand-crafted templates, prompt tuning approaches like CoOp [24] and CoCoOp [23], originally developed for classification tasks, aim to *learn* the context embeddings $\mathbf{v}$ with a frozen VLM using a supervised contrastive learning loss. To adapt these prompt learning approaches to the Grounding DINO [33] detection framework, we first construct a pseudo label set $\mathcal{D}_{\mathrm{PSL}}$ from the zero-shot detection results with UNIVERSAL and CLASS-WIDE text queries (see Section A.3 for details). The prompt learning is then supervised by the standard box regression loss $\mathcal{L}_{\mathrm{box}}$, $\mathcal{L}_{\mathrm{giou}}$ and focal classification loss $\mathcal{L}_{\mathrm{cls}}$ as implemented in [33].

3.2 Dispersing Prompt Expansion (DiPEx)

Unlike previous prompt tuning approaches, the proposed DiPEx strategy aims to iteratively grow a set of learnable prompts $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_L\}$ in a tree hierarchy of depth L. To maximize the utility of prompts and ensure minimal semantic overlap among them, we assume $\mathbf{v}$ resides on the surface of a unit-hypersphere, *i.e.,* $\|\mathbf{v}_i\|_2 = 1$. This assumption transforms the overlap minimization problem into maximizing the angular distances among the learned prompts. In the initial round, we set a single learnable parent prompt $\mathbf{P}_1 = \{\mathbf{v}\}$, which is self-trained using $\mathcal{D}_{\mathrm{PSL}}$ with the same procedure outlined above. In each subsequent round l for $l \in [1, L]$, we identify the parent prompts of highest uncertainty and grow K child prompts $\mathbf{P}_{l+1} \in \mathbb{R}^{K \times d}$ from it. The learned $\mathbf{v}_l^*$ is then frozen and stored in a parent queue $\mathbf{P}_{\mathrm{parent}}$. Prompt growth is terminated when the maximum angular coverage $\alpha_{\max}$ exceeds a certain threshold $\mathcal{T}_\alpha$.

Child Prompt Initialization. Continuing from the previous discussion, we now describe the process of child prompt initialization, which aims to inherit the semantics from parent prompts while capturing more fine-grained semantics. After the l-th round of training, we expand the parent prompt with the highest uncertainty, denoted as $\mathbf{v}_l^* \subset \mathbf{P}_l$, into a set of learnable child prompts (Figure 3). We empirically adopt the logit activation frequency of the prompts as a measure of uncertainty, visualized in Figure 6. The rationale is that if a prompt is activated for most samples, it covers overly broad semantics (*e.g.,* animals) and may need to be decomposed into narrower categories (*e.g.,* cats and dogs). To disentangle the complex semantic of $\mathbf{P}_l^*$, we set up K child prompts $\mathbf{P}_{l+1} = \{\mathbf{v}_{l+1,k}\}_{k=1}^K$ for the selected parent prompt $\mathbf{v}_l^*$. To diversify the initialized embedding for each child prompt, we introduce K random angular offsets $\boldsymbol{\Theta} = \{\theta_k\}_{k=1}^K$ to rotate $\mathbf{v}_l^*$ on the hypersphere by different angles $\theta_k \sim [-\theta, \theta]$. Given that $\mathbf{v}_l^*$ is a d-dim vector, we randomly sample two axes i and j where $i, j \sim [1, d]$ for rotation. The k-th child prompt embedding $\mathbf{v}_{l+1,k}$ is then obtained by applying the

5

corresponding rotation matrix $\mathfrak{R}_k \in \mathbb{R}^{d \times d}$, which are defined as follows:

$$\mathbf{v}_{l+1,k} = \mathbf{v}_l^* \mathfrak{R}_k, \quad \mathfrak{R}_k = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots & & \vdots \\ 0 & \cdots & \cos\theta_k & \cdots & -\sin\theta_k & \cdots & 0 \\ \vdots & & & 1 & & & \vdots \\ 0 & \cdots & \sin\theta_k & \cdots & \cos\theta_k & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}. \tag{1}$$

Here, the non-identity elements are placed at the intersections of the i-th and j-th rows and columns, corresponding to the plane of rotation, as illustrated by the grey ellipses in Figure 3. As the initialized embeddings of the child prompts are diversified while maintaining consistency with the central parent embedding (red dot), this leads to varying detection results. This enriched prediction diversity allows us to facilitate online self-training, where we adopt the predictions with the highest confidence as pseudo labels for each child prompt, which in turn supervise the next round of prompt learning with respect to $\mathcal{L}_{\mathrm{bbox}}$, $\mathcal{L}_{\mathrm{giou}}$ and $\mathcal{L}_{\mathrm{cls}}$ for the next iteration.

Optimization. We expect the learned child prompts to follow an accurate semantic hierarchy, having minimal overlap with other child tokens while maintaining semantic consistency with their original parent prompts. We leverage the following *dispersion losses* to enlarge the angular distances among the child-child and decrease the distances between child-parent prompt pairs:

$$\mathcal{L}_{\mathrm{parent\text{-}child}} = -\frac{1}{K} \sum_{i=1}^{K} \left(\frac{\mathbf{v}_i^\top \mathbf{v}_l^*}{\|\mathbf{v}_i\| \|\mathbf{v}_l^*\|} / \tau_p \right),$$

$$\mathcal{L}_{\mathrm{child\text{-}child}} = \frac{1}{K} \sum_{i=1}^{K} \log \frac{1}{K-1} \sum_{j \neq i} \exp\left(\frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} / \tau_c \right), \tag{2}$$

where $\mathbf{v}_l^*$ is retrieved from the parent prompt queue $\mathbf{P}_{\mathrm{parent}}$ as a fixed prototype. The temperature coefficients τ_p and τ_c adjust the angular separation. The overall optimization can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{parent\text{-}child}} + \gamma \mathcal{L}_{\mathrm{child\text{-}child}} + \gamma_{\mathrm{bbox}} \mathcal{L}_{\mathrm{bbox}} + \gamma_{\mathrm{giou}} \mathcal{L}_{\mathrm{giou}} + \gamma_{\mathrm{cls}} \mathcal{L}_{\mathrm{cls}}, \tag{3}$$

where γ is the loss coefficient that controls the $\mathcal{L}_{\mathrm{child\text{-}child}}$. The rest coefficients i.e., $\gamma_{\mathrm{bbox}}, \gamma_{\mathrm{giou}}, \gamma_{\mathrm{cls}}$ follows [33]. Until the optimization convergence, the prompt expansion will repeat if needed.

Expansion Termination with Maximum Angular Coverage (MAC). While prompt expansion is effective in capturing fine-grained semantics, it inevitably introduces computational overhead, impacting inference efficiency for downstream tasks. To balance the semantic richness and inference costs, we gather all learned prompts $\mathbf{P}$ and evaluate the maximum angular coverage (MAC) among all pairs. MAC is defined as:

$$\alpha_{\max} = \max_{\mathbf{v}_i, \mathbf{v}_j \in \mathbf{P}} \arccos\left(\frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} \right). \tag{4}$$

The $\alpha_{\max}$ reveals the breadth of vocabularies covered by the current prompts. Notably, our empirical study shows that as the number of expansion rounds increases, the MAC increases monotonically and eventually converges. This convergence serves as an effective *signal* to terminate prompt expansion. The overall algorithm is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct our experiments using two detection datasets: 1). **MS-COCO** [32], a large-scale object detection and instance segmentation dataset, comprising approximately 115K training images and 5K validation images across 80 classes. 2). **LVIS** [15] includes 2.2 million high-quality instance segmentation masks covering 1,000 class labels, resulting in a long-tailed data distribution. It consists of around 100K training images and 19.8K validation images. For class-agnostic object detection (CA-OD) setting, we merge all categories from both datasets into a single class to perform class-agnostic detection. To further validate the efficacy of DiPEx in downstream out-of-distribution object detection (OOD-OD) tasks, we evaluate our method using a rectified version of the OOD-OD benchmark. Unlike previous benchmarks [12], where samples that do not contain ID instances are manually selected and ID and OOD performance are evaluated separately, we tested our approach

6

Algorithm 1 The Proposed DiPEx for Class-Agnostic Object Detection

---

Input: f: vision-language model (VLM)
Output: $\mathbf{P}$: set of fine-tuned prompts for f to detect class-agnostic objects
    Initialize a single learnable parent prompt $\mathbf{P}_1 = \{\mathbf{v}_1\}$
    Optimize $\mathbf{P}_1$ using zero-shot detection results from f
    Initialize a growing set of learnable prompts $\mathbf{P} = \{\mathbf{P}_1\}$ and an empty parent queue $\mathbf{P}_{\text{parent}} = \{\}$
    for each round $l \in \{1, 2, \cdots, L\}$ **do**
        Identify the parent prompt with the highest uncertainty $\mathbf{v}_l^* \in \mathbf{P}_l$
        Freeze $\mathbf{v}_l^*$ and add it to the parent queue $\mathbf{P}_{\text{parent}}$
        Expand $\mathbf{v}_l^*$ into K learnable child prompts $\mathbf{P}_{l+1} = \{\mathbf{v}_{l+1,k}\}_{k=1}^{K}$ via Equation (1)
        Grow the set of learnable prompts: $\mathbf{P} = \mathbf{P}_l \cup \mathbf{P}_{l+1}$
        Optimize the prompts in $\mathbf{P}$ using Equation (3) with $\mathbf{P}_{\text{parent}}$
        Compute maximum angular coverage (MAC) via Equation (4)
        if MAC converges **then**
            Break; terminate the prompt growth
        end if
    end for

Table 2: **Class-agnostic object detection** on the MS-COCO dataset. [ ] indicate the prompt word for Grounding DINO. The prompting methods indicated with '*' are adapted to the OD task.

Method	Description	AR_1	AR_{10}	AR_{100}	AR@S	AR@M	AR@L	AP
Selective Search [47]	non-parametric	0.1	1.1	7.8	0.9	7.2	20.7	0.1
UP-DETR [8]	self-training	0.2	1.4	1.4	0.0	0.2	5.8	0.1
DETReg [1]	self-training	0.6	3.7	12.9	0.2	12.8	35.3	1.4
FreeSOLO [49]	self-training	3.7	9.7	12.6	0.5	12.3	34.1	4.2
Exemplar-FreeSOLO [18]	self-training	8.2	13.0	17.9	–	–	–	12.6
MOST [43]	self-training	3.1	6.4	6.4	0.1	1.6	24.5	3.3
CutLER [50]	self-training	6.8	19.6	32.8	13.7	37.5	60.0	29.6
Grounding DINO ["generic"] [9]	zero-shot	10.3	37.8	44.1	17.7	51.6	80.0	28.3
Grounding DINO+CoOp* [24]	self-training	10.4	39.1	61.3	36.4	72.7	88.8	34.6
Grounding DINO+CoCoOp* [23]	self-training	7.6	34.1	58.1	33.9	68.3	86.1	24.6
DiPEx	self-training	**10.5**	**40.8**	**63.2**	**39.2**	**74.3**	**89.8**	**35.9**

on the MS-COCO, which includes a mixture of both ID and OOD objects. While we followed the settings outlined in OOD-OD [12], with 20 base classes in PASCAL-VOC [13] designated as ID classes and the remaining classes treated as OOD. Our choice of dataset enhances the rigor of our evaluation by combining both ID and OOD instances, providing a more realistic assessment of our method's real-world conditions.

Evaluation Metrics. We report results for class-agnostic object detection on both the MS-COCO and LVIS validation splits. For evaluation, we adopt official metrics from the COCO 2017 challenge. Specifically, we report average precision (AP) at IoU thresholds from 0.5 to 0.95, along with average recall (AR) across the same threshold range. We also report AR by object scale: AR@S for small, AR@M for medium, and AR@L for large objects. Details on our implementation, including those of prior works used as *baselines*, are provided in Appendix A.2.

4.2 Main Results on Class-agnostic OD and OOD-OD

Class-agnostic OD on MS-COCO. To validate our proposed method for class-agnostic object detection, we compared it against ten different baseline methods on the MS-COCO dataset, using various metrics as reported in Table 2. We observed that non-parametric methods generally underperform compared to self-training methods due to their inability to learn and extract semantic and geometric information about objects from the dataset. In contrast, Grounding DINO, leveraging pre-trained knowledge, demonstrates strong zero-shot capabilities and achieves AR_{100} of 44.1% with a single text prompt, "generic". Furthermore, CoOp, which fine-tunes prompts for Grounding DINO, enhances class-agnostic detection performance by 39.0% in AR_{100} compared to direct zero-shot inference. Our method, which expands the learnable prompts to a wider angular distance, surpasses all baselines by achieving the highest performance across all metrics and outperforming the leading baseline, CoOp, by 3.1% in AR_{100}. Notably, for *small objects* which are challenging to localize, our method improves

7

Table 3: **Class-agnostic object detection** on the LVIS dataset. [†] indicate the model is fine-tuned on the LVIS training set by self-training without box annotations.

Method	AR_1	AR_{10}	AR_{200}	AR@S	AR@M	AR@L	AP	AP@S	AP@M	AP@L
Selective Search [47]	0.1	1.1	13.0	6.1	19.9	37.6	0.2	0.4	0.2	0.2
G-DINO ["object"] [9]	4.1	17.9	27.2	13.0	44.1	71.1	5.4	5.6	10.0	9.4
G-DINO ["generic"] [9]	3.8	16.5	20.2	6.5	34.5	67.7	9.0	4.1	17.4	30.7
G-DINO ["items"] [9]	4.0	17.8	28.0	13.9	45.3	70.7	11.6	6.3	19.6	32.0
SAM [26]	–	–	42.7	27.7	66.3	75.5	6.1	–	–	–
[†] CutLER [50]	2.4	9.3	21.8	10.8	35.1	55.5	4.5	2.7	9.1	15.1
[†] HASSOD [2]	0.2	10.6	26.9	15.6	42.2	56.9	4.9	2.8	7.9	12.2
[†] G-DINO + CoOp* [24]	4.2	19.1	40.3	23.6	63.5	83.5	14.0	8.3	23.7	32.3
[†] G-DINO + CoCoOp* [23]	4.2	19.2	40.7	24.1	63.8	84.1	13.6	8.1	22.4	30.1
[†] **DiPEx**	**4.3**	**20.1**	**48.4**	**31.9**	**72.6**	**88.2**	**15.2**	**9.3**	**25.3**	**32.8**

Table 4: The downstream **out-of-distribution object detection (OOD-OD)** on the MS-COCO dataset, where the ground truth boxes contain both known and unknown classes.

Method	KNOWN		UNKNOWN							
	AP	AP50	AR_{100}	AR@S	AR@M	AR@L	AP	AP@S	AP@M	AP@L
Selective Search [47]	–	–	8.3	1.0	8.5	23.2	0.1	0.0	0.0	0.5
MOST [43]	–	–	5.3	0.1	1.3	22.5	0.4	0.1	0.4	1.2
CutLER [50]	–	–	34.5	15.8	41.5	62.7	5.7	2.3	6.9	13.7
VOS [12]	36.6	56.7	10.0	2.2	6.1	27.1	2.8	0.8	2.2	7.2
PROB [62]	28.2	43.8	13.2	1.9	11.2	40.3	0.9	0.6	0.9	2.1
UnSniffer [30]	35.8	55.8	20.6	11.8	19.9	34.8	2.9	1.5	3.1	5.3
G-DINO ["generic"]	**46.3**	**59.7**	43.3	18.0	52.1	82.6	12.5	6.9	17.8	**25.7**
DiPEx	**46.3**	**59.7**	**59.9**	**35.8**	**72.9**	**89.7**	**15.7**	**9.7**	**21.8**	25.2

AR@S by 7.7% compared to CoOp, indicating that expanded prompts better capture a range of object sizes. Additionally, the proposed DiPEx achieved the highest AP of 35.9%, demonstrating the superior quality of class-agnostic detection.

Class-agnostic OD on LVIS. To further validate the efficacy of DiPEx, we conducted extensive experiments on the challenging LVIS dataset, which includes thousands of classes with a long-tail distribution. As shown in Table 3, prompt tuning methods such as CoOp [24] and CoCoOp [23] outperform zero-shot Grounding DINO when using hand-crafted prompts (*e.g.*, "items", "generic", "objects"). Additionally, CoCoOp surpasses multi-object discovery baselines like CutLER [50] and HASSOD [2], by 86.7% and 51.3% in AR_{200}, respectively. Notably, SAM [26], which was pre-trained on a vast of dataset containing millions of images and billions of masks, demonstrates strong zero-shot capabilities, surpassing all other baselines. In contrast, our proposed DiPEx outperforms SAM by 13.3% in AR_{200} and 21.3% in AP after only four epochs of self-training, Furthermore, DiPEx exceeds CoOp by 20.1% in AR_{200}.

Downstream OOD-OD on MS-COCO. To evaluate the generalization of our proposed DiPEx in out-of-distribution object detection (OOD-OD), we compared its performance on both known and unknown classes against various baselines. As shown in Table 4, the zero-shot Grounding DINO uses known class names as prompts, supplemented with a simple "generic" prompt for unknowns, outperforms all other non-VLM methods (*e.g.*, 25.5% higher AR_{100} compared to CutLER [50]). This improvement stems from VLMs leveraging rich semantic knowledge from language models to better comprehend object information in images. DiPEx enhances this further by expanding text prompts in embedding space, enabling it to capture and differentiate objects of varying sizes and diverse semantics from learned classes. This approach delivers a significant performance gain, achieving a 38.3% increase in AR_{100} and a 25.6% in AP increase over zero-shot predictions. Furthermore, the expanded prompts can be directly applied alongside various known class vocabularies to detect unknown objects, eliminating the need for retraining.

4.3 Ablation Study and Model Analysis

We investigate the impact of various factors on prompting performance including the learnable prompt lengths, the number of expansion rounds L, and angular coverage achieved across rounds. To facilitate model analysis, we present the distribution of prompt logit activation and visualization of detection results. Further ablation studies refers to Section A.3.

8
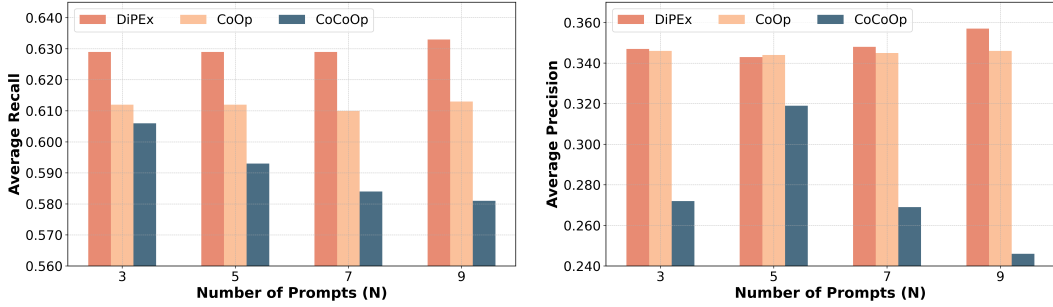
Figure 4: Impact of the prompt length on the MS-COCO dataset. The average recall (AR) and precision (AP) are reported to compare the derived DiPEx against CoOp [24] and CoCoOp [23].
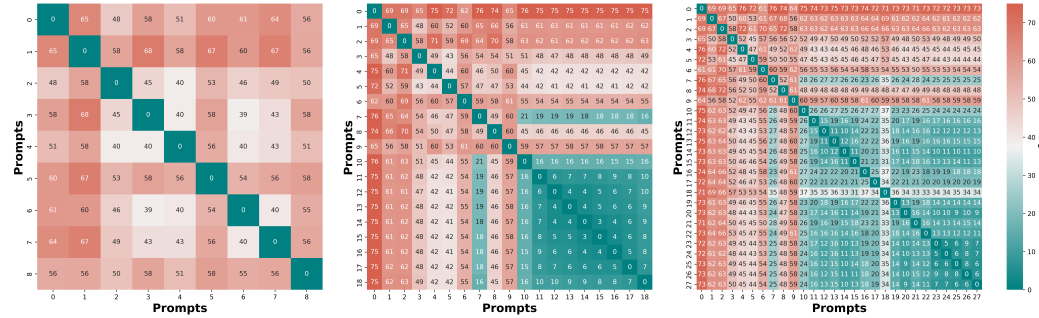
Figure 5: The heatmap visualization presents the **angular coverage** across all learned prompts through the 2nd, the 3rd, and the 4th round of training. The maximum angular coverage (MAC) monotonically increases from **67.7°** in the 2nd round to **75.95°** in the final round. The gradual reduction in rate of change in angular coverage towards the final round suggests that the model nearing convergence.

Impact on Number of Prompts. In Figure 4, we compare the impact of prompt length N for DiPEx against CoOp [24] & CoCoOp [23]. Overall, DiPEx shows consistent improvement in performance with a greater number of prompts – not merely due to quantity, but rather because a larger set fosters greater diversification, enabling the model to capture more comprehensive semantics. In contrast, CoOp's [24] performance remains constant, while CoCoOp's [23] performance declines, suggesting that more prompts do not necessarily guarantee enhanced performance.

Impact on Expansion Rounds and Angular Coverage. To substantiate our hypothesis that a higher maximum angular coverage (MAC) correlates with a broader spectrum of vocabularies, we computed the MAC using Equation (4). The coverage results are visualized as heatmaps in Figure 5. At the initial stage of expansion (leftmost heatmap), we observe that the prompts are quite uniformly distributed, with a mean coverage of 47.56°, This suggests that the prompts are actively exploring the embedding space to capture diverse semantics. As the expansion progresses to the third round (middle heatmap), the MAC increases from 67.78° to 75.70°. Specifically, row/col 7 (selected parent prompt) demonstrates the closest angular distances among the child prompts. This observation is crucial as it suggests that child prompts should not diverge excessively from the root semantics to maintain coherence. By the fourth round of expansion (rightmost heatmap), the pattern remains consistent with the third round. There is a reduced rate of change of MAC, achieving a maximum coverage of 75.95°and a mean coverage of 11.51°among the child prompts. This plateau in MAC indicates that maximum semantic expansion has been reached, suggesting that the model is approaching convergence and further expansion may not be necessary.

The Distribution of Prompt Logit Activation. We previously established prompt logit activation frequency as an uncertainty measure to guide parent prompt selection for splitting. To investigate the dynamics of expanding highly uncertain parent prompts, we visualize the activation statistics (*i.e.*, the frequency of logit activations) of tokens within the 2nd and 3rd expansion rounds. As illustrated in Figure 6, the distribution of these logits exhibits a long-tailed pattern, suggesting substantial uncertainty and numerous semantic overlaps among the mined semantics. The figure on the right demonstrates that, following the expansion of highly activated prompts, the distribution of

9
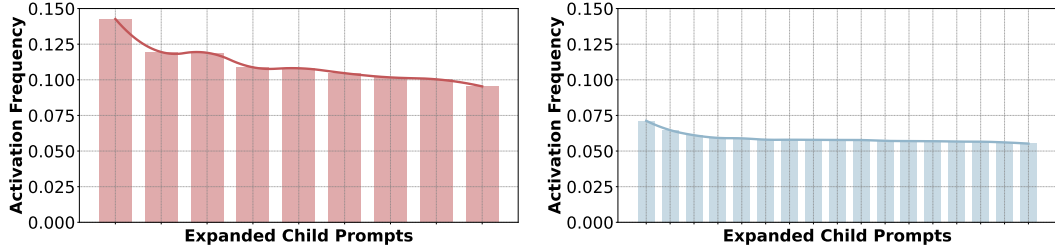
Figure 6: The distribution of **logit activation** of the learned prompts in the 2nd round (*left*) and the 3rd round (*right*). The prompt of the highest activation frequency is identified for further expansion.

| MOST | CutLER | Zero-Shot G-DINO | DiPEx | Ground-Truth |

Figure 7: Visualization of the class-agnostic detection performance by baselines and the proposed DiPEx on MS-COCO [32]. More visualizations are provided in Appendix (Figures 9 and 10).

child prompts becomes more uniform, suggesting the discovery of fine-grained semantics. These observations support our choice of uncertainty measure and verify the validity of DiPEx, indicating that expanding based on highly uncertain parent prompts effectively alleviates semantic ambiguity.

Qualitative Study. In this section, we present visualized class-agnostic box predictions on images sampled from the MS-COCO dataset [32], as shown in Figure 7. The proposed DiPEx method demonstrates a superior ability to detect more bounding boxes than all baseline methods, particularly for *small objects*. For example, people in the distance (rows 1 and 3) and some bonsai (row 2) are missed by all baselines but successfully detected by DiPEx, showcasing its strong capability in localizing challenging small objects. For *large objects*, such as a motorcycle (row 3) and two people shaking hands in the near distance (row 1), DiPEx localizes them with significantly higher confidence compared to the zero-shot predictions of Grounding DINO using the prompt "generic". Additionally, DiPEx successfully identifies objects that are not annotated in the MS-COCO ground truth, such as plates (row 1), a pillowcase (row 2), and a frame on the wall (row 2). This highlights DiPEx's ability to identify a comprehensive set of class-agnostic objects, *even* those missed in human annotations.

5 Conclusion and Limitations

This work introduces DiPEx, a novel self-supervised dispersing prompt expansion approach for class-agnostic object detection. We demonstrate through comprehensive experiments and analysis that DiPEx effectively detects a wide range of unseen objects of varying sizes and achieves broad vocabulary coverage. The progressively expanded prompt sets maintain good angular distances, promoting the formation of a semantic hierarchy and facilitating downstream detection tasks with a single inference pass. While the proposed DiPEx does not rely on box annotations, it requires self-training on the entire dataset for each round of prompt expansion, resulting in increased computational overhead. Additionally, some hyperparameters like temperature coefficients τ_p, τ_c and learnable prompt length K, may require manual tuning for optimal performance. Future research directions include exploring methods to learn hierarchical prompts at once rather than through expansion. Extensive benchmarking on additional downstream tasks, such as open-vocabulary and open-world detection, is necessary to comprehensively validate the proposed approach.

10

Acknowledgments and Disclosure of Funding

This research is partially supported by the Australian Research Council (DE240100105, DP240101814, DP230101196)

References

[1] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J. Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14585–14595, 2022.

[2] Shengcao Cao, Dhiraj Joshi, Liangyan Gui, and Yu-Xiong Wang. HASSOD: hierarchical adaptive self-supervised object detection. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[5] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[6] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr. BING: binarized normed gradients for objectness estimation at 300fps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3286–3293, 2014.

[7] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15656–15666. IEEE, 2023.

[8] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: unsupervised pre-training for object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1610, 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[10] 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 *International Conference on Learning Representations (ICLR)*, 2021.

[11] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. SIREN: shaping representations for detecting out-of-distribution objects. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[12] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations (ICLR)*, 2022.

[13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

11

[14] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

[15] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5356–5364, 2019.

[16] Weizhen He, Weijie Chen, Binbin Chen, Shicai Yang, Di Xie, Luojun Lin, Donglian Qi, and Yueting Zhuang. Unsupervised prompt tuning for text-driven object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2651–2661. IEEE, 2023.

[17] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *CoRR*, abs/2204.03649, 2022.

[18] Taoseef Ishtiak, Qing En, and Yuhong Guo. Exemplar-freesolo: Enhancing unsupervised instance segmentation with exemplars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15424–15433, 2023.

[19] Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic object detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 918–927, 2021.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, volume 139, pages 4904–4916, 2021.

[21] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. *CoRR*, abs/2403.14610, 2024.

[22] K. J. Joseph, Salman H. Khan, Fahad Shahbaz Khan, and Vineeth N. Balasubramanian. Towards open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5830–5840, 2021.

[23] Zhou Kaiyang, Yang Jingkang, Loy Chen Change, and Liu Ziwei. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[24] Zhou Kaiyang, Yang Jingkang, Loy Chen Change, and Liu Ziwei. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

[25] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122. IEEE, 2023.

[26] 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)*, pages 3992–4003, 2023.

[27] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 725–739, 2014.

[28] 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)*, pages 6478–6487, 2023.

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2022.

[30] Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, and Anlong Ming. Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3230–3239. IEEE, 2023.

12

[31] 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 *International Conference on Learning Representations (ICLR)*, 2023.

[32] 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 *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014.

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023.

[34] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5196–5205. IEEE, 2022.

[35] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14054–14063. IEEE, 2022.

[36] 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 *European Conference on Computer Vision (ECCV)*, volume 13670, pages 512–531, 2022.

[37] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8354–8365. IEEE, 2022.

[38] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023.

[40] Pedro H. O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1990–1998, 2015.

[41] Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(1):128–140, 2017.

[42] 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 *International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021.

[43] Sai Saketh Rambhatla, Ishan Misra, Rama Chellappa, and Abhinav Shrivastava. MOST: multiple object localization with self-supervised transformers for object discovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15777–15788, 2023.

[44] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[45] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *British Machine Vision Conference (BMVC)*, page 310, 2021.

13

[46] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. POUF: prompt-oriented unsupervised fine-tuning for large pre-trained models. In *International Conference on Machine Learning (ICML)*, volume 202, pages 33816–33832, 2023.

[47] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.

[48] 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)*, pages 11186–11196. IEEE, 2023.

[49] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and José M. Álvarez. Freesolo: Learning to segment objects without annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14156–14166, 2022.

[50] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3124–3134, 2023.

[51] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufrey-daz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14523–14533, 2022.

[52] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11433–11443. IEEE, 2023.

[53] Samuel Wilson, Tobias Fischer, Feras Dayoub, Dimity Miller, and Niko Sünderhauf. SAFE: sensitivity-aware features for out-of-distribution object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23508–23519. IEEE, 2023.

[54] 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)*, pages 15254–15264. IEEE, 2023.

[55] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, and Junzhi Yu. UC-OWOD: unknown-classified open world object detection. In *European Conference on Computer Vision (ECCV)*, volume 13670, pages 193–210, 2022.

[56] Yunyi Xuan, Weijie Chen, Shicai Yang, Di Xie, Luojun Lin, and Yueting Zhuang. Distilling vision-language foundation models: A data-free approach via prompt diversification. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain, editors, *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 4928–4938. ACM, 2023.

[57] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2022.

[58] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021.

[59] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *CoRR*, abs/2304.00685, 2023.

[60] Xiaowei Zhao, Yuqing Ma, Duorui Wang, Yifan Shen, Yixuan Qiao, and Xianglong Liu. Revisiting open world object detection. *IEEE Trans. Circuits Syst. Video Technol.*, 34(5):3496–3509, 2024.

14

[61] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 391–405, 2014.

[62] Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. PROB: probabilistic objectness for open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11444–11453. IEEE, 2023.

[63] Wei Li Zuwei Long. Open grounding dino:the third party implementation of the paper grounding dino. `https://github.com/longzw1997/Open-GroundingDino`, 2023.

15

A   Appendix / Supplemental Material

This supplementary material includes a comprehensive overview of related work on class-agnostic object detection, vision-language models (VLMs), and prompt tuning. Additionally, we provide detailed descriptions of baselines, implementation details for both the baselines and the proposed method, and an extensive ablation study are provided. The ablation study analyzes the impact of pseudo-labeled supervision and the effect of the hyperparameter γ. Lastly, we present more comprehensive visualizations of class-agnostic box predictions.

- **Section A.1:** Related Work
- **Section A.2:** Baselines and Implementation Details
- **Section A.3:** More Ablation Studies
- **Figures 9 and 10:** Additional Visualizations of Class-Agnostic Box Predictions

A.1   Related Work

Class-Agnostic Object Detection. Traditional bottom-up approaches [47, 27, 61, 40, 41, 6] for region proposal generation, often grapple with the constraints precision, despite high recall rates, reducing their scalability for general use in diverse environments. Recent breakthroughs in ViTs [4, 10, 39] have enabled scaling up to massive datasets for self-supervised learning, extracting both local and global semantically meaningful features. This has led to numerous methods in unsupervised object discovery and localization. LOST [45] is an early application, using a patch similarity graph and an inverse degree map to identify seed patches and extract bounding boxes. TokenCut [51] constructs an undirected graph with image tokens as nodes, applying the normalized cut algorithm [44] for foreground-background segregation. MOVE [37] builds on LOST by employing deep spectral bipartitioning, offering a more principled and effective approach. However, both LOST [45] and TokenCut [51] are limited to detecting a single object per image. MOST [43] addresses this limitation by using entropy-based box analysis (EBA) to segregate foreground tokens. Nevertheless, their performance remains *sub-optimal*, constrained by their limited capacity to generalize across diverse object categories. Closest to our work is MAVL [36], where they develop an MViT with late fusion strategy and use generic text queries like "all objects" to locate objects. However, their framework requires full supervision and multiple inferences with different textual prompts, yet still falls short of achieving optimal performance. In contrast, our approach eliminates the need for labels and achieves SOTA performance with one-pass inference with the non-overlapping prompts.

VLMs and Prompt Tuning. Recent advances in VLMs [42, 20, 57] which are pretrained on expansive image-text pairs have demonstrated significant potential in learning generic concepts. HierKD [35] introduces global language-to-visual knowledge distillation modules, which align global-level image representations with caption embeddings through contrastive loss. OV-DETR [58] pioneered the extension of the DETR framework to an open-vocabulary setting by integrating a conditional binary matching mechanism. GLIP [29] converted object detection into a grounding task, utilizing additional data to align phrase and region semantics. Recently, Grounding DINO [33] introduced a dual-encoder-single-encoder framework to integrate object detection and visual grounding within a unified architecture. Similarly, T-Rex2 [21] synergizes text and visual prompts through contrastive learning, leading to state-of-the-art performance in out-of-distribution object detection. Nonetheless, the effectiveness of VLMs is heavily influenced by the textual cues they are conditioned on, and efficiently adapting them to specific downstream applications remains a substantial challenge as manually engineering optimal prompts can often entail considerable effort and resources [59]. Prompt tuning is a simple yet effective solution to adapt models to specific tasks by optimizing a small number of soft prompts in an end-to-end manner while keeping the original model's parameters frozen. The pioneering work of CoOp [24] introduced context optimization by fine-tuning CLIP using learnable tokens. However, CoOp's generalizability was constrained, a limitation later addressed by CoCoOp [23], which conditioning input tokens on image embeddings. MaPLe [25] advanced this by introducing a multi-modal prompting technique to overcome the limitations of uni-modal prompting methods. ProDA [34] further innovated by learning a distribution of diverse prompts and employing a Gaussian model to capture visual variations. Despite these advancements, an inherent limitation persists across these methods: they all require full supervision. UPL [17] first proposed unsupervised prompt learning for image recognition task, POUF [46] later introduced a similar self-prompting mechanism to minimize entropy using optimal transport. However, these

16

Table 5: The impact on pseudo-labeled supervision on MS-COCO [32] dataset, when applying different pseudo-labels queried on Grounding DINO using different textual cues. In the main paper, we report the performance of DiPEx using merged pseudo labels for the first round of training.

Method	AR_{100}	AR@S	AR@M	AR@L	AP	AP@S	AP@M	AP@L
Grounding DINO @ ["generic"]	44.1	17.7	51.6	80.0	28.3	11.4	33.0	56.5
Grounding DINO @ [25 nouns]	40.5	16.0	46.6	75.0	12.1	4.4	12.0	26.7
Grounding DINO @ merged	51.9	24.8	61.6	85.8	19.1	7.6	19.5	42.0
DiPEx @ ["generic"]	**65.5**	**42.7**	**76.2**	**90.3**	**37.0**	**20.5**	**43.3**	62.4
DiPEx @ [25 nouns]	46.6	18.5	55.7	83.3	13.2	4.0	13.3	30.7
DiPEx @ merged	63.2	39.2	74.3	89.8	35.9	16.4	39.7	**63.8**

methods have yet to be adapted for the object detection domain. To our knowledge, UPT [16] is the only existing work that optimizes prompts using dual complementary teaching specifically for object detection tasks. Our work, DiPEx, represents the first endeavor to apply prompt learning to class-agnostic object detection through a self-training approach.

A.2 Baselines and Implementation Details

Baselines. We compare the proposed approach with fourteen baselines: 1) bottom-up selective search [47] that slides windows of different sizes to locate objects, 2) UP-DETR [8], an unsupervised pre-training method for OD that can be fine-tuned to detect class-agnostic objects. 3) DETReg [1], which learns to localize objects and encode an object's properties during unsupervised pre-training, 4) MOST [43], a multiple objects localizer based on patch correlations without any training, 5) FreeSOLO [49], which unifies pixel grouping, object localization and feature pre-training in a fully self-supervised manner, 6) Exemplar-FreeSOLO [18], an improved approach based on FreeSOLO through exemplar knowledge extraction, 7) CutLER [50], an unsupervised object detection method by encouraging the detector to explore objects missed in extracted coarse masks, 8) HASSOD [2], a clustering strategy that groups regions into object masks based on self-supervised features, 9) CoOp [24] and 10) CoCoOp [23], prompting techniques that utilize learnable vectors to model a prompt's context words, enabling zero-shot transfer to class-agnostic detection, 11) segment anything model (SAM) [26], a foundational model trained on 1 billion masks and 11 million images such that can perform zero-shot transfer to the class-agnostic OD task. For OOD-OD task, we further compare three baseline methods: 12) VOS [12] that regularizes the model's decision boundary between known and unknown classes by training with generated virtual outliers. 13) PROB [62] which utilizes a multivariate Gaussian distribution to learn objectness probability to separate known and unknown objects, 14) UnSniffer [30], which similarly introduces an object confidence, derived from learning known objects with varying degrees of overlap.

Implementation Details. Our code is developed on the Open Grounding-DINO framework [63], and operates on a single NVIDIA RTX A6000 GPU with 48 GB of memory. For our experiments, we choose a batch size of 8 for training, and set hyperparameter $\gamma = 0.1$, $\tau_p = 0.1$, $\tau_c = 0.1$, $\theta = \pm 15°$, $K = 9$, $L = 3$, and while adopting all remaining hyperparameters from the Open Grounding-DINO codebase. We empirically set the $\mathcal{T}_\alpha = 75°$ as our threshold for expansion termination. The original implementation of CoOP was developed for image classification tasks based on CLIP and supervised contrastive learning. We extend CoOP to class-agnostic object detection using pseudo labeling-based self-training, which remains consistent with our approach. All the implementation code and configurations files are provided in supplementary materials and will be publicly released upon acceptance of this work.

A.3 More Ablation Studies

Pseudo-labels Construction For pseudo-labeling, we utilize off-the-shelf Grounding DINO with a "generic" text prompt, which demonstrates considerable zero-shot performance, as illustrated in our pilot study. Additionally, we generate pseudo-boxes by concatenating all 25 beginner nouns from WordNet [14]. We then merge the predictions from these two queries and apply Soft-NMS to eliminate overlaps. In the following Section A.3, we also investigate the performance of DiPEx alongside other prompt-tuning methods on the quality of pseudo-labels.

17

Impact on Pseudo-labeled Supervision. In this section, we investigate how the quality of pseudo-labels used for self-training impacts DiPEx's performance, given our reliance on these training samples. Specifically, we generate pseudo-labels by querying the off-the-shelf Grounding Dino model with three different approaches: 1) using a "generic" text prompt, 2). the 25 beginner nouns from WordNet [14], and a combination of both. As shown in Table 5, the "generic" text prompt alone demonstrated considerable performance. However, we observed an improvement in Average Recall (AR) when merging the predictions generated by "generic" with the 25 beginner nouns, leading us to this study. Consequently, we use these pseudo-labels to self-train our model.

Effect of Loss Coefficient γ**.** To effectively separate child prompts while maintaining semantic coherence between parent and child prompts, selecting an appropriate γ is essential. As illustrated in the bar plot below, a moderate γ value typically yields optimal results. In contrast, a larger value (e.g., $\gamma = 5$) causes child prompts to diverge more significantly, distorting semantic integrity and potentially leading to over-regularization of the model.
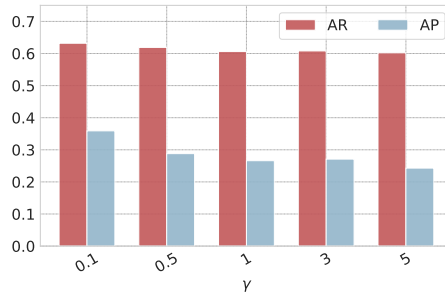
Figure 8: Study of Loss Coefficient γ

18

Figure 9: Additional visualizations of class-agnostic box predictions. Columns 1 – 4 correspond to the following methods: MOST [43], CutLER [50], zero-shot Grounding DINO ["generic"] [9], and our proposed DiPEx, respectively. The final column presents human-annotated ground truth bounding boxes from the MS-COCO dataset [32].

19

Figure 10: Additional visualizations of class-agnostic box predictions. Columns 1 – 4 correspond to the following methods: MOST [43], CutLER [50], zero-shot Grounding DINO ["generic"] [9], and our proposed DiPEx, respectively. The final column presents human-annotated ground truth bounding boxes from the MS-COCO dataset [32].

20

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. The experimental results, proposed methodology, and analysis throughout the paper substantiate the claims regarding performance improvement, novelty, and broader applicability.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the proposed methodology in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

21

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides a comprehensive disclosure of all necessary information to replicate the main experimental results. It offers a detailed, step-by-step algorithm and exhaustive experimental settings in the appendix, facilitating an accurate reproduction process. Additionally, the supplementary materials include the complete source code, along with extensive usage instructions and configuration files for various methods and datasets, thereby ensuring thorough reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

22

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: This paper presents a codebase for class-agnostic detection, featuring detailed usage instructions and configuration files for a wide range of existing methods, models, and datasets. The codebase will be made available in the supplementary materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and evaluation details are clearly outlined in the implementation details section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not report error bars, but results are based on iterative trials to ensure consistency.

23

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The implementation section includes detailed information about the specifications of the computing devices used to run the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The authors have thoroughly reviewed and adhere to the standards of the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

24

Answer: [NA]

Justification: The work performed does not involve any positive or negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not present any potential risks and therefore does not describe any safeguards for the responsible release of data or models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets used, including MS-COCO 2017 and LVIS, as well as the Grounding DINO's codebase, are properly cited and referenced throughout the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.

25

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper benchmarks class-agnostic object detection through extensive experiments, including downstream out-of-distribution object detection. The codebase is well-documented and will be provided in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve any crowdsourcing experiments with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

26

Answer: [NA]

Justification: This paper does not pose any of the above-mentioned risks.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

27