How to Evaluate the Generalization of Detection? A Benchmark for Comprehensive Open-Vocabulary Detection

Yiyang Yao[1], Peng Liu[2], Tiancheng Zhao[3], Qianqian Zhang[2], Jiajia Liao[3], Chunxin Fang[3], Kyusong Lee[3], Qing Wang[1]

[1] Northwestern Polytechnical University
[2] Linker Technology Research Co. Ltd
[3] Binjiang Institute of Zhejiang University
yaoyiyang@mail.nwpu.edu.cn, {liu_peng,zhang_qianqian}@hzlh.com,
{tianchez,liaojiajia,fangchunxin,kyusongl}@zju-bj.com, qwang@nwpu.edu.cn

Abstract

Object detection (OD) in computer vision has made significant progress in recent years, transitioning from closed-set labels to open-vocabulary detection (OVD) based on large-scale vision-language pre-training (VLP). However, current evaluation methods and datasets are limited to testing generalization over object types and referral expressions, which do not provide a systematic, fine-grained, and accurate benchmark of OVD models' abilities. In this paper, we propose a new benchmark named OVDEval, which includes 9 sub-tasks and introduces evaluations on commonsense knowledge, attribute understanding, position understanding, object relation comprehension, and more. The dataset is meticulously created to provide hard negatives that challenge models' true understanding of visual and linguistic input. Additionally, we identify a problem with the popular Average Precision (AP) metric when benchmarking models on these fine-grained label datasets and propose a new metric called Non-Maximum Suppression Average Precision (NMS-AP) to address this issue. Extensive experimental results show that existing top OVD models all fail on the new tasks except for simple object types, demonstrating the value of the proposed dataset in pinpointing the weakness of current OVD models and guiding future research. Furthermore, the proposed NMS-AP metric is verified by experiments to provide a much more truthful evaluation of OVD models, whereas traditional AP metrics yield deceptive results. Data is available at https://github.com/om-ai-lab/OVDEval

Introduction

Open vocabulary detection (OVD) models have experienced rapid development in recent years, with numerous innovative techniques being introduced to the field. Novel models such as GLIP (Li et al. 2022b), Grounding DINO (Liu et al. 2023) and OmDet (Zhao et al. 2022a) have introduced new vision-language learning methods such as modeling detection as visual grounding (Kamath et al. 2021; Li et al. 2022b), pre-training with coarse image-text pairs (Dou et al. 2022), and multi-task learning with a variety of detection tasks (Zhao et al. 2022a).

As a result, for the first time, we can achieve strong zero-shot object detection (OD) on popular datasets such as COCO (Lin et al. 2014), even surpassing the performance of some of the supervised methods (Liu et al. 2023). Users can simply use natural language to specify the desired targets and OVD models can detect the described targets on the fly, which opens doors for many new applications such as interactive image-editing (Shen et al. 2023), Augmented Reality (Li et al. 2023) and robotics (Shah et al. 2023).

Meanwhile, current common approaches to evaluate OVD models include zero-shot/few-shot testing on OD dataset with common objects like COCO (Lin et al. 2014), OD dataset with long-tail objects like LVIS (Gupta, Dollar, and Girshick 2019), grounding such as Flickr30K (Plummer et al. 2015) and referral expression comprehension (REC) such as RefCOCO (Yu et al. 2016). These datasets were challenging for traditional OD research, but no longer serve as a challenging enough benchmark for future OVD methods for the following reasons:

- **Lack of systematic probing of model's generalization ability:** An ideal OVD model should be able to understand the fine-grained semantics in the language prompt and align the language with visual features. Thus, it is required to probe the OVD model from various linguistic aspects such as object type, visual attributes, object relationship, etc., to quantify an OVD model's generalization to various degrees of prompt complexity.
- **Lack of hard negative for real-world usage:** Existing grounding and REC data assume the text prompt is paired with the image. The OVD model is only required to localize the entities mentioned in the caption without the need to discriminate against hard negatives. However, real-world usages command an OVD model to detect described object without knowing if the caption is related to the image at all.

To address the above issues, this paper introduces OVDEval to provide a comprehensive evaluation of OVD models and test their robustness against hard negatives. OVDEval is inspired by behavioral testing (Ribeiro et al. 2020; Zhao et al. 2022b), and consists of 9 large datasets that cover 6 linguistic aspects: *object*, *proper noun*, *attribute*, *position*, *relationship*, and *negation*. All of the data annotations are carefully annotated by human experts to guarantee data quality. Additionally, these sub-datasets are meticulously crafted to ensure that all negative labels are hard. As a result, OVDE-

val is able to rigorously test a model's true understanding of a given aspect, preventing them from achieving high scores on a particular dimension by taking advantage of data bias.

Besides the proposed dataset, this work also proposes a new evaluation metric named Non-Maximum Suppression Average Precision (NMS-AP). We identifies the *The Inflated AP Problem* where even with high-quality hard negatives, a poor OVD model can still achieve a deceptive high AP score due to limitations on the calculation process of AP. The proposed NMS-AP is able to effectively resolve the Inflated AP Problem issue and offers a truthful evaluation of OVD models.

We compared six strong baseline models on the proposed OVDEval dataset. Experimental results show that the current state-of-the-art (SOTA) OVD models only achieve strong results in simple object detection, and performance drop significantly on visual attribute understanding, commonsense knowledge and etc. This shows the significance to have a comprehensive and truthful benchmark to reveal the weakness of SOTA systems and guides the direction of future improvement. Analysis results also confirm the effectiveness of the proposed NMS-AP metric, whereas the conventional AP score is 30% higher than the model's actual performance. Further analysis indicates that the current OVD model is only able to detect object types reliably and shows how OVD models can deceive conventional AP metrics by predicting multiple bounding boxes for each potential target object.

The contributions of our work are summarized as follows:

- We introduce the first OVD evaluation benchmark that comprehensively tests model abilities across six linguistic aspects with complex language prompts and well-designed hard negatives.
- We identify the inflated AP problem that applies to any OVD model with traditional AP metric.
- We propose NMS-AP, a novel evaluation metric that addresses the inflated AP score problem associated with traditional AP and we show NMS-AP provides a more accurate evaluation of OVD models' performance when dealing with fine-grained described detection.
- We show extensive experiment results that reveal the limitations of current SOTA OVD models and verify the effectiveness of the proposed metric.

Related Work

Progression from Fixed Labels to Open Vocabulary Expressions: Traditional object detectors, such as Faster R-CNN (Ren et al. 2015) and YOLO (Redmon et al. 2016), rely on a closed-set vocabulary and are trained on datasets like COCO (Lin et al. 2014) and Pascal VOC (Hoiem, Divvala, and Hays 2009) with predefined categories.

Over time, the number of labels increased, with Object365 (Shao et al. 2019) introducing 365 labels and LVIS (Gupta, Dollar, and Girshick 2019) surpassing a thousand. Also, datasets like ODinW (Li et al. 2022a) focus on wilderness objects with 35 different domains. V3Det (Wang et al. 2023) further broadened object detection capabilities across an extensive range of categories, paving the way for OVD. In addition to object detection, a growing body of research is dedicated to referral expression comprehension (REC) and visual grounding. REC focuses on identifying objects based on textual descriptions provided. Notable datasets in this area include RefCOCO (Yu et al. 2016), PhraseCut (Wu et al. 2020), Flickr30K (Plummer et al. 2015) and Visual Genome (Krishna et al. 2017). The Described Object Detection (DOD) introduced recently, combines the principles of object visual detection and REC, with the goal of detecting objects across various described categories. However, the above-mentioned datasets often lack hard negatives, which can lead to models detecting objects based on general terms rather than recognizing fine-grained details. Moreover, existing datasets have not investigated the model's ability to utilize common sense knowledge for detecting objects such as landmarks, logos, and celebrities.

Endeavor for Systematic Model Evaluations: Benchmark scores often do not provide a comprehensive understanding of a model's capabilities, as they tend to present a superficial evaluation that can be difficult to interpret. Consequently, researchers have sought to scrutinize machine learning (ML) models with greater precision and granularity. In the realm of natural language processing (NLP), CheckList (Ribeiro et al. 2020) evaluates a wide range of linguistic competencies, revealing the limitations of numerous leading NLP models. For computer vision, the Vision CheckList (Du et al. 2022) assists system developers in understanding a model's potential by introducing various transformation techniques to generate an extensive array of test samples. In the vision-language multimodal domain, VL-Checklist (Zhao et al. 2022b) serves as a framework for examining the proficiency of vision-language processing (VLP) models.

In the field of OD, studies often report conventional Average Precision (AP) scores. However, without an in-depth analysis, these scores can be challenging to understand. To address this limitation, we propose a novel evaluation approach that investigates a model's proficiency across clearly defined dimensions. Additionally, we introduce an evaluation metric designed to tackle the problem of deceptively high AP scores.

OVDEval Benchmark

The utilization of commonly employed OD datasets is associated with certain limitations. Firstly, evaluating OD performance solely based on AP across all labels in these datasets provides only a basic assessment. The specific capabilities of the model, such as accurately identifying object positions, have not been thoroughly evaluated. Moreover, in order to maintain linguistic label diversity and comprehensiveness, the distinctions between labels within the same dataset are typically coarse-grained and easily distinguishable. However, the OD task in the real world is much more challenging than merely detecting obvious objects or expressions. It is crucial to include hard negative samples that possess similar linguistic meanings but refer to different objects. Considering these concerns, we propose a new comprehensive benchmark dataset called OVDEval. OVDEval is divided into 9 sub-datasets, each focusing on evaluating the OD capabilities across 6 aspects: *object*, *proper noun*, *attribute*, *position*,

relationship, and *negation*. The utilization of this benchmark dataset offers 3 significant benefits:

- Detailed understanding of OD models: By evaluating OD models across different linguistic aspects, we can gain a more detailed understanding of their performance. This allows us to gain insights into the strengths and weaknesses of OD models, thereby facilitating the identification of areas for improvement.
- Commonsense understanding performance: OVDEval is specifically designed with linguistic queries, including commonsense knowledge-based labels, which enable us to assess the model's commonsense capabilities in the context of multimodal OVD. This evaluation sheds light on how well the model interprets knowledge.
- Fine-grained hard negative labels: we have carefully selected hard negative samples that conflict with the ground truth labels for each object, which provide a straightforward assessment of the model's performance in specific aspects.

Dataset Description

We comprise the benchmark dataset from three viewpoints. Firstly, in line with existing datasets that primarily focus on evaluating the detection of common objects, we employ the COCO dataset to assess the models' general ability in this domain. Additionally, we aim to investigate the models' capacity to leverage external knowledge and common sense. Therefore, *landmark*, *logo*, and *celebrity*, which require knowledge in both vision and language are implemented. Samples for the three aspects are shown in Figure 1. Furthermore, to delve into the models' proficiency in localizing fine-grained details, we divide the dataset into attributes (*color* and *material*), *relationship*, *position*, and *negation* aspects. Figure 2 shows the detail-oriented dataset samples with corresponding fine-grained hard negative samples, which essentially raise the detection difficulty. Finally, 9 sub-dataset across 6 aspects are collected and described as follows:

Figure 1: Samples of Proper noun datasets.

- ***Object*** is utilized to evaluate the general capability in identifying common objects on the COCO Val 2017 (Lin et al. 2014), which covers 80 common object categories.
- ***Proper noun*** can unveil a model's comprehension of commonsense knowledge including famous landmarks, renowned logos, and celebrities.
- ***Attribute*** is used to assess OVD model's proficiency of distinguishing object characteristics. Specifically, *color* and *material* are employed as representing attribute aspects.

Figure 2: Detail-oriented dataset samples. Ground-truth labels are annotated with red color, and fine-grained hard negative samples are shown in black.

- ***Position*** aims to evaluate identifying specific objects among multiple visually similar items within a given image. The evaluation entails determining the target object based on the spatial relationships with other described object expressions.
- ***Relationship*** involves the examination of interactions between humans and other objects to comprehend both active and passive relationships among multiple objects.
- ***Negation*** focuses on identifying objects expressed negatively, like spotting kitchen staff not wearing gloves. This checks the model's skill in detecting objects expressed in a negated context.

Dataset Collection Process

Image Collection We collected varied images from three main sources. We used popular datasets, notably COCO and HICO (Chao et al. 2018). For evaluation, the COCO Val 2017 was directly used. For *relationship*, the HICO dataset was the key source. After selecting the top-most frequent interaction label and excluding it, this selection process was repeated 10 times. We also changed active expressions to passive ones, ensuring two distinct labels for each sample image.

For the *color* sub-dataset, we identified the top 50 objects from the visual genome (VG) dataset (Krishna et al. 2017), labeling them using Oscar (Li et al. 2020). This enabled labeling objects from VG with colors. We concentrated on six object categories and six distinct colors, leading to 36 object-color combinations. Images were then randomly chosen from VG based on Oscar's labels.

For other datasets (*landmark*, *logo*, etc.), images came from the Laion-400m dataset (Schuhmann et al. 2021). We began by identifying key terms for each subset. Using CLIP (Radford et al. 2021), a top-tier image-text match model, images were sourced based on these keywords. To ensure variety, we crafted specific search prompts, considering context

and diversity. For *position* and *negation*, we added terms like "multiple" to get images with several similar items.

Hard Negative Labels We have implemented a novel approach that incorporates fine-grained hard negative labels for each linguistic aspect. These carefully selected hard negative labels are specifically designed to challenge the models and prevent them from achieving high scores on particular aspects without a genuine understanding.

- For *color*, variations in colors with the same object category are used to serve as negative labels. This approach exposes the OVD models to different color representations of the same object, thereby testing their ability to accurately distinguish and classify objects based on color.
- For *material*, we maintain consistency in the object category while introducing variations in materials.
- For *relationship*, we maintain the same subject and object entities but alter the verbs used to describe their relationship.
- In *position*, we introduce changes in position words to serve as negative labels. For example, the words left can be replaced with right, above, under, front, back, and in.
- For *negation*, we remove the word "not" from the positive labels as negative labels.

The datasets with detail-oriented negatives essentially challenge the OVD models toward the advancement of object understanding in natural language processing.

Manual Annotation To ensure the accuracy and reliability of the dataset, we engaged a team of OD annotation experts to manually annotate the collected images with a rigorous annotation process and a thorough quality inspection process. During the annotation process, any images with ambiguous labels were carefully identified and filtered out, guaranteeing the integrity of the final dataset. All the bounding boxes for the corresponding objects were annotated.

Statistics

As shown in Table 1, the full OVDEval dataset comprises 9 distinct sub-datasets, collectively offering a total of 20K high-quality images accompanied by 3K meticulously annotated labels. The statics of each sub-dataset is provided in Table 1. Notably, each sub-dataset encompasses a range of 1K to 5K images, ensuring the diversity and representatives of samples. While some sub-datasets feature proper nouns with a limited number of labels, it is important to highlight that all other sub-datasets can be considered as open set labels. Moreover, these sub-datasets incorporate extremely hard negative labels, further pushing the boundaries of model performance and evaluation. The inclusion of open set labels and hard negative labels within the majority of the sub-datasets enhances the dataset's realism and reflects the complexities encountered in real-world scenarios.

The Proposed Evaluation Metric

The Inflated AP Problem

AP is defined as the area under the precision-recall curve. This metric evalautes a model's performance by considering the trade-off between precision and recall. Recent OD research has predominantly used the COCO AP as the major benchmark metric (Zhang et al. 2022; Zong, Song, and Liu 2023). In the COCO mean Average Precision (mAP) calculation, a 101-point interpolated AP definition is utilized. Specifically, for COCO, AP is determined as the average across multiple Intersection-over-Union (IoU) thresholds that determine a positive match. AP@[.5:.95] represents the average AP for IoU values ranging from 0.5 to 0.95, with a step size of 0.05.

Considering a scenario where an OVD model demonstrates good zero-shot performance in detecting objects but totally does not understand contextual descriptions, the model can deceive traditional AP metrics and obtain a high score by generating multiple predicted bounding boxes for the target object with all candidate labels. Assuming an image with 2 annotated ground-truth instances, which are labeled as *red car* and *blue car*, respectively. Then, the aforementioned model predicts 4 bounding boxes, generating 2 for each target object and assigning both candidate labels to each box. The IoUs between predictions and corresponding ground-truth instances are assumed to be greater than 0.95. As a result, the precision and recall for each category can be derived using the following equation:

$$Precision = \frac{TP}{TP + FP} = \frac{1}{1+1} = 0.50 \quad (1)$$

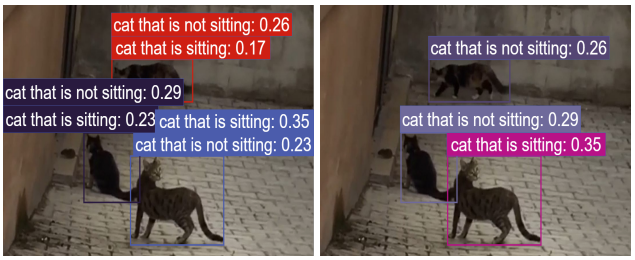
$$Recall = \frac{TP}{GTnum} = \frac{1}{1} = 1.0 \quad (2)$$

Here, TP is the number of correctly predicted instances for a specific category, while FP is the number of instances that were incorrectly predicted as belonging to that category. GTnum represents the total number of ground-truth instances in the image. In the given scenario, where the IoU of predictions is assumed to be greater than 0.95, we can ignore the AP calculation process for IoU values ranging from 0.5 to 0.95. Therefore, we can calculate the average AP of each category as 0.50. Consequently, the mAP would also be 0.50. In this case, the model deceives traditional AP metrics to get an mAP score of 0.50, even though it only detects the target objects without comprehending their descriptions. In this case, the conventional COCO AP metric demonstrates a vulnerability that we refer as *The Inflated AP Problem*. During the stage of matching predictions with ground truth to count TP and FP, it only considers predictions that have the same label as the ground truth. As a result, OVD models can obtain inflated AP scores by simply predicting multiple bounding boxes on a single object with all possible labels.

The inflated AP problem can lead to misleading evaluations of OVD models, as it fails to capture the accuracy of the descriptive labels assigned to the objects. Therefore, it is essential to develop alternative evaluation metrics that consider both object detection and the understanding of linguistic descriptions to provide a more robust assessment of OVD models.

	Object	Attribute		Proper noun			Relationship	Position	Negation
	COCO	Color	Material	Landmark	Logo	Celebrity			
Images	5,000	1,170	2,124	1,533	1,935	2,244	2,169	2,109	1,858
Bboxes	36,781	3,421	5,358	1,709	2,329	2,244	8,190	2,150	3,785
Labels	80	36	90	9	9	10	319	7,301	2,414
Avg. negative labels	-	5.01	8.73	8.00	8.00	9.00	7.65	3.06	1.00
Avg. label tokens	6.03	11.56	11.01	14.37	11.24	12.15	24.14	47.08	27.34
Avg. label words	1.10	2.00	2.03	2.66	2.00	2.12	4.48	9.67	5.35

Table 1: Statistics of OVDEval for the 9 sub-datasets. OVDEval provides fine-grained annotations with hard negatives.



(a) before NMS (b) after NMS

Figure 3: Examples of predictions from GLIP before and after class-ignored NMS, showing the limitation of current OVD models.

A Simple Fix: NMS-AP

To address the aforementioned issue, we propose a simple fix for the COCO AP metric, which we refer to as NMS-AP. It extends the traditional COCO AP metric by incorporating NMS (Girshick 2015), a technique that is used in OD tasks to eliminate redundant bounding box predictions by selecting the most relevant ones based on their confidence scores and suppressing overlapping bounding boxes based on IoU. Specifics of NMS-AP are outlined below Algorithm 1.

Algorithm 1: NMS-AP Metrics

Input: $preds$: predictions
Input: GT: ground-truth
$pickedPreds = keepPreds = []$
for k *in* GT **do**
  for p *in* $preds$ **do**
    if IoU(p, k) >0.5 **then**
      $pickedPreds = pickedPreds \cup p$
    else
      $keepPreds = keepPreds \cup p$
    end if
  end for
  $keepPreds = keepPreds \cup$ *C-NMS*$(pickedPreds)$
end for
$mAP = AP(keepPreds, GT)$
return mAP

In NMS-AP, instead of considering only the prediction with the highest confidence score for each object, we apply a class-ignored NMS (C-NMS) to remove redundant predictions that match ground truth. To be specific, we employed class-ignored NMS on the predictions that exhibited an IoU >0.5 when compared to the ground-truth instances. This ensures that multiple bounding boxes predicted for the same object are appropriately handled and only use the prediction with the highest confidence. In an ideal scenario with a flawless OVD model, it should predict bounding boxes with the correct label and the highest confidence score for each ground-truth instance. Consequently, the application of class-ignored NMS will solely remove false positives, ensuring that this model achieves a perfect score of 1.0. However, in the case of a subpar model that struggles to comprehend complex linguistic descriptions, the application of class-ignored NMS may lead to a decrease regarding true positives and NMS-AP scores (Figure 3). This is because of the failure of accurately predict the bounding boxes that correspond to the ground-truth instances due to its limited understanding of the linguistic context.

Note that NMS-AP is model-agnostic and can be apply to any OVD models. It simply takes a set of predictions and ground-truth bounding boxes, and removes overlapping predictions adjacent to the ground-truths.

Results and Analysis

We conducted experiments on 9 datasets across 6 aspects using several leading publicly available models: Detic (Zhou et al. 2022), MDETR (Kamath et al. 2021), GLIP (Li et al. 2022b), FIBER (Dou et al. 2022), OmDet (Zhao et al. 2022a) and Grounding DINO (Liu et al. 2023). We provide these detailed model information such as pretraining data, backbone, and the number of parameters (Table 2).

Main Results on NMS-AP on OVDEval

The experimental results, as presented in Table 3, show that current models generally perform satisfactorily on the *object* task, with the exception of MDETR. This observation is consistent with earlier work that reported MDETR's low performance on the COCO dataset (Cai et al. 2022). This indicates that most existing models possess strong capabilities in detecting objects. However, we observe that all current models exhibit poor performance on the *logo*, *landmark*, and *celebrity* tasks in *proper noun* aspect. Especially the NMS-AP values are close to 0% in celebrity tasks. Notably, Detic demonstrates impressive results on the *logo* and *landmark* tasks, even without employing a complex fusion strategy, while its performance is relatively weak on tasks involving longer descriptions.

Model	Pre-train Data	Backbone	Params
Detic	ImageNet-21K,COCO,LVIS	Swin-B	141.6M
MDETR	VG,Flickr30k,COCO image-text pairs	ResNet-101	185M
GLIP	FourODs,GoldG,CC3M+12M,SBU	Swin-L	430.42M
FIBER	Flickr30k, MixedNoCOCO, O365	Swin-B	252.06M
OmDet	O365,GoldG,PhraseCut,HOI-A,VAW,RefCOCO	ConvNext-B	241.5M
Grounding DINO	COCO,O365,GoldG,Cap4M,OpenImage,ODinW-35,RefCOCO	Swin-B	232.9M

Table 2: The relevant information of different models include pre-train data, backbone, and parameters.

Aspects	Sub-datasets	GLIP	FIBER	Grounding DINO	Detic	MDETR	OmDet
		NMS-AP/AP	NMS-AP/AP	NMS-AP/AP	NMS-AP/AP	NMS-AP/AP	NMS-AP/AP
Object	COCO	48.90 / 51.30	46.80 / 49.30	52.50* / 55.30*	45.30* / 45.80*	1.60 / 3.20	**54.68** / 57.50
	Logo	10.20 / 17.61	6.30 / 9.05	**10.30** / 14.60	9.60 / 9.60	0.90 / 4.60	6.10 / 11.00
	Landmark	20.30 / 36.36	11.00 / 16.99	15.10 / 23.40	**30.00** / 30.08	1.80 / 7.80	26.30 / 32.38
	Celebrity	**4.60** / 8.24	0.80 / 3.31	0.70 / 2.00	0.00 / 0.00	1.10 / 4.80	1.80 / 6.36
Proper Noun	Avg	11.70 / 20.74	6.03 / 9.78	8.70 / 13.33	**13.20** / 13.23	1.27 / 5.73	11.40 / 16.58
	Color	3.70 / 6.70	6.80 / 9.40	9.40 / 12.41	3.90 / 4.14	3.10 / 7.30	**22.90** / 24.56
	Material	7.40 / 15.87	12.40 / 17.72	9.00 / 15.50	9.20 / 9.75	2.50 / 10.70	**16.30** / 22.59
Attribute	Avg	5.55 / 11.28	9.60 / 13.56	9.20 / 13.96	6.55 / 6.94	2.80 / 9.00	**19.60** / 23.58
Position		30.90 / 48.10	34.30 / 48.20	**67.50** / 77.40	12.20 / 14.40	34.00 / 48.80	21.20 / 47.75
Relationship		10.00 / 33.20	14.50 / 31.40	10.70 / 35.30	6.10 / 7.20	8.20 / 29.40	**41.98** / 51.98
Negation		29.30 / 51.80	28.70 / 57.20	**52.50** / 67.30	27.90 / 29.70	28.30 / 41.10	35.10 / 55.86
Total Average		18.37 / 29.91	17.96 / 26.95	25.30 / 33.69	16.02 / 16.74	9.06 / 17.52	**25.86** / 39.15

Table 3: The NMS-AP and traditional AP evaluation results (%), * represents supervised score, otherwise it's zero-shot. Total average is averaged over the 9 subtasks.

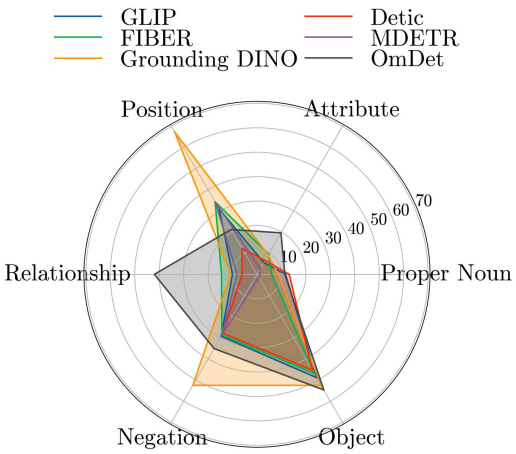


Figure 4: Radar chart of NMS-AP results on 6 aspects. Most models successfully worked on *object* but failed on others.

For datasets with hard negatives, the labels often involve some descriptions and require a more fine-grained linguistic understanding for models. We found that all models exhibit poor performance on *color* and *material* tasks. In contrast, OmDet performs more favorably overall on these tasks, largely due to its use of the VAW (Pham et al. 2021) dataset with attributes during pre-training. Meanwhile, the overall performance of existing models on the *position*, *relationship*, and *negation* tasks is similar, with generally low NMS-AP values. This indicates that the current models have limited capability in handling tasks with fine-grained descriptions. However, we note that Grounding DINO significantly outperforms the other models in *position* task. This can be attributed to its utilization of the RefCOCO dataset with orientation data during pre-training, which provides the model with specific knowledge related to the position and improves its performance on this task. Moreover, OmDet performs better than other models on the relationship task, which can be attributed to its use of the HOI-A (Liao et al. 2020) dataset with relation attributes during pre-training, providing the model with specific knowledge related to the relationship and improving its performance on this task. While minor differences exist, all models display a similar trend when we represent the 6 aspects on a radar chart (Figure 4). All models successfully worked on the common object task (*object*). However, they all failed on the hard tasks from the proposed datasets, which require the use of external/commonsense knowledge and fine-grained localization ability. Therefore, it is evident that a dataset with fine-grained labels is necessary to establish a better benchmark to provide a clear optimization direction for improving the model's performance on challenging tasks.

Comparing NMS-AP with Traditional AP

To validate the Inflated AP problem, we performed the evaluation of traditional AP on our OVDEval dataset and compared it with the NMS-AP results. Table 3 shows that the difference between NMS-AP and AP on classical OD datasets such as COCO is small, e.g., 52.50 vs. 55.30 for Grounding DINO because the probability of mutual exclusion of the predicted labels in this task is small and its impact on the AP calculation is negligible. On the other hand, the difference between NMS-AP and AP becomes much more significant for more difficult aspects including attribute, position and etc. For example, the *relationship* AP of Grounding DINO decreased from 35.30 to 10.70. The above results confirmed that our hypothesis about the Inflated AP Problem exists for the compared OVD models. To visually illustrate our hypothesis and investigate the cause of the large NMS-AP and AP difference, we have plotted several bounding boxes obtained from the GLIP predictions, as illustrated in Figure 3.

From the examples depicted in Figure 3, it is evident that GLIP tends to generate multiple bounding boxes on the same object. Notably, the labels assigned to these bounding boxes are mutually exclusive. For instance, in the case of a cat, the predicted bounding boxes include both "cat that is sitting" and "cat that is not sitting". This inconsistency matches our hypothesis about the inflated score problem by deceiving traditional AP. That is although Grounding DINO has a poor performance in understanding *negation* it can still obtain a high AP score.

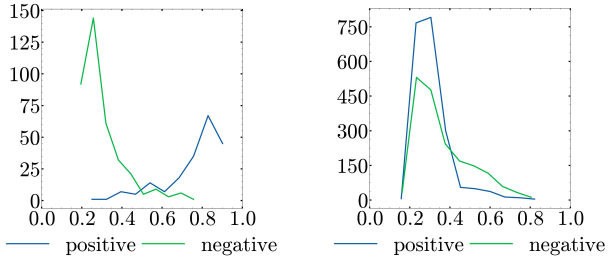
On the other hand, by employing our NMS-AP algorithm, we effectively retain only one bounding box with the highest confidence for each ground-truth instance while disregarding other false bounding boxes during the AP calculation. This approach helps mitigate the inflated AP problem caused by multiple bounding boxes. The decrease in scores that we observed earlier can be primarily attributed to models predicting the highest confidence on false labels, indicating a failure in comprehending fine-grained descriptions. Note that among all the models, Detic suffers the least from NMS-AP and AP difference because its model architecture already applies NMS internally to the region proposal network (RPN) that remove the duplicated boxes over the same region (Zhou et al. 2022).

Therefore, utilizing NMS-AP to evaluate OVD models on our benchmark provides a more suitable approach for assessing their performance on intricate linguistic descriptions. This method helps address the limitations of the models and provides a more accurate evaluation metric.

Limitations of Current OVD Models

We have also noticed a recurring issue among all the OVD models, where they tend to generate multiple bounding boxes for the same object but assign inconsistent labels to them. Moreover, these predicted labels are often mutually exclusive, and it is worth mentioning that the predictions with the highest confidence scores are frequently incorrect. This issue is particularly pronounced in models with a large number of output bounding boxes, such as Grounding Dino. This observation further strengthens our previous hypothesis that the current models demonstrate exceptional performance in learning straightforward object tasks such as COCO. However, they encounter difficulties in comprehending the intricacies of detailed descriptions.

To further support our hypothesis, we plot the distribution of predicted confidence score for the *object* and *negation* aspects. Figure 5 from GLIP illustrates the distribution of confidence scores. The distribution of *object* is obtained from the model predictions on a subset of images in the COCO validation dataset. To calculate these distributions, we tally the number of positive and negative labels from the predictions that have an IoU greater than 0.9 with the ground truth.



(a) object aspect (b) negation aspect

Figure 5: Distribution analysis of predicted confidence for object and negation aspects in GLIP. X-axis is prediction confidence and Y-axis is the number of predictions.

Based on the results in Figure 5, it is clear that in the *object* task, positive predictions tend to be spread out across the high confidence range, while negative predictions are mostly concentrated in the low confidence range. This indicates that most models have successfully learned to accurately identify objects. However, in the *negation* task, the confidence distribution of positive and negative samples exhibits a similar trend. Meanwhile, the predictions predominantly appear in the low confidence region. These findings further support our hypothesis that existing models struggle to comprehend certain nuanced semantic information in fine-grained tasks.

Conclusion

This paper presents a novel benchmark OVDEval, testing the generalization of open-vocabulary detectors. We carefully create the dataset with challenging hard negatives and annotate 20K images with human experts. We also identified the Inflated AP problem for conventional AP calculation and introduce a new metric NMS-AP to deal with it. Our assessment validates the OVDEval's effectiveness in revealing the pros and cons of current SOTA open-vocabulary models. Lastly, OVDEval provides promising future research questions. How can we incorporate better training objectives so OVD models can acquire better discriminate abilities against hard negatives in both visual and linguistic input? What are the better pre-training data to inject more common sense knowledge in vision-language alignment? In summary, solving OVDEval is an important step for future general-purpose object detectors.

Acknowledgements

This research is supported by National Key R&D Program of China under grant (2022YFF0902600) and Key R&D Program of Zhejiang under grant (2023C01048). Y.Y. Yao and Q. Wang are supported by NSFC under grant 62031023.

References

Cai, Z.; Kwon, G.; Ravichandran, A.; Bas, E.; Tu, Z.; Bhotika, R.; and Soatto, S. 2022. X-detr: A versatile architecture for instance-wise vision-language tasks. In *European Conference on Computer Vision*, 290–308. Springer.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, 381–389. IEEE.

Dou, Z.-Y.; Kamath, A.; Gan, Z.; Zhang, P.; Wang, J.; Li, L.; Liu, Z.; Liu, C.; LeCun, Y.; Peng, N.; et al. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35: 32942–32956.

Du, X.; Legastelois, B.; Ganesh, B.; Rajan, A.; Chockler, H.; Belle, V.; Anderson, S.; and Ramamoorthy, S. 2022. Vision checklist: Towards testable error analysis of image models to help system designers interrogate model capabilities. *arXiv preprint arXiv:2201.11674*.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Hoiem, D.; Divvala, S. K.; and Hays, J. H. 2009. Pascal VOC 2008 challenge. *World Literature Today*, 24(1).

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Yang, J.; and Liu, Z. 2023. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Li, C.; Liu, H.; Li, L. H.; Zhang, P.; Aneja, J.; Yang, J.; Jin, P.; Hu, H.; Liu, Z.; Lee, Y. J.; and Gao, J. 2022a. ELEVATER: A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models. *Neural Information Processing Systems*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 121–137. Springer.

Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; and Feng, J. 2020. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 482–490.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2021. Learning To Predict Visual Attributes in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13018–13028.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shah, D.; Osiński, B.; Levine, S.; et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, 492–504. PMLR.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Wang, J.; Zhang, P.; Chu, T.; Cao, Y.; Zhou, Y.; Wu, T.; Wang, B.; He, C.; and Lin, D. 2023. V3det: Vast vocabulary visual detection dataset. *arXiv preprint arXiv:2304.03752*.

Wu, C.; Lin, Z.; Cohen, S.; Bui, T.; and Maji, S. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10216–10225.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85. Springer.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhao, T.; Liu, P.; Lu, X.; and Lee, K. 2022a. Omdet: Language-aware object detection with large-scale vision-language multi-dataset pre-training. *arXiv preprint arXiv:2209.05946*.

Zhao, T.; Zhang, T.; Zhu, M.; Shen, H.; Lee, K.; Lu, X.; and Yin, J. 2022b. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368. Springer.

Zong, Z.; Song, G.; and Liu, Y. 2023. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6748–6758.