The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding

Lorenzo Bianchi[1,2], **Fabio Carrara**[1], **Nicola Messina**[1], **Claudio Gennaro**[1], **Fabrizio Falchi**[1]
[1]CNR-ISTI, Pisa, Italy [2]University of Pisa, Italy
`<name.surname>@isti.cnr.it`

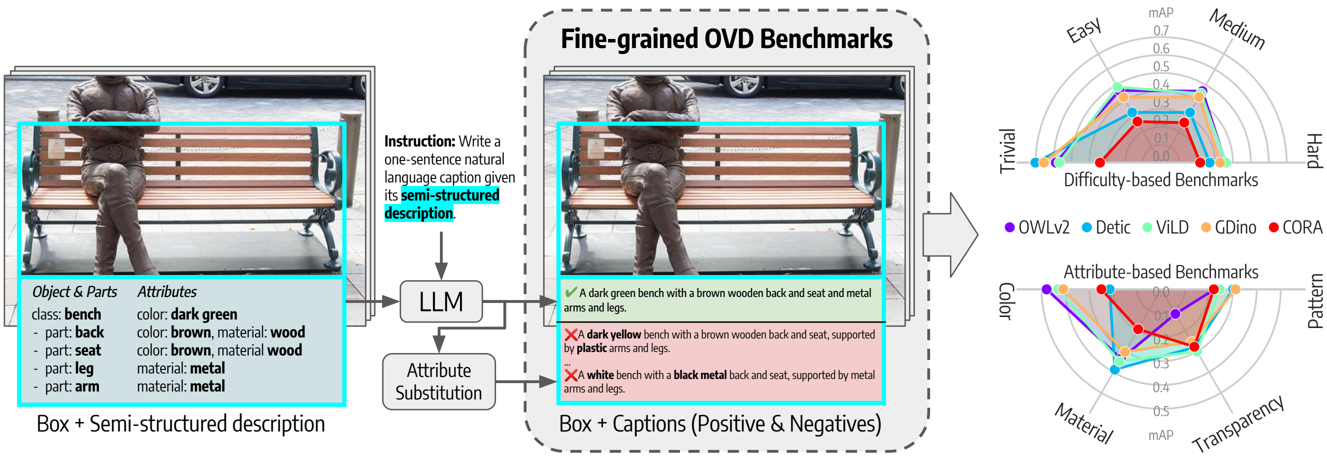

Figure 1. We propose a benchmark suite to evaluate **fine-grained open-vocabulary detection (FG-OVD)**. We build several sets of dynamic object-specific vocabularies, comprised of one positive and several negative captions, to probe the ability of open-vocabulary detectors to discern detailed object properties, like color, pattern, or material. We craft positive captions from semi-structured descriptions of objects and their parts employing a Large Language Model (**LLM**), while negative captions of different difficulty levels are built via attribute substitution. By manipulating negative sets according to their difficulty levels or the types of attributes altered — categorized as **Difficulty-based** and **Attribute-based** benchmarks — we acquire a nuanced comprehension of each detector's capabilities across various scenarios.

Abstract

Recent advancements in large vision-language models enabled visual object detection in open-vocabulary scenarios, where object classes are defined in free-text formats during inference. In this paper, we aim to probe the state-of-the-art methods for open-vocabulary object detection to determine to what extent they understand fine-grained properties of objects and their parts. To this end, we introduce an evaluation protocol based on dynamic vocabulary generation to test whether models detect, discern, and assign the correct fine-grained description to objects in the presence of hard-negative classes. We contribute with a benchmark suite of increasing difficulty and probing different properties like color, pattern, and material. We further enhance our investigation by evaluating several state-of-the-art open-vocabulary object detectors using the proposed protocol and find that most existing solutions, which shine in standard open-vocabulary benchmarks, struggle to accurately capture and distinguish finer object details. We conclude the paper by highlighting the limitations of current methodologies and exploring promising research directions to overcome the discovered drawbacks. Data and code are available at https://lorebianchi98.github.io/FG-OVD/ .

1. Introduction

Open-vocabulary object detection (OVD) consists of recognizing objects not present at training time, therefore solving the limitations imposed by traditional detectors that could only recognize a fixed pool of object classes. In the last years, open-vocabulary detectors have captured large research attention thanks to their wide flexibility in many potential downstream applications like autonomous driving [20], extended reality [19], and robotics [9, 10, 16]. The core idea behind open-vocabulary detectors is to establish a semantic connection between object regions and object labels chosen from a possibly very large vocabulary by (i) employing some focused prompting on the object label to transform it into a natural sentence and (ii) relying on vision-language interaction — possibly employing large pre-trained matching methods like CLIP [29] — for associating the most relevant object label with the specific image region.

Given the vision-language nature of open-vocabulary detectors, the elements in the labels' vocabulary are no longer restricted to simple categorical values. Instead, labels can be seamlessly substituted with any natural language sentence, ranging from compound objects possibly containing modifying attributes — i.e., for distinguishing a *brown bear* from a *polar bear* — to very complex sentences describing extrinsic object properties like colors, patterns, or materials. Therefore, we may ideally pretend that a pre-trained open-vocabulary detector is also able — to some extent — to discern a *dark brown wooden lamp* from a *gray metal lamp*. This is what we call **F**ine-**G**rained[1] **O**pen-**V**ocabulary object **D**etection (**FG-OVD**).

Apart from some marginal attempts [3, 30], no work has deeply investigated the ability of open-vocabulary detectors to discern fine-grained object properties. Some emerging downstream tasks start to require this ability, like episodic memory tasks in egocentric videos[2], where it is asked to retrieve object *instances* having particular attributes.

However, it is difficult to evaluate the ability of an open-vocabulary detector to discern specific object properties. As of now, most works evaluate the effectiveness of open-vocabulary detectors using established benchmarks like COCO and LVIS, which are already widely used to train and test traditional object detectors. Given their closed-vocabulary nature, these benchmarks primarily focus on very generic class labels and do not explore the capabilities of these detectors when the input text is more elaborate and includes fine-grained characteristics of the object.

In this paper, we propose a novel evaluation protocol and a benchmark suite for measuring the discriminative power of open-vocabulary detectors against fine-grained object descriptions. We build a set of object detection benchmarks that provide each object with a rich and intricate caption — generated by a Large Language Model (LLM) fed with a structured description of objects in the scene — that encapsulates its complex extrinsic characteristics like color, material, or pattern. For each positive caption, we also include a series of incorrect variants obtained through slight attribute modifications, acting as negative labels.

By carefully varying the number and types of attributes in the negative examples for each object, our benchmarks, together with the proposed evaluation protocol, enable us to analyze the weaknesses of the most recent open-vocabulary detectors across different analytical dimensions, identifying areas for improvement and providing valuable directions for advancing this extremely relevant field.

In summary, our key contributions are as follows:

- We introduce the novel challenging task of Fine-grained Open Vocabulary Detection (FG-OVD).
- We propose a novel evaluation protocol for FG-OVD, complimented by some relevant metrics, for quantitatively assessing the fine-grained discriminative power of open-vocabulary detectors.
- We introduce a novel set of benchmarks specifically crafted to evaluate the ability of open-vocabulary detectors to discern extrinsic object properties.
- We perform extensive experimentation, testing some state-of-the-art pre-trained open-vocabulary detectors and demonstrating that even the most recent ones struggle to distinguish fine-grained object properties.

2. Related work

2.1. Zero-Shot Open-vocabulary Object detection

Zero-shot (ZS) generically refers to scenarios where the model, at inference time, is not exposed to specific object classes seen during its training process. In the context of object detection, we follow the one given by [1], where ZS refers to never seeing, during training, even a single annotated bounding box of the class of interest at test time. [3]

Early approaches in zero-shot object detection, such as Bansal et al. [2], proposed replacing the last classification layer with language embeddings, such as GloVe [26], representing class names. However, recent advancements in large-scale image-text encoder models — such as CLIP [29] and ALIGN [11], trained on millions of image-text pairs — enabled a strong semantic interaction between vision and language. Their cross-modal alignment capabilities have been used in a plethora of recent open-vocabulary detectors by either substituting the final class features with their text embeddings [5, 14, 22, 23, 37], by learning better ROI-Align head [8] using knowledge from pre-trained vision-language backbones [4, 6, 14, 23, 36], or by directly repurposing the CLIP model itself to work as an open-vocabulary detector [22, 23, 35].

Some recent works [12, 15, 18, 34] started to unify open-vocabulary detection with the Referring Expression Comprehension (REC) [21, 33] and Phrase Grounding (PG) [25, 28] tasks, already well-known in literature. While both REC and PG are given a single possibly complex sentence,

[1] Rather than closed-set fine-grained categorization [31] focusing on small inter-class variation (e.g., *golden retriever* vs *labrador retriever*), we focus on intra-class extrinsic attributes (e.g., *brown dog* vs *black dog*).

[2] https://github.com/EGO4D/episodic-memory

[3] In our case, we deal with free-form sentences as labels and not classical categorical ones, which makes all the possible phrasings of every combination of attributes almost unique. Therefore, we deem the proposed task lies in the ZS regime since we have as labels, in the test set, a set of sentences comprising attributes unlikely seen during training.

REC's objective is to locate the single correct object in the image, while PG requires locating all the entities appearing in the text. Despite the similarities with the FG-OVD task and their higher need for associating articulated phrases to image regions, there are some key differences: (i) REC and PG tasks assume that the given unambiguous sentence is surely grounded somewhere in the image, making the development of advanced discriminative abilities almost unnecessary; (ii) networks trained to solve these tasks are not evaluated on their discriminative abilities when posed with difficult choices. Therefore, although some REC or PG networks can also work as open-vocabulary detectors — like the recent GroundingDino [18] or GLIP [15, 34] — they tend to inherit the same drawbacks when they are required to distinguish fine-grained object characteristics.

2.2. Open-Vocabulary Detection Benchmarks

The COCO [17] and LVIS [7] datasets are widely recognized as benchmarks for evaluating the localization and classification capabilities of object detectors. The COCO dataset, originally largely employed for evaluating standard closed-set detectors, has been repurposed to address zero-shot detection [2] and open-vocabulary detection [6, 18, 35]. In this setup, it is composed of a vocabulary of 48 base categories for training and 17 novel categories for testing. On the other hand, LVIS contains a large and diverse collection of object categories grouped by frequency of appearance (common, frequent, and rare). Many works [6, 14, 18, 35] used the frequent and common categories as base categories and rare categories as novel categories for testing. While these benchmarks evaluate the model's ability to perform zero-shot open-vocabulary detection, they do not explicitly evaluate the model's ability to recognize specific object characteristics.

The datasets that most closely align with our objectives are OVAD [3] and VAW [27]. Despite focusing on object attributes and proposing to use negatives to challenge existing detectors, there are major differences with our work: (i) they mainly benchmark attribute detectors, which usually have a separate head to specifically infer attributes besides the object classes; (ii) they only need structured annotations and not natural language sentences describing each object, limiting the evaluation of current state-of-the-art vision-language foundation models; (iii) they do not craft challenging negative examples, limiting the analysis of the current detector's limitations. The dataset from which we drew inspiration is PACO [30]. PACO is built upon COCO and comes with object bounding boxes annotated with a structured JSON-like representation that carries information about attributes and parts information of the object. PACO carries extrinsic and categorized object properties, like object colors, materials, and patterns. Our benchmarks suite is crafted from PACO, by creating fine-grained sentences from structured object descriptions using LLMs, empowering the growing trend of using LLMs to create targeted, high-quality, and diverse annotations [24].

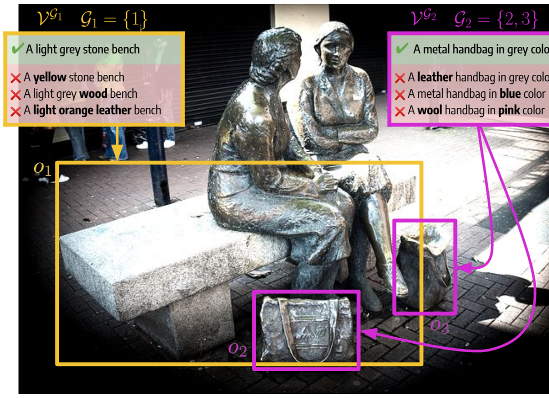

Figure 2. **Examples of Dynamic Vocabularies:** the image I features two distinct object groups $\mathcal{G}_1$ and $\mathcal{G}_2$, each one associated with a set of captions. The positive captions c^{pos} (marked with ✓) – *A light grey stone bench* and *A metal handbag in grey color* –, are juxtaposed with three negative captions c^{neg} (indicated by ✗). These positive and negative captions collectively form two vocabularies, namely $\mathcal{V}^{\mathcal{G}_1}$ assigned to o_1, and $\mathcal{V}^{\mathcal{G}_2}$ assigned to o_2 and o_3. The open-vocabulary detector is then applied to I two times, once for each vocabulary: $\psi(I, \mathcal{V}^{\mathcal{G}_1})$ and $\psi(I, \mathcal{V}^{\mathcal{G}_2})$.

3. Methodology

We propose an evaluation protocol to probe the capabilities of open-vocabulary object detectors to discern fine-grained characteristics of objects to be detected. In subsection 3.1, we formulate the problem and propose a suitable evaluation protocol. Then, in subsection 3.2, we describe the novel suite of benchmarks aimed at challenging open-vocabulary detectors to distinguish fine-grained object labels.

3.1. Evaluation Protocol

OVD Formalization. Let $I \in \mathbb{R}^{W \times H \times C}$ an image and $\mathcal{V} = \{c_j\}_{j=1}^{T}$ a vocabulary composed by T arbitrary sentences describing the objects we want to detect. An open-vocabulary object detector ψ takes in input I and $\mathcal{V}$ and produces a set of predictions $\mathcal{D} = \psi(I, \mathcal{V}) = \{d_i\}_{i=1}^{m}$, where each prediction $d_i = (\mathbf{b}_i, \mathbf{s}_i)$ is comprised by a bounding box $\mathbf{b}_i \in \mathbb{R}^4$, and the scores $\mathbf{s}_i \in \mathbb{R}^T$, where each element $s_{i,j}$ is the confidence assigned to vocabulary object c_j [4]. The highest-score caption in $\mathcal{V}$ is selected as the predicted one. In classic zero-shot and open-vocabulary evaluation settings, the vocabulary $\mathcal{V}$ is fixed for all images in the test dataset, and for each image, predictions $\mathcal{D}$ are compared with the ground truth object instances $\mathcal{O} = \{o_i\}_{i=1}^{m'}$, $o_i =$

[4]Usually, captions are obtained by prompting a vocabulary of class labels (e.g., *dog* → *a photo of a dog*).

$(\mathbf{b'}_i, l'_i) \in \mathbb{R}^4 \times \mathcal{V}$ and evaluated by means of standard detection metrics, e.g., mAP, mAR, etc.

Dynamic Vocabularies for FG-OVD. For evaluating FG-OVD we suggest utilizing a carefully crafted dictionary $\mathcal{V}_i$ for each ground truth object o_i. This approach allows us to have finer control over the selection of negatives to be used for each object. More formally, each object o_i is given a detailed visual description denoted as c_i^{pos} (referred to as the positive caption) that accurately describes an object. The object o_i is also associated with several other descriptions $c_{i,1}^{\text{neg}}, \dots, c_{i,N}^{\text{neg}}$ (referred to as negative captions) that semantically differ — up to varying degrees — from c_i^{pos}. In this way, we define a new ground truth composed of objects $\tilde{o}_i = (\mathbf{b'}_i, l'_i, \mathcal{V}_i)$, where $\mathcal{V}_i$ is a per-object vocabulary filled with $\{c_i^{\text{pos}}, c_{i,1}^{\text{neg}}, \dots, c_{i,N}^{\text{neg}}\}$, and the label $l'_i = c_i^{\text{pos}}$. The detector draws the predicted label from the vocabulary $\mathcal{V}_i$ during inference. From a practical perspective, notice that if we perform inference for each object o_i over its vocabulary $\mathcal{V}_i$ inside the same image I, we may produce duplicated outputs in the case where there are two objects o_i and o_j both correctly described by the same c^{pos}. In order to solve this problem, we perform a single inference for each one of the K set of objects $o_{\mathcal{G}_k}$ in the image I sharing the same positive caption c^{pos}, where $\mathcal{G}_k$ is the k-th set of object indexes satisfying this property. Therefore, we perform K separate inferences for each set $\mathcal{G}_k$ by performing the forward passes $\{\psi(I, \mathcal{V}^{\mathcal{G}_k})\}_{k=1}^{K}$, where $\mathcal{V}^{\mathcal{G}_k}$ is the vocabulary associated with the objects $o_{\mathcal{G}_k}$. We report an example of our ground truth arrangement in Figure 2.

Post-processing. Class-aware non-maximum suppression (NMS) is typically employed in object detection post-processing to discard near-duplicate predictions of the same class insisting on the same location, while different-class predictions of the same object are left untouched. In our formulation, classes in the vocabulary are mutually exclusive, and we are interested in evaluating only the most confident prediction per object independently from the class label. Traditional implementations of detection metrics, such as COCO mAP, do not penalize incorrect predictions of non-occurring classes, such as the negative ones in our vocabularies, i.e., the presence of an incorrect higher-confidence prediction in the same location as a correct prediction is not penalized. To circumvent this issue, we apply *class-agnostic* NMS to ensure only one prediction per location is present, regardless of class label. This guarantees that an incorrect higher-confidence prediction suppresses the lower-confidence correct prediction, correctly bringing the mAP metric to work as desired in our evaluation setup.

Metrics. The post-processed predictions are rigorously evaluated using the COCO mean Average Precision (mAP) metric, which provides a comprehensive assessment of object localization accuracy and caption assignment correctness across varying detection confidence levels.

Alongside mAP, we incorporate the Median Rank metric. Given an object $o_i = (\mathbf{b'}_i, l'_i)$ and its corresponding vocabulary $\mathcal{V}_i$, let $d_j^i = (\mathbf{b}_j, \mathbf{s}_j)$ the prediction satisfying the condition of Intersection over Union (IoU) ≥ 0.5 with o_i [5]. We sort the confidence scores of each element of the vocabulary $\mathbf{s}_j$ in descending order and record the position of the score corresponding to the correct caption l'_i in the ranked list. Then, we report the median rank over all objects in the benchmark.

Unlike mAP, which only considers the maximally activated label, the median rank helps to better quantify the confidence of each detector in predicting the correct label among the other choices available in the dictionary.

3.2. Dataset

Evaluating open-vocabulary object detectors through the aforementioned evaluation protocol requires a carefully crafted benchmark, carrying different vocabularies for each object. A benchmark implementing this novel evaluation protocol should pay particular attention to the quality and variability of negative examples in each object's vocabulary. To this aim, we introduce an extension to the PACO dataset [30], which comprises an innovative suite of benchmarks meticulously crafted to tackle the task of FG-OVD. Our benchmark suite offers a comprehensive evaluation through eight distinct scenarios, categorized into **Difficulty-based** and **Attribute-based** benchmarks. Difficulty-based benchmarks enable the assessment of detector performance across different difficulty levels obtained by varying the hardness of negative captions. On the other hand, Attribute-based benchmarks enable the precise selection of attribute types to facilitate the evaluation of detectors' capabilities in recognizing specific attributes. For each of the four attribute types, we mostly inherit the possible values from the PACO dataset, for a total of *29 colors*, *14 materials*, *8 patterns*, and *3 transparencies* modes (more details in the supplementary material). In the following paragraphs, we dive into the construction details of such benchmarks.

Positive Caption Generation. In our quest to describe objects with precision, we assume each object is characterized by at least one attribute and may possess one or more parts, where each part may have its own attributes. As we require text entries in open-vocabulary detectors, we exploit this structured definition of each object to generate a fine-grained textual description carrying attributes and parts details. To achieve this, we harnessed the capabilities of a

[5] In rare edge cases, we might have multiple predictions insisting on the same ground truth box not removed by NMS. In those cases, we select the prediction with the maximum highest confidence.

Table 1. Statistics of the benchmarks for each different negative set comprising the number of images (Imgs), the number of annotated objects (Objs), objects-to-image ratio (Objs/Img), positive captions, positive captions per image, negative captions per positive caption, and objects per positive caption.

Name	Negative Set Strategy	Imgs	Objs	Obj/Img	✓Caps	✓/Img	✗/✓	Objs/✓
Hard	Random attribute subst. $(\times 1)$	1707	3545	2.1	2349	1.4	9.9	1.5
Normal	Random attribute subst. $(\times 2)$	1537	2968	1.9	2034	1.3	10.0	1.5
Easy	Random attribute subst. $(\times 3)$	853	1299	1.5	971	1.1	10.0	1.3
Trivial	Random captions	1707	3545	2.1	2349	1.4	9.9	1.5
Color	Color attribute subst.	1599	3119	2.0	2126	1.3	10.0	1.5
Material	Material attribute subst.	1577	3193	2.0	2128	1.3	10.0	1.5
Pattern	Pattern attribute subst.	321	467	1.5	337	1.0	7.4	1.4
Transparency	Transparency attribute subst.	230	409	1.8	238	1.0	2.2	1.7

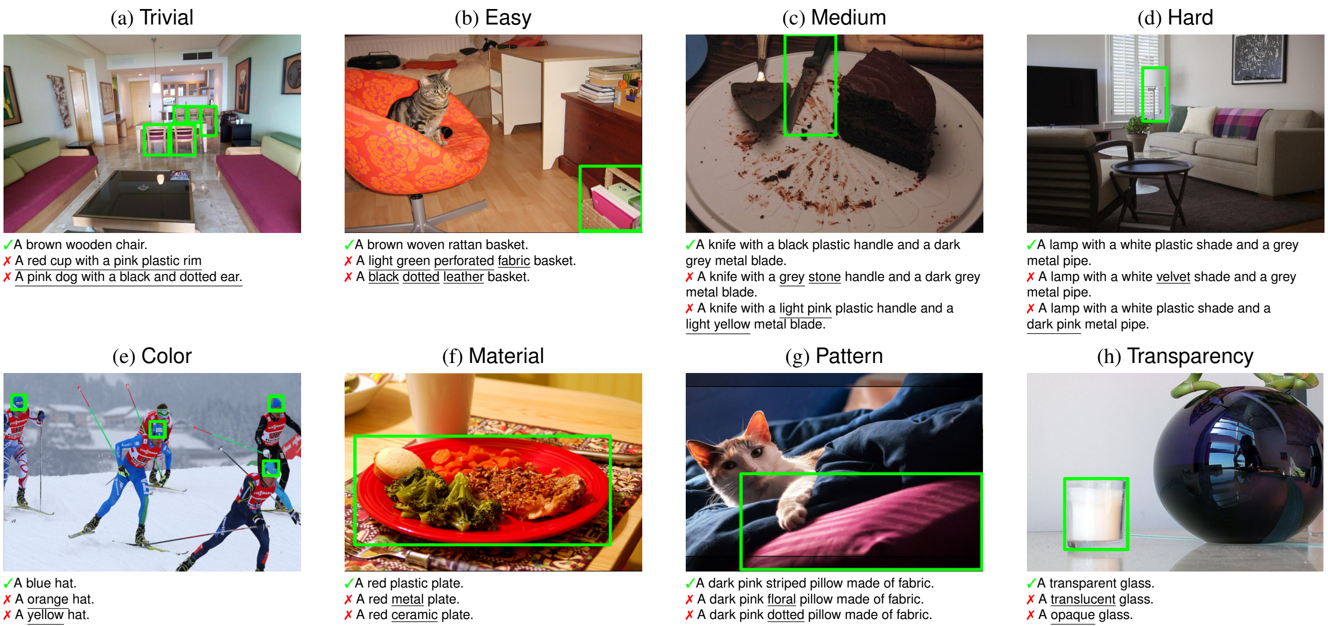

Figure 3. **Benchmarks Examples**: each benchmark tests different properties by crafting negative captions via attribute substitution.

Large Language Model (LLM), by prompting it with structured object descriptions from PACO – as shown in Figure 1 – and forcing it to mimic some carefully crafted natural language examples. We then performed a quick manual check to ensure the required consistency. Notice that PACO already provides straightforward natural language captions for a limited number of objects. We still preferred the ones generated through an LLM, considering the beneficial quality-to-quantity ratio compared with the already available ones. Detailed comparisons with the original PACO captions are given in supplementary material. We decided to employ OpenAssistant-LLAMA-30B [13] as an LLM, as it provides finer control over inference and better reproducibility guarantees over other available LLMs.

Negative Captions Generation. To thoroughly assess open-vocabulary object detection models, we introduce diverse benchmarks featuring vocabularies with challenging negative captions derived from the generated captions. We aim to create negative captions that are semantically different from the positive ones while maintaining structural similarities. We do so through attribute substitution, which involves replacing attributes in the positive captions text to preserve syntactic consistency. We preferred this approach to altering the object's structured description and re-querying the LLM, which could introduce significant syntactic variations and unwanted hallucinations.

The choice of negative captions aims to assess the model's adaptability to a range of attribute variations, covering changes in attribute types and quantities. This allows for a comprehensive evaluation of its resilience across diverse negative scenarios. The benchmarks fall into two classes: **(i) Difficulty-based** (Trivial, Easy, Medium, Hard): These benchmarks explore open-vocabulary detectors' capabilities in ascending difficulty scenarios. Starting with the Trivial benchmark, where negative captions are randomly sampled from other objects, we progress through Easy, Medium, and Hard benchmarks, in which negative captions are generated by randomly replacing 3, 2, and 1 attributes, respectively. As the number of attributes replaced decreases, the distinctions between captions become less pronounced and the task more challenging (see Figure 3, top row). **(ii) Attribute-based** (Color, Material, Transparency, Pattern): This class of benchmarks delves into open-vocabulary detectors' capabilities in recognizing specific types of attributes. By precisely replacing only one attribute of a specific type in each positive caption (depending on the specific category of the benchmark), we aim to evaluate the detector's proficiency in recognizing that particular attribute type (see Figure 3, bottom row). Table 1 reports statistics of each described benchmark.

4. Experiments

4.1. Evaluated Models

In our experiments, we evaluated the performance of the following state-of-the-art open-vocabulary detectors.

ViLD [6], **Detic** [37], and **CORA** [32] underwent evaluation using our standard protocol outlined in subsection 3.1 for all the images within the FG-OVD benchmarks. Specifically, for every object group i in the image I, we perform an inference $\psi(I, \mathcal{V}^{\mathcal{G}_i})$.

OWL [22] and **OWLv2** [23] adhered to the same standard protocol. However, due to the positional embedding layer of these models being trained with a limit of 16 tokens, evaluation was conducted exclusively on captions respecting this maximum token constraint (about 80% of the available ones). We then checked that applying this same inference constraint to all the detectors did not consistently affect the overall evaluation. We report the details of this sanity check in the supplementary material.

GroundingDino [18], is mainly a REC model and thus, in contrast to other detectors, cannot accept a vocabulary of captions as input. The authors claim that specific prompts enable the given input textual expression to be automatically split – e.g., after each "." character. However, we found that this works only when single words are given as labels, while complex sentences are sometimes erroneously split in their middle. Therefore, we turned GroundingDino into an open-vocabulary detector by making a forward pass for each of the captions in the vocabulary $\mathcal{V}_i$ and merging the results before evaluation. More details about the specific inference process are given in the supplementary material.

	FG-OVD				LVIS
Detector	Hard	Medium	Easy	Trivial	Rare
OWL (B/16)	26.2	39.8	38.4	53.9	20.6
OWL (L/14)	**26.5**	39.3	**44.0**	65.1	31.2
OWLv2 (B/16)	25.3	38.5	40.0	52.9	29.6
OWLv2 (L/14)	25.4	**41.2**	42.8	63.2	34.9
Detic	11.5	18.6	18.6	**69.7**	**39.9**
ViLD	22.1	36.1	39.9	56.6	16.8
GDino	16.6	27.9	30.1	62.7	18.1
CORA	13.8	20.0	20.4	35.1	22.2

Table 2. mAP on Difficulty-based benchmarks ($N = 5$) and on rare categories of the standard LVIS benchmark (The mAP values for LVIS rare are taken from the original papers or GitHub repos).

	Color	Material	Pattern	Transp.
OWL (B/16)	45.3	37.3	26.6	**34.1**
OWL (L/14)	43.8	**44.9**	**36.0**	29.2
OWLv2 (B/16)	45.1	33.5	19.2	28.5
OWLv2 (L/14)	**53.3**	36.9	23.3	12.2
Detic	21.5	38.8	30.1	28.0
ViLD	43.2	34.9	24.5	30.1
GDino	41.0	30.2	31.2	25.4
CORA	25.0	19.3	22.0	27.9

Table 3. mAP on Attribute-based benchmarks ($N = 2$).

4.2. Results

Performance vs Negative Difficulty. Table 2 shows the mAP of the tested state-of-the-art detectors on our difficulty-based benchmarks alongside reference results from existing benchmarks for standard open-vocabulary detection. We set the number of negative captions in the dynamic vocabularies $N = 5$ for our benchmarks. In general, all detectors have acceptable performance in localizing and recognizing objects from detailed descriptions in the absence of confounding labels (Trivial), but the discriminative power in hard-negative settings drops drastically. For example, Detic is the best-performing detector on Trivial while the worst one in the Hard setting. Note also that performance in standard benchmarks does not positively correlate with performance in our hard-negative tasks. In this context, Detic is the detector performing better on LVIS while being the worst in the Hard scenario. Generally, the best mAP in Hard is obtained by methods like OWL and ViLD that carefully embed image-language contrastively learned features into the detector heads. If properly managed inside the architecture, this information is crucial in giving the object detection heads discriminative abilities. Instead, Detic bases its strength on training with large image-level

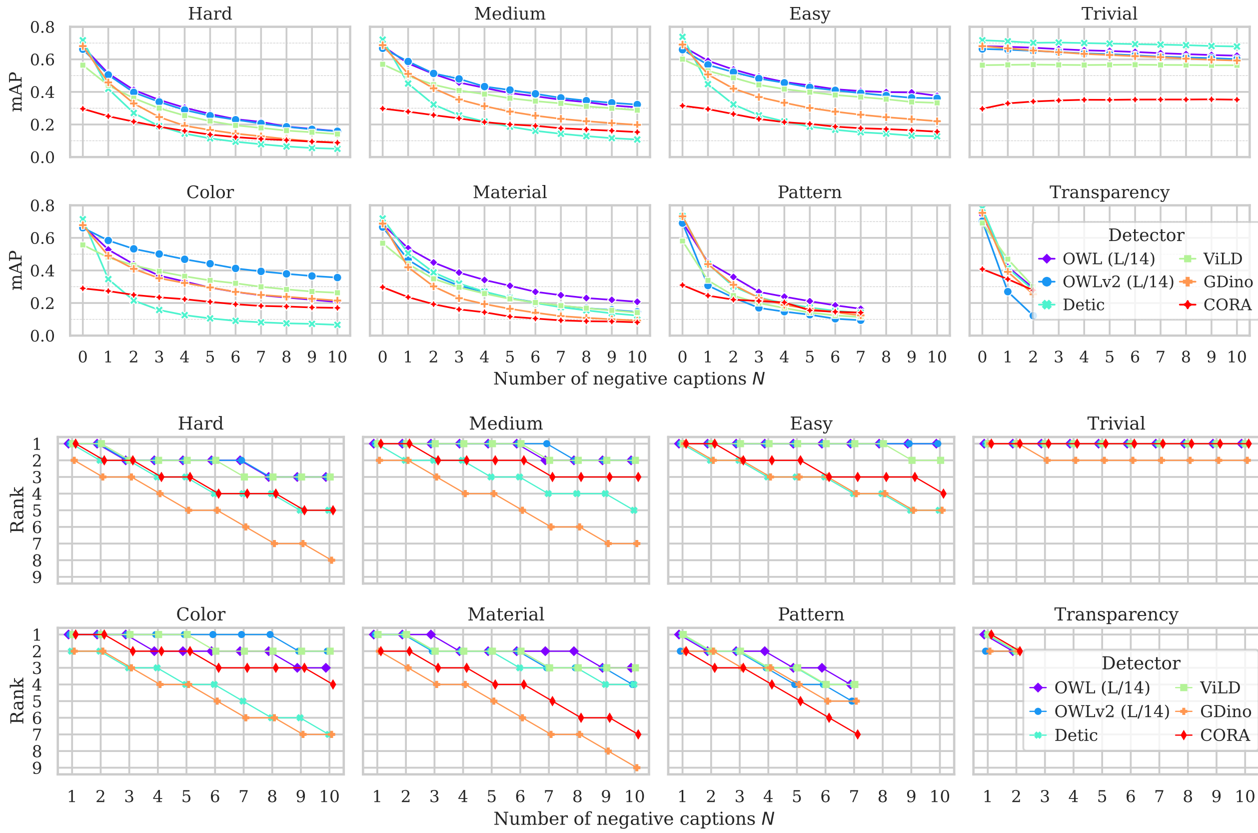

Figure 4. **Effect of the number of negative captions.** For each one of the eight proposed benchmarks, we report the mAP (rows 1-2) and the Rank (rows 3-4) varying the number N of negative captions for the different probed detectors. Notice that for Pattern and Transparency, we have a limited number of possible negatives (7 and 2, respectively).

datasets, which add strong class-wise discriminative skills while largely sacrificing fine-grained attribute recognition abilities. This conclusion also seems motivated by the almost zero gain that OWLv2 – trained in a self-supervised manner on web-scale data (10B images) – has with respect to OWL. The training approach of OWLv2 possibly adds visual robustness to the already known classes but fails to inject proper discriminative attribute information despite the massive amount of data employed. Two qualitative examples in the first row of Figure 5 show how the scores entropy from the different detectors drastically increases when captions in the vocabulary start to contain very hard negatives.

Performance vs Attribute Type. In Table 3, we show the performance of each tested detector in discerning particular attribute types, by crafting hard negatives where only a particular attribute is changed. Note that we simplified the task for these results by adding only $N = 2$ hard-negative captions to the vocabulary so that mAP values across different attributes are comparable[6]. We can notice that color is the easiest feature for most detectors, with OWLv2 reaching the best mAP. Color, in fact, is probably the attribute more present in web-scale image-text pairs (image and corresponding alt-text) used to train CLIP-like contrastive image-language models and to produce the N-grams vocabulary entries employed during the self-supervised training stage of OWLv2. On the other hand, it is possible that transparency and pattern properties are hardly found in image labels or alt-text captions collected in currently employed datasets. Despite good architectures and clever pre-training strategies, these attributes also undermine very recent open-vocabulary detectors like CORA. Two qualitative examples reported in the second row of Figure 5 show the sensitivity of different detectors to different attribute types.

Performance vs Vocabulary Size. Figure 4 reports mAP and median rank for increasing the number of negative captions N in the vocabulary up to 10 for each benchmark in the proposed suite. For Pattern and Transparency, the maximum number of negative captions N is set to the number of available values for that attribute. We note that, as expected, the detectors' performances degrade with increas-

[6]Different attributes imply different number of negatives, but N=2 is always guaranteed by our benchmark.

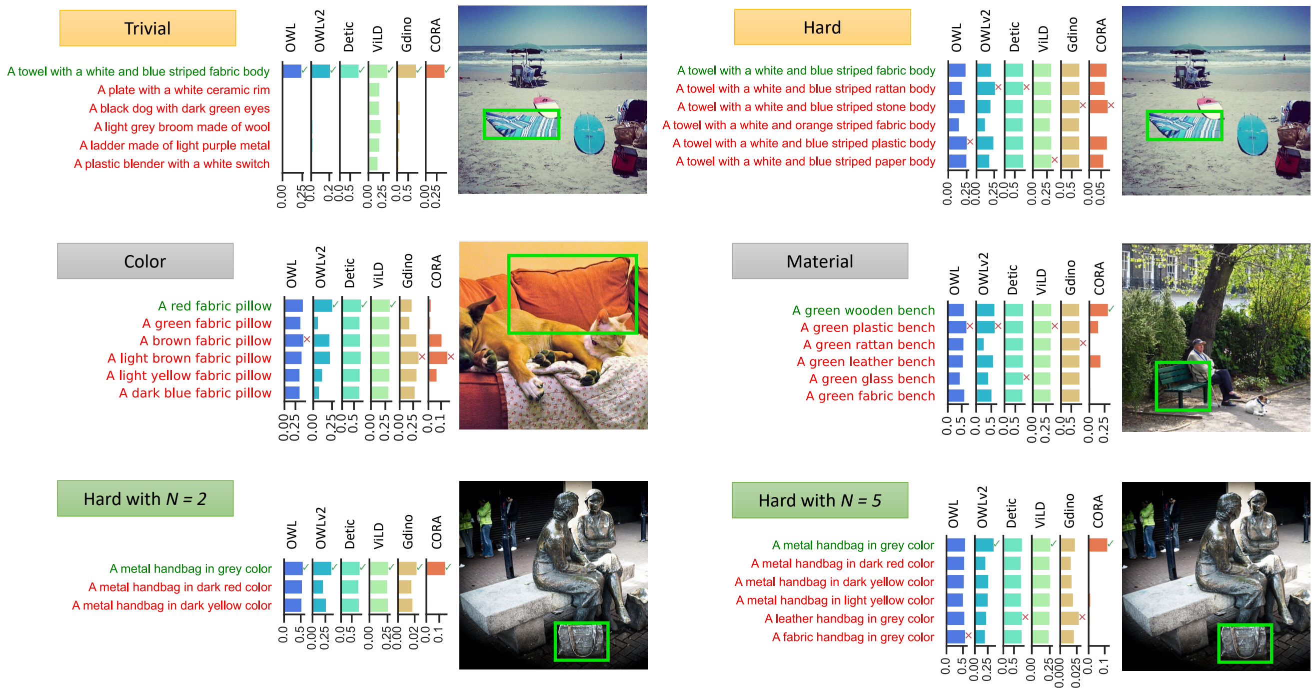

Figure 5. **Output scores from the probed detectors**. We report examples of how detectors score vocabulary entries for a specific object. The first green caption is positive, while the other red ones are the negatives. **First row: Hard vs Trivial** – we show the difference in score distributions when detectors are challenged with Hard (left) or Trivial (right). **Second row: Attributes** – we show the behavior when changing specific attributes, like color (left) or material (right). **Third row: Varying the number of negatives** – we show how increasing the number of negatives (left: $N = 2$, right: $N = 5$) strongly challenges the fine-grained discriminative abilities of many detectors.

ing N values, with different decreasing rates. Detic seems to have the steepest performance degradation, both in terms of mAP and Rank, as the number of negatives increases. Differently, despite an initial physiological decline, OWLv2 and ViLD seem to be the most robust on the Hard setting for the Color and Material attributes, despite remaining among the worst on Pattern and Transparency – probably due to the absence of these attributes in their training and pre-training data. Notably, CORA seems challenged even for a small number of negatives, evidencing the trend that recent open-vocabulary detectors still suffer from major limitations when asked to distinguish fine-grained labels. Two qualitative examples reported in the third row of Figure 5 show how a higher number of negatives increases the scores distribution entropy, bringing higher misclassification rates.

5. Conclusions

In this paper, we explored Fine-grained Open Vocabulary Detection (FG-OVD), by presenting a comprehensive evaluation protocol and benchmarks suite designed to scrutinize the fine-grained discriminative power of open-vocabulary detectors. The presented evaluation protocol, accompanied by meaningful metrics, challenges these models with rich captions encapsulating complex extrinsic characteristics.

To implement the presented protocol, we prepared an ad-hoc benchmark, starting from structured object descriptions and employing an LLM to generate diverse and high-quality captions. By slightly changing one or more attributes in the generated caption, we produced a spectrum of difficulty levels, which enabled us to systematically analyze the weaknesses of recent open-vocabulary detectors across various analytical dimensions. Our experiments revealed a notable gap in the detectors' ability to effectively capture and distinguish fine-grained object properties, with the most recent ones often performing the worst.

In the near future, we plan to fine-tune existing open-vocabulary detectors in a few-shot contrastive manner, tackling downstream tasks like episodic memory retrieval in egocentric datasets. Furthermore, we aim to exploit the proposed benchmarks to study latent information about fine-grained object attributes learned by diffusion models.

Acknowledgments This work was partially supported by the following projects: SUN – Social and hUman ceNtered XR (101092612), NextGenerationEU (FAIR PE00000013, ITSERR B53C22001770006, MUCES B53D23026090001, and EKEEL P20227PEPK).

References

[1] Relja Arandjelović, Alex Andonian, Arthur Mensch, Olivier J Hénaff, Jean-Baptiste Alayrac, and Andrew Zisserman. Three ways to improve feature alignment for open vocabulary detection. *arXiv preprint arXiv:2303.13518*, 2023. 2

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, pages 384–400, 2018. 2, 3

[3] Maria A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *CVPR*, pages 7041–7050, 2023. 2, 3

[4] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pages 14084–14093, 2022. 2

[5] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, pages 701–717. Springer, 2022. 2

[6] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2, 3, 6

[7] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 3

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017. 2

[9] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3d value maps for robotic manipulation with language models. In *Proceedings of The 7th Conference on Robot Learning*, pages 540–562. PMLR, 2023. 1

[10] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd. Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, K. Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. 1

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 2

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 2

[13] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. OpenAssistant Conversations - democratizing large language model alignment. In *NeurIPS*, 2023. 5

[14] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2023. 2, 3

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 2, 3

[16] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023. 1

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 3

[18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DIN: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 6

[19] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *CVPR*, pages 1190–1199, 2023. 1

[20] Zeyu Ma, Yang Yang, Guoqing Wang, Xing Xu, Heng Tao Shen, and Mingxing Zhang. Rethinking open-world object detection in autonomous driving scenarios. In *ACM MM*, pages 1279–1288, 2022. 1

[21] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 2

[22] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, pages 728–755. Springer, 2022. 2, 6

[23] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *NeurIPS*, 2023. 2, 6

[24] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *ICCV*, pages 15579–15591, 2023. 3

[25] Zongshen Mu, Siliang Tang, Jie Tan, Qiang Yu, and Yueting Zhuang. Disentangled motif-aware graph learning for phrase grounding. In *AAAI*, pages 13587–13594, 2021. 2

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[27] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, pages 13018–13028, 2021. 3

[28] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[30] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: parts and attributes of common objects. In *CVPR*, pages 7141–7151, 2023. 2, 3, 4

[31] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021. 2

[32] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, pages 7031–7040, 2023. 6

[33] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 2

[34] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *NeurIPS*, pages 36067–36080, 2022. 2, 3

[35] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 2, 3

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[37] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368. Springer, 2022. 2, 6