Investigating the Role of Attribute Context in Vision-Language Models for Object Recognition and Detection

Kyle Buettner[1], Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh, PA, USA
buettnerk@pitt.edu, kovashka@cs.pitt.edu

Abstract

Vision-language alignment learned from image-caption pairs has been shown to benefit tasks like object recognition and detection. Methods are mostly evaluated in terms of how well object class names are learned, but captions also contain rich attribute context that should be considered when learning object alignment. It is unclear how methods use this context in learning, as well as whether models succeed when tasks require attribute and object understanding. To address this gap, we conduct extensive analysis of the role of attributes in vision-language models. We specifically measure model sensitivity to the presence and meaning of attribute context, gauging influence on object embeddings through unsupervised phrase grounding and classification via description methods. We further evaluate the utility of attribute context in training for open-vocabulary object detection, fine-grained text-region retrieval, and attribution tasks. Our results show that attribute context can be wasted when learning alignment for detection, attribute meaning is not adequately considered in embeddings, and describing classes by only their attributes is ineffective. A viable strategy that we find to increase benefits from attributes is contrastive training with adjective-based negative captions.

1. Introduction

Natural language has been shown to provide a strong signal for training visual representations. A visual-text alignment model can be pretrained with image-caption data and used for downstream tasks like object recognition, detection, and retrieval. While the text embeddings that represent object nouns are often used as classifier weights, the impact and utility of other caption context, especially attributes, are less clear. Consider the Fig. 1 caption: "A very **large furry brown** bear on a rock by the water." The model can learn grounding using only nouns (underlined), but *bear* can also be learned in the context of its attributes (bolded adjectives). Do alignment models use attribute context to learn *bear*? Do they distinguish between *brown bear* and *black bear*?

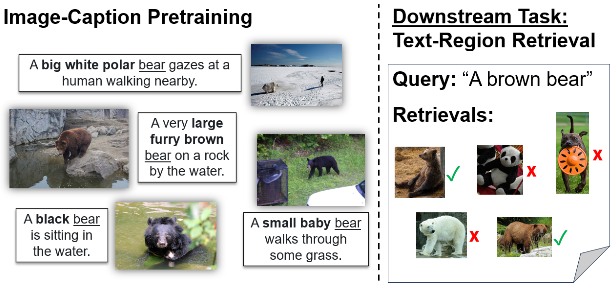
Image-Caption Pretraining
A **big white polar** bear gazes at a human walking nearby.
A very **large furry brown** bear on a rock by the water.
A **black** bear is sitting in the water.
A **small baby** bear walks through some grass.
Downstream Task:
Text-Region Retrieval
Query: "A brown bear"
Retrievals:
✓
✗
✗
✗
✓

Figure 1. **Do vision-language models effectively leverage attribute context in captions?** In captions, objects (*e.g.* bear) are often described with rich contextual information (*e.g.* attributes like **big**, **furry**, **black**). We evaluate the impact and utility of attributes in VL modeling through tasks such as text-region retrieval.

Attributes in captions can aid object recognition and detection in various ways. Attributes can serve as a proxy for a fine-grained category which is not explicitly mentioned (*e.g.* a *small, young cat* is a *kitten*). They can ensure that the alignment model is paying attention to the right features, rather than dataset artifacts (*e.g.* that a *bear* is being grounded as such because it is *brown/black*, rather than because its background is a *forest*). They can be used to specify subcategories, *e.g.* when a user desires detections that match a certain property (*e.g.* a *red car*, but not a *blue car*).

With this motivation, our goal is to better understand the connection between attributes and objects in vision-language (VL) models. In particular, we explore considerations such as whether models leverage attributes in captions as important signals for object learning and whether VL models can use object and attributes effectively in fine-grained tasks (*e.g.* recognizing "a large bird with a brightly colored bill"). We refer to a model's capabilities in using attribute information as *attribute sensitivity*.

We offer an extensive sensitivity analysis of two popular alignment paradigms for VL models, region-word grounding and whole image-text alignment, which we study with OVR-CNN [55] and CLIP [37], respectively. We investigate if model decisions consider attribute presence and meaning, specifically testing how attribute perturbations affect unsupervised phrase grounding and classification via description tasks. *We find an overall lack of sensitivity to attribute meaning*, inspiring an investigation into adjective-based contrastive negative sampling of captions. Our exploration results in strategies that increase the benefits from attribute context, exhibited through improvements in practical use cases of open-vocabulary object detection, fine-grained text-region retrieval, and object attribution.

Our main contributions are insights into these questions:

1. Does attribute context play an impactful role in VL pretraining for object detection?
2. Does learning to ground objects to contextualized word embeddings utilize attribute meaning?
3. Do VL models perform well at tasks where objects are described with/in terms of attributes?
4. Can contrastive negative sampling of captions increase a model's ability to use attribute context?
5. What sampling mechanisms are most effective?

2. Background and related work

Visual representation learning with language Common VL tasks such as phrase grounding and visual question answering leverage objectives that align and/or merge image and text features [1, 25, 27, 28, 44, 45, 52, 53]. Alignment achieved with large-scale contrastive learning [16] has powered the traditional vision task of image classification by enabling impressive zero-shot capability (*e.g.* ALIGN [22], CLIP [37]). The primary mechanism for adapting models like CLIP to recognition is through creating prompts (*e.g.* "a photo of a [classname]") for all classes and using the text encoder to convert prompts into classifier weights. Recent methods have exploited this open-vocabulary capability of CLIP to provide attribute context with LLM-based class descriptions [32, 36]. It is still unclear the extent to which VL models for recognition can consider attributes. As such, we conduct more in-depth experiments within the "classification via description" task of [32], highlighting limited utility of attributes in zero-shot recognition with CLIP.

Our analysis also hones in on object detection, which typically entails a predefined class list and bounding box annotations. The use of text embeddings has expanded the detection vocabulary [3, 24, 29, 56], and large-scale image-caption datasets and weakly supervised objectives have enabled "cheaper" supervision [9, 51]. Open-vocabulary detection [55], which involves training on base classes and using region-text alignment to extend to novel classes, has become especially popular. Recent open-vocabulary detectors leverage CLIP through mechanisms such as distillation, prompting, and pseudo-labeling [2, 12–14, 46, 47, 57, 58]. Our work impacts this area as we gauge the attribute sensitivity of the CLIP model widely used with these approaches. Additionally, through [55], we study fine-grained region-word alignment [8, 25, 29], which is tailored to region-level tasks [50]. We in particular examine the impact of attribute context in detection through comparing results to [55].

Bias and sensitivity measurement of embeddings in VL tasks Our work relates to efforts to understand the biases in embeddings from VL models. Such probing has highlighted that grounded/aligned embeddings encode social biases [41] and lack sensitivity to composition and word order [43, 54]. Our investigation more thoroughly analyzes embeddings with respect to attributes. For example, we gauge whether visual embeddings are sensitive to attribute presence and meaning when grounding to contextualized word embeddings (from [10]). The work of [4] relates as it involves use of contextualized object embeddings to detect object states (*e.g. sliced tomato*, *tomato in a bowl*). Our work instead explores if *enhancing the attribute sensitivity* of contextualized object embeddings impacts more general object detection and fine-grained text-region retrieval tasks.

Contrastive negative sampling "Hard" negative samples can benefit contrastive learning [23, 40]. We explore negative sampling of captions to enhance attribute context in region-word pretraining and CLIP finetuning. Past works have created negatives by replacing nouns [15] and by perturbing word order [54]. We alternatively test more attribute-tailored strategies by replacing only adjectives in captions, randomly/plausibly based on a dataset. We exhibit that order perturbations [54] do *not* help fine-grained text-region retrieval, while adjective negatives do. We also show that adjective negatives improve vs. a generic caption sampling baseline on Visual Genome Attribution [54], and that adjective negatives benefit detection in region-word pretraining. Concurrent work [11] has also shown the value of adjective negatives, though our work uniquely shows value in detection. [6] also leverages attribute perturbations, but alternatively with synthetic visual data.

Attributes in vision tasks Our work considers object attributes, described through adjectives in captions, with respect to object learning and model capabilities. Past work has explored attributes with respect to direct prediction [35], compositional zero-shot recognition with objects [33, 42], and use as a bridge between base and novel classes in zero-shot classification [26, 49]. With respect to localization, attributes have served as signals to spatially constrain object learning [21, 48] and as part of an open-vocabulary attribute detection task (detecting all attributes with an object) [5]. Our work is unique as we analyze attributes *as context for objects*, gauging impact in tasks like retrieval and detection.

3. Methodology

When VL models learn alignment for recognition and detection, the *utility* of attributes in captions is considerably underlooked. We study object representation sensitivity to attributes through case studies in region-word grounding (OVR-CNN [55]) and image-text alignment (CLIP [37]). This section outlines these frameworks, our measurement methods, and our strategies to enhance attribute context.

3.1. Vision-language frameworks of study

We study contrastive frameworks that learn in each iteration using a batch $\mathcal{B}$ of image-caption pairs ($\mathcal{B}_I$ and $\mathcal{B}_C$ for just images and captions, respectively). A score $\langle I, C\rangle$ is computed to quantify the relative matching between an image I and a caption C. Image-to-text and text-to-image contrastive objectives are as shown in Eqs. 1 and 2:

$$\mathcal{L}_{I\to T}(I) = -\log \frac{\exp\langle I, C\rangle}{\sum_{C'\in\mathcal{B}_C}\exp\langle I, C'\rangle} \tag{1}$$

$$\mathcal{L}_{T\to I}(C) = -\log \frac{\exp\langle I, C\rangle}{\sum_{I'\in\mathcal{B}_I}\exp\langle I', C\rangle} \tag{2}$$

Methods may differ in terms of how the scoring function is defined and whether losses include additional components such as temperature or normalization constants.

3.1.1 Case study: Region-word grounding

We explore region-word grounding to learn fine-grained alignment for detection. We specifically consider OVR-CNN [55], which employs a weakly supervised, region-word grounding pretraining task to learn class embeddings for open-vocabulary detection with Faster R-CNN [39]. The model resembles PixelBERT [19], using ResNet-50 [17], a pretrained BERT [10], and a BERT-like multimodal model. The total loss $\mathcal{L}(I, C)$ comprises four objectives: masked language modeling ($\mathcal{L}_{MLM}$), image-to-text matching ($\mathcal{L}_{ITM}$), and two contrastive grounding terms ($\mathcal{L}_G(C)$ and $\mathcal{L}_G(I)$). $\langle I, C\rangle$ for OVR-CNN is defined with Eqs. 3 and 4, where the dot product is taken between each word token e^C_j (n_C total produced from BERT's input layer) and each region token e^I_i (n_I total from ResNet then projected into the language embedding space with a V2L layer):

$$\langle I, C\rangle = \frac{1}{n_C}\sum_{j=1}^{n_C}\sum_{i=1}^{n_I} a_{i,j}\langle e^I_i, e^C_j\rangle \tag{3}$$

$$a_{i,j} = \frac{\exp\langle e^I_i, e^C_j\rangle}{\sum_{i'=1}^{n_I}\exp\langle e^I_{i'}, e^C_j\rangle} \tag{4}$$

Notably, this default alignment mechanism uses *context-free* word embeddings (e^C_j in BERT), which do not change with surrounding language context (*e.g*. *orange* has the same embedding in the captions "orange basketball" and "eating an orange"). We reason that this type of grounding contributes to misalignment of concepts, potentially inhibiting the benefits of attribute context. More recent models (*e.g*. CLIP [37]) also align visual regions to text embeddings contextualized through transformers. For expansive insights, we experiment with contextualization in OVR-CNN by altering Eq. 3 to use f^C_j, which are BERT's *output* embeddings that change with context (unlike BERT's e^C_j which are static). Eq. 5 shows this change:

$$\langle I, C\rangle = \frac{1}{n_C}\sum_{j=1}^{n_C}\sum_{i=1}^{n_I} a_{i,j}\langle f^I_i, f^C_j\rangle \tag{5}$$

Since word embeddings are dynamically contextualized, visual regions for an object are grounded to a *collection* of embeddings instead of one. Naive integration of such embeddings into detection results in poor performance. We use the following training recipe to effectively use contextualized embeddings in detection: (1) using a prompt "A/an <objName>." when changing a class embedding for object k from $e^{\mathcal{C}}_k$ to $f^{\mathcal{C}}_k$, (2) allowing the language encoder to update in the grounding pretraining task, and (3) allowing the V2L layer to update in finetuning. These strategies provide the training flexibility needed to thoroughly evaluate attribute sensitivity with contextualized embeddings.

3.1.2 Case study: CLIP image-text alignment

Open-vocabulary detectors that have come after OVR-CNN notably leverage CLIP [2, 12–14, 46, 47, 57, 58]. Their ability to use attribute context is thus highly dependent on the attribute sensitivity of CLIP. We study CLIP's attribute sensitivity for insights that generalize to various methods. The alignment objective of CLIP notably differs from OVR-CNN in that it aligns embeddings corresponding to entire images and text descriptions rather than to regions and words. More specifically, an image I and caption C are processed by CLIP's image and text encoders to produce normalized feature representations z^I_i and z^C_j. $\langle I, C\rangle$ for CLIP is defined in Eq. 6, where $\langle z^I_i, z^C_j\rangle$ is a dot product:

$$\langle I, C\rangle = \langle z^I_i, z^C_j\rangle \tag{6}$$

A temperature τ is also used with the losses in Eqs. 1 and 2. Due to CLIP's large size and scale, we focus on finetuning representations, rather than pretraining from scratch.

3.2. Analyzing model sensitivity to attributes

We aim to measure how influential attribute context is to a model's decision (*e.g*. classification, grounding). We reason that in an attribute-sensitive model, the *presence* of attributes should help decisions, as this information is complementary to objects. Additionally, the *meaning* of attributes should be respected. Object representations should be more aligned when correct attributes are used than when incorrect attributes are used. Our mechanism for exploring these considerations is through *removing* and *changing* attribute context in the text for a task, as removal tests presence, and changing tests meaning. In this section, we outline our measurement methodology, for which we explore prior tasks that can show attribute sensitivity while fitting each alignment mechanism. In particular, we use unsupervised phrase grounding [34] for OVR-CNN and classification via description [32] for CLIP, as shown in Fig. 2.

Isolating objects and attribute context For analysis of OVR-CNN (and for training as outlined in Sec. 3.3), we define a vocabulary $\mathcal{V}$ to be the nouns corresponding to objects in a dataset $\mathcal{D}$. In our study, $\mathcal{D}$ is COCO [7], with 118,287 images and 5 captions per image. We build $\mathcal{V}$ from the synonym list of COCO class names provided in [31], with plural terms added. The vocabulary $\mathcal{V}$ captures various terms for each class (*e.g. jet*, *aircraft*, *planes* for *airplane*). We identify a class attribute as any adjectival modifier ("amod") with dependency on a class synonym in $\mathcal{D}$, detected with [18]. The unique adjectives for each class make up respective **plausible** sets, containing attribute properties across the dataset (*e.g.* a *frisbee* is *red*/*green*/etc.). Unique adjectives across all classes make up the **random** set. We provide further details and statistics in the supp. material.

Measuring attribute sensitivity in region-word grounding OVR-CNN is analyzed using unsupervised phrase grounding [34], a task that returns a bounding box b for a text query t. Given an image-caption pair (I, C), we ask: if I has a *red car*, are visual regions for that car grounded better when using the car embedding in the caption "a red car..." than when using the embedding in "a blue car..." or "a car..."? Put another way, we test if the model leverages attribute meaning when grounding object regions to contextualized word embeddings. While a model could align visual regions for *car* independently of attributes (*e.g.* with context-free embeddings), we reason that bag-of-words behavior may result since embeddings are the same in cases like "a red car and blue truck"/"a blue car and red truck". Also, the model would not be fully leveraging capabilities of contextualized embeddings, where a region-word objective can encode attribute information within a contextualized object grounding, such that the model dynamically learns to represent *red car* vs. *blue car*.

In this setup, we test four grounding scenarios: (1) using the *baseline caption*, containing ground-truth attributes in adjective form (*e.g.* "a *yellow* banana on the table"); (2) using a caption that has object adjectives *removed* (*e.g.* "a banana on the table"); (3) using a caption that has object adjectives changed *plausibly* according to our sets (*e.g.* "a *rotten* banana on the table"); and (4) using a caption that has adjectives changed *randomly* to be any intra-corpus (*e.g.* "a *red* banana on the table"). In an attribute-sensitive model, we expect the top-performing grounding to have the most information (*e.g.* "yellow banana"). We expect removal performance to drop vs. this baseline as objects are less specified. We reason that changing adjectives should make attributes *incorrect* and thus hurt vs. the baseline. In the plausible case, we expect the dataset to cover disjoint states (*e.g. wooden* vs. *plastic* spoon). While multiple attributes could be valid for an object, in practice, we find such cases rare. On 100 random samples, we find that 84% of captions changed plausibly and 92% changed randomly are not reasonably correct. We expect changing plausibly to thus result in a smaller drop from the baseline vs. changing randomly.

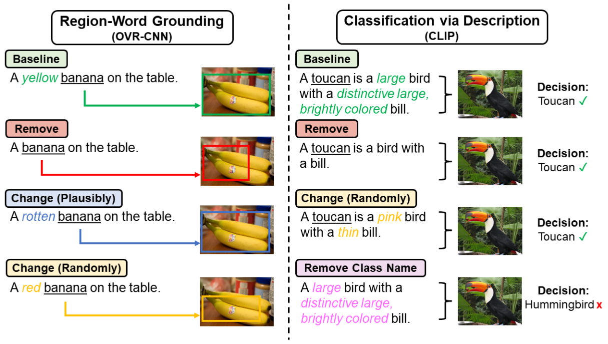
Region-Word Grounding
(OVR-CNN)
Baseline
A yellow banana on the table.
Remove
A banana on the table.
Change (Plausibly)
A rotten banana on the table.
Change (Randomly)
A red banana on the table.
Classification via Description
(CLIP)
Baseline
A toucan is a large bird with a distinctive large, brightly colored bill.
Decision: Toucan ✓
Remove
A toucan is a bird with a bill.
Decision: Toucan ✓
Change (Randomly)
A toucan is a pink bird with a thin bill.
Decision: Toucan ✓
Remove Class Name
A large bird with a distinctive large, brightly colored bill.
Decision: Hummingbird x

Figure 2. **Our attribute sensitivity measurement methodology.** We remove/change attributes/class names in text for grounding and classification tasks to measure if model decisions are sensitive to attributes. We show example predictions when attribute meaning is ignored (*e.g. rotten* banana, toucan is a *pink* bird).

To compute groundings, for each caption token j, if it matches an object term in $\mathcal{V}$, one or more bounding boxes are generated from the binary map of region-word similarity $\langle f_i^I, f_j^C \rangle$ such that $\langle f_i^I, f_j^C \rangle \geq th_{sim}$. In the supp. material, we test at values th_{sim}=5, 10, 15 and show trends are not sensitive to this threshold. Then for all captions which mention that object, these boxes are compared to the groundtruth at various IoU thresholds, producing AP@t values. We use t=30,40,50 as non-aggressive thresholds suitable for unsupervised inference. The average AP@IoU=30:10:50 is reported over all classes in the COCO validation set.

Measuring attribute sensitivity in CLIP image-text alignment We analyze CLIP's attribute sensitivity through classification via description [32], which adds attribute context to object prompts to aid zero-shot inference. In [32], for each class c in a dataset $\mathcal{D}$, GPT-3 is prompted to produce a list of descriptors $D(c)$. The descriptors contain attributes relevant to the object, along with c to condition the attributes. For instance, the descriptors produced for *toucan* are "a/an toucan which (is/has/etc) *large*, *brightly colored* bill.", "a/an toucan which (is/has/etc) *long*, *pointed* wings.", etc. To classify an image I, each descriptor d serves as a prompt. The score for each class is computed using the average CLIP logits, $\phi(I, d)$, over each d, shown in Eq. 7:

$$s(c, I) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(I, d) \tag{7}$$

We select $\mathcal{D}$ to be ImageNetV2 [38] and use GPT-3 (*davinci-002*) to produce descriptors. We also test producing only a single-sentence description to simulate the related method [36] (*e.g.* "A toucan is a *large* bird with a *distinctive large, brightly colored* bill."). With both setups, we test sensitivity through removing and changing, detected as "ADJ" with [18], but unlike OVR-CNN, we only use random changing (and not plausible) since the descriptors do not have a vocabulary like COCO. Then to further stress test CLIP, in a given inference, we remove *all* class names by replacing them with "a/an object". This experiment gauges whether CLIP can interpret objects from attribute-only descriptions (*e.g.* a *small*, *white*, *round* object with *red seams* is a *baseball*). In the supp. material, we provide specific details, examples, and linguistic properties for descriptors.

3.3. Enhancing model sensitivity to attributes

We also hypothesize that *enhancing* attribute context's role can help tasks like object detection and text-region retrieval. We specifically experiment with *adjective-based negative caption sampling*, where a negative for a caption C includes the same words, just with an adjective replaced (*e.g.* for "*blue* car in the street", *blue* is replaced with *red*). We reason that these negatives can encourage models to capture attribute meaning when learning objects, increasing the model's fine-grained utility. In pretraining specifically, another benefit is that attributes may help "guide" object grounding to the correct regions (*e.g.* a *car* to a *red* region).

We test negative sampling in OVR-CNN pretraining and CLIP finetuning, both with COCO. We explore two replacement methods: (1) choosing a *random* adjective from the corpus and (2) choosing a *plausible* adjective for a noun, such that it is mentioned intra-dataset with the respective class term. Through these strategies, we aim to gauge whether it is beneficial to contrast disjoint states in a dataset with *plausible* (*e.g.* *wooden* vs. *metal spoon*) or if simple *random* adjectives suffice. Table 1 shows examples of plausible and random captions. To implement in training, for each caption in $\mathcal{B}_C$ with an adjective detected, a negative caption is added to a batch $\mathcal{B}_N$. The loss in Eq. 1 becomes:

$$\mathcal{L}_{I\rightarrow T}(I) = -\log \frac{\exp\langle I, C\rangle}{\sum_{C' \in \mathcal{B}_C + \mathcal{B}_N} \exp\langle I, C'\rangle} \tag{8}$$

A potential shortcut with region-word grounding is that a model can solve the task by grounding just adjectives rather than object words. To encourage OVR-CNN to consider objects and attributes, we use noun negatives (using the same caption, but replacing nouns with random ones from $\mathcal{D}$). For CLIP, if no adjective-noun pair is detected, we add a random caption to $\mathcal{B}_N$. We also compare the plausible and random strategies to order-perturbing sampling [54], since order perturbations can influence attention to attributes (*e.g.* "red car and blue truck" vs. "red truck and blue car"). We test this strategy by perturbing order when possible (i.e. the caption has adjectives and nouns); as with other strategies, we sample a random caption otherwise.

Caption	A bunch of **green** bananas growing in a tree.
Plausible Neg.	A bunch of ***rotten*** bananas growing in a tree.
Random Neg.	A bunch of *pink* bananas growing in a tree.

Table 1. Examples of negative adjective captions.

4. Evaluation

We evaluate context enhancement on one object-focused task (open-vocabulary detection) and two fine-grained tasks (text-region retrieval/object attribution).

Datasets For image-caption training, we use COCO Captions [7], and for finetuning open-vocabulary detection, we use COCO Objects [30], (2017 train/val for both). The class split for open-vocabulary detection is the same as [3,55] (48 base and 17 target classes). For retrieval, we use the 2,000 image COCO val subset with object and attribute annotations from OVAD [5]. For attribution, we use ARO Visual Genome Attribution (VGA) [54], with 28,748 examples.

Open-vocabulary object detection This task considers *base*/*target* class sets with/without bounding box annotations. A detector (i.e. Faster R-CNN [39]) is trained only on base classes, and there are three evaluation settings: *base* classes only, *target* only, and *generalized*. As in [55], base only and target only classify over the respective set, while in generalized, prediction is performed over the union of base and target classes, and results are reported within each group and overall. We report AP_{50} as the metric, as in [55].

Text-region retrieval We pose fine-grained text-to-region retrieval as a use case where attribute-object understanding is needed. We input a set of texts $\mathcal{T}$ for which each text t contains an attribute a from the set $\mathcal{A}$ and an object o from the set $\mathcal{O}$ (*e.g.* *red car*). The goal is to return as output top-scoring regions that are correct if they contain the correct attribute *and* object (*e.g.* for *red car*, non-red cars would be incorrect). We select $\mathcal{A}$ to be colors, patterns (striped, dotted, etc.), and materials (metal, wooden, etc.) in OVAD [5] and $\mathcal{O}$ to contain all COCO objects. Since OVAD's annotations are dense, we exclude attribute-object pairs that are not described in language due to being inherent (*e.g.* *metal car*) and use attribute-object pairs with greater than 10 annotations. Overall, we use 323 attribute-object pairs (273 with colors, 42 materials, and 8 patterns). In evaluation, every ground-truth box in OVAD is considered a possible retrieval (≈14,300 samples). For CLIP, we input crops for each GT box to the image encoder and use similarity between image and text features to rank retrievals. For OVR-CNN, we compute the region embedding f_i^I for each box. Then for all text t in $\mathcal{T}$, we compute the dot product between the average word embedding f_j^t of its attribute and object text tokens (*e.g.* average(f_{red}, f_{car})) and every f_i^I. We report recall@k (a true positive is when at least one retrieval of the correct attribute and object is within the top k). We also report precision@k, the proportion of correct retrievals in the top k. We do not directly evaluate on OVAD since the task has the different goal of detecting all attributes rather than differentiating between categories described with attributes.

Object attribution For CLIP, we consider object attribution (with the VGA dataset [54]) as a relevant benchmark for image-text matching. This task involves selecting the correct text for an image, given two choices with different order (*e.g.* "the crouched cat and the open door" vs. "the open cat and the crouched door"). Note that this task is *complementary* to the retrieval task, in that they both test attribute understanding, but text-region retrieval focuses more on fine-grained differentiation among *plausible* attribute-object pairs (*blue car* vs. *red car* vs. *blue truck*), while attribution focuses on intra-caption ordering where negative pairs are often *implausible* (*e.g.* "crouched door").

Training Full-scale comparison to [55] uses 8 Quadro RTX 5000 GPUs and settings from [55]. For other OVR-CNN results, we pretrain using 4 NVIDIA GeForce GTX 1080 Ti with memory 11 GB. Pretraining uses 80k iter., batch size (BS) 16, and learning rate (LR) 0.01 that scales down 10x after 40k/70k steps. For COCO finetuning, we use 4 GPUs, 75k iter., BS 8, and LR 0.005 that scales down after 30k/60k steps. CLIP finetuning is performed using OpenCLIP [20], for 5 epochs using BS 64 and LR 1e-6 on 1 Quadro RTX 5000. CLIP's image encoder is ViT-B/32.

5. Experimental results and analysis

In Section 5.1, we analyze the attribute sensitivity of VL alignment. For OVR-CNN region-word grounding, we test removing context in the captions used for pretraining detection (Fig. 3) and perturbing captions in unsupervised phrase grounding (Fig. 4). For CLIP image-text alignment, we test perturbing the text prompts for classification via description (Fig. 5). In Section 5.2, we further evaluate how attribute context sensitivity impacts practical downstream tasks. We evaluate the impact of attribute sensitivity on an *object-focused* task, in particular open-vocabulary detection with OVR-CNN (Table 2/3). We also evaluate models on two *fine-grained* tasks that require attribute knowledge, namely, text-region retrieval and object attribution (Table 4/5).

5.1. Gauging the role of attribute context

Attribute context has limited impact in region-word pretraining for object detection. We first examine the role of attributes through *removing all "amod" from captions* during VL pretraining with OVR-CNN. Open-vocabulary detection results for baseline OVR-CNN [55] are shown in Fig. 3. Note that the max. drop from training with to without adjectives is -0.36 AP_{50} (base), and there are not discernible drops in target/generalized settings. *These results point to attribute context being wasted and not helpful when learning object grounding*, and thus serve as inspiration for our investigation of ways to boost use of attribute context.

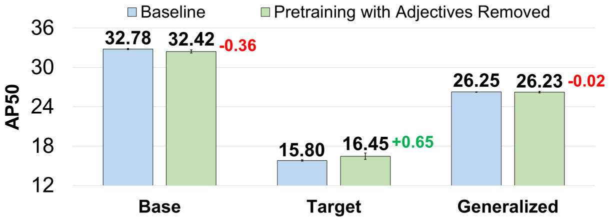

Figure 3. **Effects from removing adjectives in OVR-CNN pretraining on COCO open-vocabulary detection**. Across settings, the maximum drop from removing adjectives from training is only -0.36 AP_{50}. These results indicate that attribute context has limited benefit in detection. Error bars show std. error (3 trials).

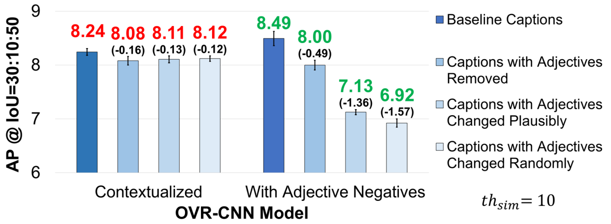

Figure 4. **Measuring attribute sensitivity in contextualized object grounding**. We find limited sensitivity to attribute meaning in default contextualized grounding, but enhanced sensitivity with (plausible) adjective negatives added. This observation is supported by AP differences (in black) with incorrect adjectives used for the adjective negative vs. default contextualized models. The drops are discernibly larger with adjective negatives: -1.36% vs. <0.2% from baseline captions to changing plausibly and -1.57% vs. <0.2% from baseline captions to changing randomly. Values are avgs. over 3 training runs. Bars show std. error.

Contextualizing object grounding does not result in embeddings with high sensitivity to attribute meaning. As outlined in Sec. 3.1.1, we contextualize grounding in OVR-CNN as one strategy to integrate attribute context. Then through unsupervised phrase grounding, we gauge sensitivity to attribute meaning and analyze whether the attributes contextualizing an object noun (*e.g.* "a *red* car") impact performance. Fig. 4 shows AP@IoU=30:10:50 for (1) OVR-CNN with contextualization and (2) OVR-CNN with contextualization *and* plausible adjective/noun negatives, on the four region-word grounding scenarios of interest (baseline grounding, removing adjectives, changing adjectives plausibly, and changing adjectives randomly). On the left of Fig. 4, we find that with default contextualization, changing adjectives plausibly/randomly yields similar AP to using baseline captions or captions with removed adjectives

Adjective	Noun	Grounding	LE/PL	Base-Only		Target-Only		Generalized					
Negative	Negative	Type	Trained	AP_{50}	Δ	AP_{50}	Δ	All AP_{50}	Δ	Base AP_{50}	Δ	Target AP_{50}	Δ
-	-	Context-Free	-	32.8 ± 0.08	–	15.8 ± 0.11	–	26.3 ± 0.04	–	31.4 ± 0.15	–	11.8 ± 0.28	–
Plausible	✓	Contextualized	✓	**35.8** ± 0.09	**+3.0**	17.7 ± 0.38	+1.9	**28.8** ± 0.17	**+2.5**	**33.9** ± 0.24	**+2.5**	14.2 ± 0.34	+2.4
Random	✓	Contextualized	✓	35.7 ± 0.31	+2.9	18.0 ± 0.25	+2.2	28.6 ± 0.33	+2.3	33.5 ± 0.30	+2.1	**14.5** ± 0.41	**+2.7**
-	✓	Contextualized	✓	35.3 ± 0.19	+2.5	17.8 ± 0.18	+2.0	28.3 ± 0.12	+2.0	33.3 ± 0.13	+1.9	14.2 ± 0.13	+2.4
-	-	Contextualized	✓	35.2 ± 0.13	+2.4	16.7 ± 0.26	+0.9	28.3 ± 0.20	+2.0	33.6 ± 0.16	+2.2	13.1 ± 0.30	+1.3
-	-	Contextualized	-	31.8 ± 0.14	-1.0	10.5 ± 0.28	-5.3	22.7 ± 0.83	-3.6	28.0 ± 0.98	-3.4	7.5 ± 0.47	-4.3
Plausible	✓	Context-Free	✓	34.1 ± 0.21	+1.3	**19.3** ± 0.29	**+3.5**	28.4 ± 0.17	+2.1	33.4 ± 0.19	+2.0	14.3 ± 0.38	+2.5
-	-	Context-Free	✓	34.1 ± 0.01	+1.3	19.1 ± 0.72	+3.3	28.3 ± 0.27	+2.0	33.2 ± 0.12	+1.8	14.4 ± 0.70	+2.6

Table 2. **Adapting OVR-CNN [55] with attribute context enhancement strategies (Sec. 3.1.1/3.3): adjective/noun negative caption sampling, contextualized grounding, language encoder/projection layer training (LE/PL)**, AP_{50} mean over 3 trials ± std error, Δ=change vs. default OVR-CNN [55] (top row). Using adjective negatives *with* contextualization yields base/generalized AP_{50} increases, and top base/generalized AP_{50} overall, as the model is able to take into account attribute meaning in object embeddings.

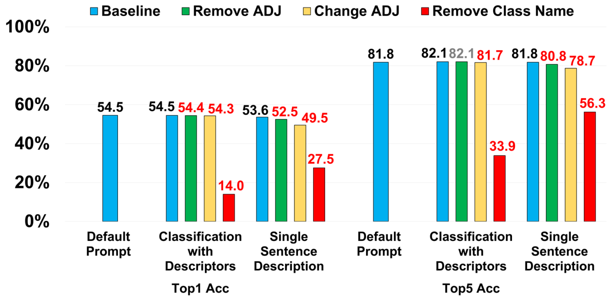

Figure 5. **Perturbing attributes and object names in CLIP descriptions used for ImageNetV2 classification.** Removing and changing adjectives have small effects on accuracy. When classes are described without object names, accuracy significantly drops.

(a max. difference of 0.13 AP@IoU=30:10:50). *These observations are counterintuitive*, as embeddings can be contextualized by incorrect adjectives, yet ground similarly to when there are correct adjectives. We posit that the model may be sensitive to caption structure, where object embeddings with different adjectives are close together, and the model does not have an incentive to differentiate them. Such lack of sensitivity to attribute meaning motivates our exploration of *adjective negatives*; we show the effects on the right in Fig. 4. Contextualization aptly becomes less aligned with incorrect adjectives, reaching notable drops when changing plausible/randomly with respect to the baseline (-1.36%/-1.57% respectively). In Sec. 5.2, we show the importance of sensitivity in detection and retrieval.

Describing classes in terms of attributes alone is ineffective. We measure CLIP's sensitivity to attributes with classification via description. As outlined in Sec. 3.2, zero-shot inference is performed on ImageNetV2 using CLIP default prompting, LLM-based sets of object feature descriptions [32], and LLM-based single-sentence descriptions of objects [36]. In Fig. 5, we show the results of removing/changing adjectives and removing class names in terms of top1/5 accuracy. Removing/changing adjectives results in *insignificant drops* vs. the baseline with [32], and slightly bigger drops with the [36]-like method (-4.1% drop baseline to changing), potentially as a result of more adjective-dense descriptions (supported in the supp.). However, *removing class names* results in close to *ten times more substantial* drops (max -40.5% top1 accuracy). These results bring into question the model's ability to leverage attribute descriptions since *class names drive performance*. Such results also limit the appeal of using attribute descriptions for new/custom objects with names not in the pretraining set.

5.2. Evaluating context enhancement strategies

Enhancing context sensitivity helps open-vocabulary detection in base and generalized settings. We evaluate OVR-CNN with four strategies to boost attribute context in region-word pretraining: (1) contextualized grounding, (2) adjective negative sampling (plausible/random), (3) noun negative sampling (random), and (4) language encoder/projection layer training, with results shown at an experimental scale in Table 2. Compared to the baseline [55], combining all strategies, in both plausible and random adjective negative cases, provides the largest gains in base-only and generalized (all-class) settings (*e.g.* +3.0 and +2.5 AP_{50} respectively with plausible). In Table 3, we also present a proof-of-concept showing that enhancing attribute context improves the results reported in [55] in 4/5 settings (+0.9-1.0 AP_{50} in base-only and all generalized settings). *Such results highlight value in better using context, especially attributes, when learning grounding for detection.*

Breaking down Table 2, a key observation is that plausible/random adjective negatives, when used with contextualized grounding, result in (comparable) base and generalized gains over all other baselines (+0.5 and +0.4 AP_{50} with plausible). These results can be ascribed to increased attention to attribute meaning that is obtainable with contextualized grounding, but not with context-free grounding

Method	Base	Target	Generalized		
			Base	Target	All
OVR-CNN [55]	46.8	**27.5**	46.0	22.8	39.9
+ Context Enhancement	**47.7**	26.5	**46.9**	**23.8**	**40.8**

Table 3. **OVR-CNN at full scale with various context enhancement strategies** (plausible adjective/noun negatives, contextualized grounding, language encoder/projection layer training), compared to baseline reported in [55]. AP_{50} reported on COCO.

since embeddings do not vary with context. *There is marked benefit to learning to ground objects with attribute signals for detection.* Still, there is a tradeoff between contextualized and context-free grounding. Contextualized models result in top AP_{50} in base-only and all generalized settings, but context-free results in top AP_{50} in target-only. These results can be attributed to using contextualized embeddings and *not* adjective negatives, since all contextualized methods obtain worse target performance than the best context-free method. We reason that the drop is due to the need for a prompt: we use a simple "A/an <objName>." (see Sec. 3.1.1), but this prompt may be suboptimal to represent the large variance of contextualized embeddings for an object. Training with box annotations in base may allow visual embeddings to adjust to prompts, explaining base gains, but with no target training, adjustment cannot occur. We surmise that recent work in context optimization [12] can overcome this challenge. The noun negatives notably improve target-only vs. contextualized (+1.1 AP_{50}), showing that differentiating nouns in the same context may also help.

We further inspect the plausible case by comparing class-by-class results using models in row 2/4 of Table 2. Notably, the classes with top AP_{50} gains are *oven* (+4.6), *bear* (+4.3), *horse* (+3.6), and *frisbee* (+3.4). Upon inspection of the corpus, these are commonly described in captions with visually distinctive adjectives that may help grounding such as colors (*e.g.* "yellow frisbee"). Overall, we observe that 32/48 classes improve in AP_{50} with adjective negatives.

Adjective negatives increase CLIP's fine-grained utility in multiple tasks. We use text-region retrieval and attribution as fine-grained tasks to evaluate attribute-object understanding. Table 4 shows these results comparing strategies for finetuning CLIP on COCO: (1) choosing a random negative caption, (2) order-perturbing adjectives/nouns [54], (3) random adjective sampling, and (4) plausible adjective sampling. On retrieval, random adjective sampling is generally most effective across values of k, plausible is second, and both strategies outperform a random caption baseline and the order-perturbing captions of [54]. *The fine-grained differentiation needed for retrieval is aided best by adjective negatives.* On the attribution task, the order-perturbing negatives perform best, which makes sense given that attribution involves determining the correct order of adjectives and nouns. It is notable that adjective negatives improve on this task *and* retrieval vs. a random caption baseline, *unlike the order-perturbing captions*. This shows adjective negatives achieve more generalizable attribute-object understanding across tasks. Adjective negatives similarly improve in retrieval for OVR-CNN (Table 5). Plausible and random adjective sampling are more competitive in this scenario, though random sampling has highest R@1/P@1 and plausible sampling P/R@5/10. We surmise that random adjective sampling may solidify easier retrievals by comparing to a wide array of adjectives, while plausible sampling may help the model differentiate between tougher cases as plausible adjectives serve as more realistic, *harder* negatives.

Method	R@1	R@5	R@10	P@1	P@5	P@10	VGA
Default CLIP	48.92	82.97	90.40	48.92	42.66	37.62	62.82
Random Neg.	57.59	87.62	**94.12**	57.59	50.96	44.37	64.64
Order-Based Neg. [54]	56.97	85.76	92.88	56.97	48.73	42.79	**73.87**
Plausible Adj. Neg.	58.82	86.69	93.81	58.82	50.96	44.77	67.94
Random Adj. Neg.	**60.06**	**88.24**	92.26	**60.06**	**51.76**	**44.98**	67.93

Table 4. **Fine-grained utility of CLIP finetuned with negative sampling strategies, on T2R retrieval and Visual Genome Attribution (VGA) [54].** Recall/precision@k=1,5,10 are reported for T2R retrieval and accuracy for VGA. Best=**bold**, second=underlined, results > random baseline (row 2) in green. Note that adjective sampling offers improvements across *both* attribute tasks, while order only helps on the order-based VGA task.

Method	R@1	R@5	R@10	P@1	P@5	P@10
Contextualized Baseline	13.21	37.36	52.01	13.21	12.84	11.75
Plausible Adjective Negative	16.82	**39.83**	**53.35**	16.82	**14.14**	**12.63**
Random Adjective Negative	**17.44**	39.42	52.53	**17.44**	13.70	12.32

Table 5. **Fine-grained utility of OVR-CNN, pretrained with adjective negatives, in text-region retrieval of attribute-object concepts.** Recall/precision@k=1,5,10 are reported over 3 trials.

6. Conclusion

We answer these questions (Sec. 1): (1) Attribute context can show limited impact in region-word pretraining for detection. (2) Grounding objects to contextualized word embeddings increases attribute consideration only to a limited degree. (3) Describing CLIP's classes by only their attributes results in poor accuracy. Also, models struggle at fine-grained retrieval. (4) Adjective-based negative caption sampling is promising to increase model sensitivity to attribute meaning and especially boosts fine-grained retrieval. (5) Plausible and random adjective sampling are competitive in detection/retrieval following OVR-CNN grounding; with CLIP, random sampling has higher retrieval gains.

Acknowledgements: This work was supported by a National Science Foundation Grant No. 2006885.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 2022. 2

[2] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022. 2, 3

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 2, 5

[4] Gedas Bertasius and Lorenzo Torresani. COBE: Contextualized object embeddings from narrated instructional video. *Advances in Neural Information Processing Systems*, 33:15133–15145, 2020. 2

[5] Maria A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7041–7050, 2023. 2, 5

[6] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165, 2023. 2

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4, 5

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 2

[9] Karan Desai and Justin Johnson. VirTex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 3

[11] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023. 2

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 2, 3, 8

[13] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 266–282. Springer, 2022. 2, 3

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *International Conference on Learning Representations, ICLR*, 2022. 2, 3

[15] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE, 2006. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[18] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. SpaCy: Industrial-strength Natural Language Processing in Python. 2020. 4, 5

[19] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 3

[20] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021. 6

[21] Achiya Jerbi, Roei Herzig, Jonathan Berant, Gal Chechik, and Amir Globerson. Learning object detection from captions via textual scene attributes. *arXiv preprint arXiv:2009.14558*, 2020. 2

[22] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[23] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020. 2

[24] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[26] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2013. 2

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 2

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[30] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[31] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018. 4

[32] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *International Conference on Learning Representations, ICLR*, 2023. 2, 4, 7

[33] Nihal V Nayak, Peilin Yu, and Stephen H Bach. Learning to compose soft prompts for compositional zero-shot learning. *International Conference on Learning Representations, ICLR*, 2023. 2

[34] Giacomo Nebbia and Adriana Kovashka. Doubling down: Sparse grounding with an additional, almost-matching caption for detection-oriented multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4642–4651, June 2022. 4

[35] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021. 2

[36] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? Generating customized prompts for zero-shot image classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 5, 7

[37] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[38] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 5

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 3, 5

[40] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *International Conference on Learning Representations, ICLR*, 2021. 2

[41] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online, June 2021. Association for Computational Linguistics. 2

[42] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022. 2

[43] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[44] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[45] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *International Conference on Learning Representations, ICLR*, 2022. 2

[46] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 2, 3

[47] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7031–7040, 2023. 2, 3

[48] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017. 2

[49] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020. 2

[50] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. DetCLIPv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023. 2

[51] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2Det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9686–9695, 2019. 2

[52] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 2

[53] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[54] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *International Conference on Learning Representations, ICLR*, 2023. 2, 5, 6, 8

[55] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1, 2, 3, 5, 6, 7, 8

[56] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, 2022. 2

[57] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2, 3

[58] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 2, 3