VEIL: Vetting Extracted Image Labels from In-the-Wild Captions for Weakly-Supervised Object Detection

Anonymous ACL submission

Abstract

The use of large-scale vision-language datasets is limited for object detection due to the negative impact of label noise on localization. Prior methods have shown how such large-scale datasets can be used for pretraining, which can provide initial signal for localization, but is insufficient without clean bounding-box data for at least some categories. We propose a technique to "vet" labels extracted from noisy captions, and use them for weakly-supervised object detection (WSOD), without any bounding boxes. We analyze the types of label noise in captions, and train a classifier that predicts if an extracted label is actually present in the image or not. Our classifier generalizes across dataset boundaries and across categories. We compare the classifier to nine baselines on five datasets, and demonstrate that it can improve WSOD without label vetting by 30% (31.2 to 40.5 mAP when evaluated on PASCAL VOC).

1 Introduction

Freely available vision-language (VL) data has shown great promise to advance vision tasks (Radford et al., 2021; Mahajan et al., 2018; Jia et al., 2021). Unlike smaller, curated vision-language datasets like COCO (Lin et al., 2014), captions on the web (Ordonez et al., 2011; Desai et al., 2021; Changpinyo et al., 2021) only *partially* describe the corresponding image, and often describe the *context* behind it, including objects that do not appear in the image. We hypothesize this poses a greater challenge for weakly-supervised object detection (WSOD) than learning cross-modal representations for image recognition (e.g. as in CLIP). WSOD involves learning to localize objects, i.e. predict bounding box coordinates along with the corresponding semantic label, from image-level labels only (i.e. using weaker supervision than the outputs expected at test time). WSOD has primarily been applied (Ye et al., 2019a; Fang et al.,

Figure 1: Extracted labels from captions raise challenges such as missing objects or defects, annotated in our dataset, Caption **La**bel **N**oise. **None of the underlined objects are clearly visible.** We propose a method to detect such noise and compare it to alternatives.

2022) to smaller paid-for crowdsourced vision-language datasets like COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014).

Unlike captions written by annotators for the purpose of faithfully describing an image, captions on the web go beyond a redundant, descriptive relationship with their corresponding image. For example, a word can be used in literal or metaphorical ways ("that was a piece of cake") or have multiple senses, of which only one corresponds to the sense intended by the object detection vocabulary. A caption could share a story by including context that goes beyond the visual contents of the image but mention an object name, by providing location names and unpictured interactions with objects as shown in Figure 1. All of this is relevant as narration for the image but not as supervision for precise localization. On the visual side, user-uploaded content frequently features diverse object presentations, including intriguing atypical or hand-drawn objects or photos taken from within vehicles ("in my car"). We refer to image-level labels extracted from captions, that are incorrect (object not present in corresponding image), as **v**isually **a**bsent **e**xtracted **l**abels (VAELs). We show VAELs pose a challenge for weakly-supervised detection.

To cope with this challenge, we propose **VEIL**, short for **V**etting **E**xtracted **I**mage **L**abels, to directly learn whether a label is clean or not from *caption context*. We first extract potential labels from each caption using substring matching or exact match (Ye et al., 2019b; Fang et al., 2022). We then use a transformer to predict whether each extracted label is visually present or absent. We refer to this prediction *task* as extracted label vetting. We bootstrap labels from an ensemble of two pretrained object recognition models (Jocher et al., 2021; Zhang et al., 2021), to predict image-level pseudo-ground-truth visual presence labels on a variety of large-scale, noisy datasets: Conceptual Captions (Sharma et al., 2018), RedCaps (Desai et al., 2021), and SBUCaps (Ordonez et al., 2011). While these detectors are trained on COCO and similar datasets, they generalize well to estimating extracted label visual presence on in-the-wild VL datasets; however, their predictions are better used as targets for VEIL, rather than directly for vetting. Once we vet the extracted labels, we use them to train a weakly-supervised object detector.

We investigate sources of noise across three in-the-wild datasets from diverse sources: photo-sharing platform, social media platform, and images with alt-text (typically used for VL pretraining). We collect and will release a small dataset with annotations on object visibility (label noise) and object appearance defects (visual noise such as atypical appearance). To support using language context to filter object labels, we annotate linguistic indicators of noise which explain why a VAEL is absent from the image but included in the caption, such as describing context outside the image, non-literal use, different word sense, etc. We compare our label vetting method to nine baselines, including standard cross-modal alignment prediction methods (CLIP), adaptive noise reduction methods, pseudo-label prediction, simple rule-based methods, and no vetting. Our method improves upon the baselines both in terms of predicting extracted label visual presence (measured with F1) and producing cleaner training data for object detection leading to an improvement of +10 mAP over Large Loss Matters (Kim et al., 2022) and +3 mAP improvement over using CLIP (Radford et al., 2021) for filtering. We show a significant improvement when training WSOD with both clean (annotated in Pascal VOC 07) and noisy, but vetted labels from SBUCaps (51.31 mAP) compared to naively combining clean with noisy labels without vetting (42.06 mAP) or only using clean labels (43.48 mAP). Lastly, VEIL generalizes and its gains persist across datasets, object vocabulary, and scale.

To summarize, our contributions are as follows:

1. We propose VEIL, a transformer-based extracted label, visual presence classifier.
2. VEIL outperforms language-conditioned and language-agnostic label noise detection/correction approaches in vetting labels from a wide set of in-the-wild datasets for weakly-supervised object detection.
3. VEIL enables effective combination of extracted noisy and clean labels.
4. Even when VEIL is trained on one dataset/-category, but applied to another, it shows advantages over baselines.
5. We construct the **C**aption **La**bel **N**oise dataset.

2 Related Work

Vision-language datasets include crowdsourced captions (Young et al., 2014; Lin et al., 2014; Huang et al., 2016; Krishna et al., 2016) and alt-text written by users to aid visually impaired readers (Sharma et al., 2018; Changpinyo et al., 2021; Radford et al., 2021; Schuhmann et al., 2021), widely used for vision-language grounding due to abundance and assumed visual-text alignment. There are also large in-the-wild datasets sourced from social media like Reddit (Desai et al., 2021) and user-uploaded captions for photos shared on Flickr (Ordonez et al., 2011). We show the narrative element found in these in-the-wild datasets, captured by the linguistic cues we investigate, impact the ability to successfully train an object detection model.

Weakly-supervised object detection (WSOD) is a multiple-instance learning problem to train a model to localize and classify objects from image-level labels (Bilen and Vedaldi, 2016; Tang et al., 2017a; Wan et al., 2019; Gao et al., 2019; Ren et al., 2020). Cap2Det was the first work to leverage unstructured text accompanying an image for WSOD by predicting pseudo image-level labels from captions (Ye et al., 2019b; Unal et al., 2022). However, Cap2Det cannot operate across novel categories as it directly predicts image-level labels. Further, Cap2Det targets false negatives (visually present, not extracted labels), not visually absent extracted labels. Detic (Zhou et al., 2022) uses weak supervision from ImageNet (Deng et al., 2009) and extracts labels from Conceptual

Captions (CC) to pretrain an open vocabulary object detection model with a CLIP classifier head. While these approaches succeed in leveraging relatively clean, crowdsourced datasets like COCO, Flickr30K and ImageNet, both see lower performance in training with CC (Unal et al., 2022; Zhou et al., 2022). Other prior work (Gao et al., 2022) uses a pretrained vision-language model to generate pseudo-bounding box annotations, but always requires clean data (COCO), and does not explicitly study the contribution of in-the-wild datasets.

Vision-language pre-training for object detection. Image-text grounding has been leveraged as a pretraining task for open vocabulary object detection (Rahman et al., 2020a,b; Zareian et al., 2021; Gu et al., 2022; Zhong et al., 2022; Du et al., 2022; Wu et al., 2023), followed by supervision from bounding boxes from base classes. Some methods distill knowledge from existing pretrained vision-language grounding models like CLIP and ALIGN (Jia et al., 2021) to get proposals (Shi et al., 2022) and supervision for object detection (Du et al., 2022; Zhong et al., 2022); however the latter do not compare clean vs noisy supervision in a setting without bounding boxes. In contrast, we perform weakly-supervised object detection (WSOD) using noisy image-level labels from captions only. WSOD is a **distinct task** from open-vocabulary detection and has the **advantage** of not requiring expensive bounding boxes on base classes. We focus on **rejecting labels** harmful for localization.

Adaptive label noise reduction in classification. Adaptive methods reject or correct noisy labels ad-hoc during training. These methods exploit a network's ability to learn representations of clean labels earlier in training, thus assuming there are no clear visual patterns in the noisy samples corresponding to a particular corrupted label, and these associations are learnt later in training (Zhang et al., 2017). We instead show diverse real-world datasets contain naturally occurring *structured* noise, where in many cases there are visual patterns to the corrupted label. Large Loss Matters (Kim et al., 2022) is representative of such adaptive noise reduction methods and we find that it struggles with noisy labels extracted from in-the-wild captions.

3 Label Noise Analysis and Dataset

We analyze what makes large in-the-wild datasets a challenging source of labels for object detection.

Datasets analysed. RedCaps (Desai et al., 2021) consists of 12M Reddit image-text pairs collected from a curated set of subreddits with heavy visual content. **SBUCaps** (Ordonez et al., 2011) consists of 1 million Flickr photos with text descriptions written by their owners. Captions were selected if at least one prepositional phrase and 2 matches with a predefined vocabulary were found. Conceptual Captions (**CC**) (Sharma et al., 2018) contains 3M image-alt-text pairs after heavy post-processing; named entities in captions were hypernymized and image-text pairs were accepted if there was an overlap between Google Cloud Vision API class predictions and the caption. While less in-the-wild, it is still less clean than COCO. These datasets exhibit very low precision of the extracted labels, ranging from 0.463 for SBUCaps, 0.596 for RedCaps, to 0.737 for CC, all much lower than the 0.948 for COCO (see appx).

Extracted object labels. Given a vocabulary of object classes, we extract a label for an image if there is exact match between the object name and the corresponding caption ignoring punctuation, as in (Ye et al., 2019b; Fang et al., 2022).

Gold standard object labels. We use pseudo-ground-truth *image-level* predictions from a pretrained image recognition model to *estimate* visual presence *gold standard* labels because the in-the-wild datasets do not have object annotations. We use an object recognition ensemble with the X152-C4 object-attribute model (Zhang et al., 2021) and the Ultralytic YOLOv5-XL (Jocher et al., 2021). This ensemble achieves strong accuracy, 82.2% on SBUCaps, 85.6% on RedCaps, and 86.8% on CC (see appx). We extract VAELs by selecting images where extracted and gold-standard labels disagree. Note we never use bounding-box pseudo labels, only image-level ones. Our cross-category experiments show we do not require labels for all classes.

Noise annotations collected. To understand the label noise distribution, we select 100 VAEL examples per dataset (RedCaps, SBUCaps, CC) and annotate four types of information:

- (Q1: Label Noise) How much of the VAEL object is present (<u>vis</u>ible, <u>partial</u>ly visible, completely <u>abs</u>ent);
- (Q2: Similar Context) If the VAEL object is completely absent, whether traditionally co-occurring context ("boat" and "water"), or a semantically similar object (e.g. "cake" and "bread", "car" and "truck") is present;
- (Q3: Visual Defects) If visible/partially visible, whether the VAEL object is occluded, has

	Label noise			Similar context		Visual defects			Linguistic indicators						
Dataset	%Vis	%Part	%Abs	%Co-occ	%Sim	%Occl	%Parts	%Atyp	%Beyond	%Past	%Non-lit	%Prep	%Mod	%Sense	%Named
S	21.5	20.0	58.5	42.5	13.2	61.6	46.3	44.6	26.0	3.0	11.0	40.5	32.0	12.0	5.0
R	29.2	12.8	57.5	15.0	4.0	21.8	22.2	49.0	19.8	3.1	9.3	5.7	26.6	18.2	10.9
CC	32.8	16.6	50.5	30.9	12.8	36.3	24.2	57.3	27.6	2.6	5.7	31.3	25.0	8.3	2.1

Table 1: Label noise distributions; "other"/uncommon categories skipped. Similar context is only annotated for absent objects agreed by both annotators. Visual defects are annotated over examples with full or partial visibility. Linguistic indicators are annotated over examples with visual defects or partial/no visibility. S = SBUCaps, R = RedCaps, and CC = Conceptual Captions.

key parts missing, or atypical appearance (e.g. knitted animal); and

- (Q4: Linguistic Indicators) What linguistic cues explain why the VAEL object is mentioned but absent, e.g. the caption discusses events or information beyond what the image shows ("beyond" in Tab. 1), describes the past the extracted label is part of a prepositional phrase and likely to describe setting not objects (e.g. "on a <u>train</u>"), is a noun modifying another noun, is used in a non-literal way, has a different word sense (e.g. "bed" vs "river bed"), or is part of a named entity.

Two authors provide the annotations, with high agreement: 0.76 for Q1, 0.33 for Q2, 0.45 for Q3, and 0.58 for Q4. We calculate Cohen's Kappa for each option and aggregate agreement through a weighted average for each question, with weights derived from average option counts between the two annotators across the three datasets. We label the dataset Caption Label Noise, or CLaN.

In Table 1, we show what fraction of samples fall into each annotated category, excluding "Other", "Unclear" and uncommon categories. We average the distribution between the two annotators.

Statistics: Label noise. We first characterize the visibility of objects flagged as VAELs by the recognition ensemble. We find that SBUCaps has the highest rate of completely absent images (58.5%), followed closely by RedCaps. SBUCaps also has the highest rate of partially visible objects (20%). CC has the highest full visibility (32.8%), defined as the object from a given viewpoint having 75% or more visibility. The high rate of absent and partially-visible objects justifies the use of pseudo-ground-truth labels from the recognition ensemble; these both constitute poor training data for WSOD.

Statistics: Similar context. Certain images with absent objects may be more harmful than others. Prior work has shown that models exploit co-occurrences between an object and its context which helps overall recognition accuracy, but can hurt performance when that context is absent (Singh et al., 2020). We hypothesize the inclusion of images with this context bias without the actual object present could affect localization especially when supervising detection *implicitly*, and semantically similar context may blur decision boundaries. Different annotators may have different references for similarity or co-occurrence frequency, but our annotators achieve fair agreement ($\kappa = 0.33$). In Table 1, we find high rates of co-occurring contexts in samples with completely absent VAELs for SBUCaps (42.5%) and CC (30.9%). Across all datasets, we see a similar rate, 12%-15%, of similar context being present instead of the VAEL.

Statistics: Visual defects. We hypothesize there may be visual defects which caused the recognition ensemble to miss fully-visible objects. Over the fully or partially visible subset, in CC 79% of fully or partially visible objects have a visual defect, 87% for SBUCaps, and 69% for RedCaps. The most common defect for RedCaps and CC is atypical (49% and 57.3%); we argue atypical examples constitute poor training data for WSOD. We find the caption context (e.g. "acrylic illustration of the funny mouse") may indicate the possibility of a visual defect, further motivating the VEIL design.

Statistics: Linguistic indicators. Noun modifier is one of the most frequently occurring indicators. Prepositional phrase is also significant in SBUCaps (40.5%) and CC (31.3%). All datasets contain many VAELs mentioned in contexts going beyond the image, e.g.: "just got back from the river. friend **sank his truck pulling his <u>boat</u> out**. long story short, rip this beast" (RedCaps). We find prevalent structured noise (pattern to the images associated with a particular noisy label) for indicators like "noun modifier" and "prepositional phrase".

4 Method

Vetting labels (VEIL). The extracted label vetting task aims to predict binary visual presence targets

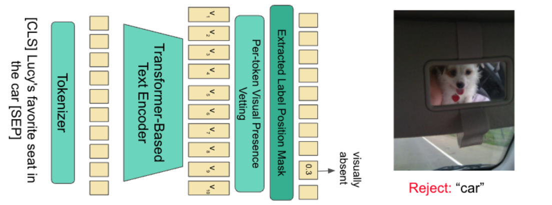

Figure 2: VEIL model architecture. After the vetting layer, the masking layer masks visual presence predictions for tokens not corresponding to an extracted label.

(present/absent) for *each* extracted label in the caption using **only** the caption context. We hypothesize there is enough signal in the caption to vet the most harmful label noise without the additional processing cost of adding the visual modality or distractions from the visual modality (similar context). The method is overviewed in Fig. 2. Given a caption, WordPiece (Wu et al., 2016) produces a sequence of subword tokens C; each token is mapped to corresponding embeddings, resulting in $e \in \mathbb{R}^{d \times C}$. These embeddings are passed through a pretrained language model (BERT (Devlin et al., 2019)), h, which includes multiple layers of multi-head self-attention over tokens in the caption to compute token-level output embeddings $v \in \mathbb{R}^{d \times C}$. An MLP is applied to these embeddings and the output is a sequence of visual presence predictions per token, $r \in [0,1]^C$.

$$v = h(e) \tag{1}$$

$$r = \sigma(W_2(\tanh(W_1 v))) \tag{2}$$

where $W_1 \in \mathbb{R}^{d \times d}$ and $W_2 \in \mathbb{R}^{1 \times d}$.

Not all predictions in r correspond to an extracted label, so we use a mask, $M \in [0,1]^C$, such that only the predictions associated with the extracted labels are used in binary cross entropy loss. To train this network, the pseudo-label targets are present, $y_i = 1$, if a pretrained image-level object recognition model also predicts the extracted label.

$$L_i = M_i \Big[y_i \log r_i + (1 - y_i) \log(1 - r_i) \Big] \tag{3}$$

$$L = \frac{1}{M^T M} \sum_{i=1}^{C} L_i \tag{4}$$

During *inference*, if an extracted label was mapped to multiple tokens (e.g. "teddy bear"), the predictions are averaged for a single presence prediction.

Weakly-supervised object detection. To test the ability of extracted label filtering or correction methods for weakly-supervised object detection, we train MIST (Ren et al., 2020). MIST extends WSDDN (Bilen and Vedaldi, 2016) and OICR (Tang et al., 2017b) which combine class scores for a large number of regions in the image to compute an image-level prediction (used for training). VEIL uses image-level pseudo training data from the in-the-wild datasets to train the vetting model, and we want to see how its ability to vet labels for WSOD generalizes to unseen data. Thus, we use the test splits of the in-the-wild datasets to train MIST, as they are unseen by all vetting methods. We do not evaluate the WSOD model on these in-the-wild datasets, but on disjoint datasets which have bounding boxes (PASCAL VOC and COCO).

5 Experiments

We show the ability of VEIL to exceed language-agnostic filtering and image-based filtering methods in extracted label vetting, to vet noisy extracted labels prior to weakly-supervised object detection training and to remove structured noise. We also benchmark the generalization ability of VEIL in cross-dataset and cross-category settings.

5.1 Experiment Details

We use three in-the-wild image-caption datasets: SBUCaps (Ordonez et al., 2011), RedCaps (Desai et al., 2021), Conceptual Captions (Sharma et al., 2018); and three crowdsourced datasets that fall into descriptive: COCO (Lin et al., 2014), VIST-DII (Huang et al., 2016)) and narrative: VIST-SIS (Huang et al., 2016). Each in-the-wild dataset and VIST are reduced to a subset of image-caption pairs where there is an substring match with a COCO category. This subset is split into 80%-20% train-test; see appx for image-caption counts. The WSOD models are trained on SBUCaps with labels vetted by different methods, and evaluated on PASCAL VOC 2007 test (Everingham et al., 2010) and COCO val 2014 (Lin et al., 2014).

5.2 Methods Compared

For VEIL, we use the convention VEIL-DatasetX to signify that VEIL is trained on the train-split of DatasetX. We group the methods we compare against into language-based, visual-based, and visual-language methods. They are category-agnostic, except for Cap2Det (Ye et al., 2019b) and Large Loss Matters (LLM) (Kim et al., 2022) which must be applied on closed vocabulary.

No Vetting accepts all extracted labels (*recall*=1).

	Method	SBUCaps	RedCaps	CC	VIST	VIST-DII	VIST-SIS	COCO	AVG
	No Vetting	0.633	0.747	0.849	0.853	0.876	0.820	**0.973**	0.822
VL	Global CLIP (Radford et al., 2021)	0.604	0.583	0.569	0.668	0.625	0.683	0.662	0.628
	Global CLIP - E (Radford et al., 2021)	0.594	0.569	0.534	0.654	0.613	0.660	0.640	0.609
V	Local CLIP (Radford et al., 2021)	0.347	0.651	0.363	0.427	0.476	0.418	0.464	0.449
	Local CLIP - E (Radford et al., 2021)	0.760	0.840	0.597	0.759	0.695	0.812	0.788	0.750
	Reject Large Loss (Kim et al., 2022)	0.667	0.790	0.831	0.782	0.794	0.743	0.896	0.786
L	Accept Descriptive	0.491	0.413	0.740	0.687	0.844	0.264	0.935	0.625
	Reject Noun Mod.	0.618	0.703	0.814	0.823	0.847	0.788	0.906	0.786
	Cap2Det (Ye et al., 2019b)	0.639	0.758	0.846	0.826	0.854	0.774	0.964	0.809
	VEIL-Same Dataset	**0.809**	**0.890**	**0.909**	0.871	**0.892**	0.816	**0.973**	**0.884**
	VEIL-Cross Dataset	0.716	0.793	0.850	**0.875**	**0.892**	**0.830**	0.958	0.842

Table 2: Extracted label vetting F1 Performance. **Bold** indicates best performance in each column, and underlined second-best. (V) signifies method uses the visual modality and (L) signifies use of language.

Global CLIP and CLIP-E use the ViT-B/32 pretrained CLIP (Radford et al., 2021) model. To enhance alignment (Hessel et al., 2021), we add the prompt "A photo depicts" to the caption and calculate the cosine similarity between the image and text embeddings generated by CLIP. We train a Gaussian Mixture Model with two components on dataset-specific cosine similarity distributions. During inference, we accept image-text pairs with predicted components aligned with higher visual-caption cosine similarity. For the ensemble variant (CLIP-E), we prepend multiple prompts to the caption, and use maximum cosine similarity.

Local CLIP and CLIP-E use cosine similarity between the image and the prompt "this is a photo of a" followed by the extracted label. Only extracted labels are filtered rather than entire captions, making this image-conditioned, not image-language conditioned vetting like Global CLIP. Local CLIP-E ensembles prompts.

Reject Large Loss. LLM (Kim et al., 2022) is a language-agnostic adaptive noise rejection and correction method. To test its vetting ability, we simulate five epochs of WSOD training (Bilen and Vedaldi, 2016) and consider label targets with a loss exceeding the large loss threshold as "predicted to be visually absent" after the first epoch. The threshold uses a relative delta controlling the rejection rate (set as 0.002 in (Kim et al., 2022)).

Accept Descriptive. We train a logistic regression model to predict whether a VIST (Huang et al., 2016) caption comes from the DII (descriptive) or SIS (narrative) split. The input vector to this logistic regression model consists of part of speech tags (e.g. proper noun, adjective, verb - past tense, etc) present in the caption. We accept extracted labels from captions with descriptiveness over 0.5.

Reject Noun Mod. Since an extracted label could be modifying another noun ("car park"), a simple baseline is to reject an extracted label if the POS label is an adjective or is followed by a noun.

Cap2Det. We reject a label if it is not predicted by the Cap2Det (Ye et al., 2019b) classifier.

5.3 Extracted Label Vetting Evaluation

VEIL selects cleaner labels compared to no vetting and other methods, even when not trained on target data. Tab. 9 shows the F1 score which combines the precision and recall of their vetting (shown separately in appx). Most language-based methods improve or maintain the F1 score of No Vetting, even though it has perfect recall, except Accept Descriptive. Rule-based methods and Cap2Det perform strongly, but are outperformed by both VEIL-Same Dataset (trained and tested on the same dataset) and VEIL-Cross Dataset (trained on a different dataset than that shown in the column; we show the best cross-dataset result). VEIL-Cross Dataset outperforms other language-based approaches, showing VEIL's generalization potential, except on COCO where Cap2Det does slightly better. Image-and-language-conditioned approaches (Global CLIP/CLIP-E) make label decisions based on the overall caption, so certain language can affect the alignment even if the object is actually visually present. Among image-based approaches for label vetting, Local CLIP benefits significantly from using an ensemble of prompts compared to Global CLIP; ensembling is well documented in improving zero-shot image recognition in prior work (Radford et al., 2021). Reject Large Loss has the strongest F1 score among the image-based methods, but worse than VEIL.

Using CLaN, we find that VEIL is stronger than CLIP-based vetting at rejecting different

Data	Vetting Method	Label noise		Similar context		Visual defects			Linguistic indicators					
		%Part	%Abs	%Co-occ	%Sim	%Occl	%Parts	%Atyp	%Mod	%Prep	%Non-lit	%Sense	%Named	%Beyond
SBUCaps	VEIL-Same Dataset	**85.0**	**94.7**	**87.0**	**80.0**	**81.1**	**90.6**	**87.2**	**95.2**	**93.9**	**90.6**	**100.0**	**100.0**	**88.8**
	LocalCLIP-E	51.5	80.7	71.3	70.0	52.7	52.1	65.6	63.8	70.6	82.9	96.2	62.5	82.4
RedCaps	VEIL-Same Dataset	**91.7**	74.1	**71.4**	**85.7**	**83.3**	**89.0**	**68.3**	**74.8**	**90.0**	66.7	**88.9**	80.9	76.3
	LocalCLIP-E	52.8	**78.4**	40.0	38.1	47.0	45.0	23.2	68.4	63.3	**70.8**	70.6	**90.0**	**76.7**
CC	VEIL-Same Dataset	**60.6**	83.0	**81.2**	55.0	**54.9**	**53.6**	**56.3**	64.2	**73.7**	81.7	**100.0**	-	77.4
	LocalCLIP-E	45.0	**89.1**	74.9	**57.5**	49.9	50.0	24.1	**73.3**	63.9	**91.7**	**100.0**	-	**86.8**

Table 3: VAEL recall on CLaN. Bold indicates best performance per column/dataset. We omit named entity results for CC as it substitutes them with predefined categories (e.g. person, org.).

forms of label noise. Captions alone contain cues about noise. We hypothesize that LocalCLIP-E would do well at vetting VAELs explained by linguistic cues like non-literal and beyond the image as they are likely to have low image-caption cosine similarity. We also hypothesize that VEIL would do better than LocalCLIP-E at vetting VAELs that are noun modifiers or in prepositional phrases, which can be easily picked up from the caption. Further, similar context can sometimes be explained by noun modifiers and prepositional phrases, but LocalCLIP-E may be oblivious to the context differing from the VAEL category. We evaluate these hypotheses on the CLaN dataset in Tab. 3. We omit "visible" VAEL samples as these may be pseudo-label errors, and the "past" linguistic indicator due to too few samples. We find that VEIL vets truly absent objects for SBUCaps much better than LocalCLIP-E, and comparably for RedCaps or CC. It vets partially visible objects better than LocalCLIP-E by a significant margin; these can be harmful in WSOD which is already prone to part domination (Ren et al., 2020). VEIL also recognizes that similar context to, rather than the actual VAEL category, are present. VEIL performs better at vetting visible objects that have visual defects which can be mentioned in caption context ("acryllic illustration of dog"). As expected, we find that for all datasets, VEIL vets VAELs from prepositional phrases better than LocalCLIP-E, and noun modifiers for SBUCaps and RedCaps. LocalCLIP-E does better on "beyond the image" and non-literal VAELs except on SBUCaps where VEIL excels.

VEIL generalizes across training sources and is complementary to CLIP-based vetting. We train VEIL on one dataset (or multiple) and evaluate on an unseen target. We find that combining multiple sources improves precision (Tab. 4). We also try ensembling by averaging predictions between LocalCLIP-E and VEIL-Cross Dataset, and find that its precision and recall is highest among the VEIL variants and LocalCLIP-E. This means

Method	Train Dataset	Prec/Rec	F1
No Vetting	-	0.463 / 1.000	0.633
VEIL	SBUCaps	0.828 / 0.791	0.809
VEIL	RedCaps (R)	0.668 / 0.759	0.710
VEIL	CC	0.585 / 0.846	0.692
VEIL	R, CC	0.689 / 0.722	0.705
LCLIP-E	WIT	0.708 / 0.820	0.760
VEIL+LCLIP-E	R,CC,WIT	0.733 / 0.848	0.786

Table 4: Source generalization of VEIL; vet on SBUCaps. LCLIP-E is LocalCLIP-E. CLIP trained on WIT.

Method	Prec/Rec	F1
No Vetting	0.323 / 1.000	0.488
ID	0.651 / 0.656	0.654
OOD	0.585 / 0.556	0.570

Table 5: VEIL category generalization on SBUCaps.

that VEIL and LocalCLIP-E can be used together. There is still a significant gap between VEIL-Same Dataset and even the ensembled model in terms of precision and F1. We leave improving source generalizability to future research.

VEIL produces cleaner labels even on unseen object categories. We define an in-domain category set (ID) of 20 randomly picked categories from COCO (Lin et al., 2014), and an out-of-domain category set (OOD) consisting of the 60 remaining categories. We restrict the labels using these limited category sets and create two train subsets, ID and OOD from SBUCaps *train* and one ID test subset from SBUCaps *test*. We find that transferring VEIL-OOD to unseen categories improves F1 score compared to no vetting as shown in Table 5. We hypothesize training on more categories could improve category generalization, but leave further experiments to future research.

5.4 Impact on Weakly Sup. Object Detection

We select the most promising vetting methods from the previous section and use them to vet labels from an in-the-wild dataset's, SBUCaps, unseen (*test*) split and then train WSOD models using the vetted labels. Then, these WSOD models are evaluated on detection benchmarks like VOC-07 and COCO-

Method	VOC Det. mAP_{50}	VOC Rec. mAP	COCO Det mAP_{50}
GT* (upper bound)	40.0	69.0	9.2
No Vetting	31.2	65.3	7.7
Large Loss (Kim et al., 2022)	30.9	65.3	7.5
LocalCLIP-E (Radford et al., 2021)	37.1	70.7	7.9
VEIL-R,CC	37.8	71.4	8.6
VEIL-SBUCaps	**40.5**	**74.3**	**10.4**

Table 6: Impact of vetting on WSOD performance on VOC-07 and COCO-14. (GT*) directly vets labels using the pretrained recognition models used to train VEIL.

14. We show two different VEIL methods, VEIL-SBUCaps and VEIL-RedCaps,CC to demonstrate the generalizability of VEIL on WSOD. Note that Large Loss Matters (Kim et al., 2022) has been relaxed to *correct* visually absent extracted labels, in addition to unmentioned but present objects (false negatives). After vetting, we remove any images without labels and since category distribution follows a long-tail distribution, we apply weighted sampling (Mikolov et al., 2013). We train MIST (Ren et al., 2020) for 50K iter. with batch size 8.

VEIL vetting leads to better detection and recognition capabilities than vetting through CLIP, an adaptive label noise correction method (Large Loss Matters) or even directly using its bootstrapped data. We find that VEIL-SBUCaps performs the best as shown in Tab. 6. In particular, it boosts the detection performance of No Vetting by 9.3% absolute and 29.8% relative gain (40.5/31.2% mAP) on VOC-07 and by 35% relative gain (10.4/7.7% mAP) on COCO. Interestingly, VEIL-SBUCaps and VEIL-Redcaps,CC have a similar performance improvement, despite VEIL-Redcaps,CC (best VEIL cross-dataset result on SBUCaps) having poorer performance than Local CLIP-E in Tab. 4. Additionally, directly using predictions from the pretrained object recognition model (used to produce visual presence targets for VEIL at the image level) to vet (GT* method in the table) performs worse than VEIL in both detection and recognition showing **VEIL's generalization from its bootstrapped data**.

Structured noise negatively impacts localization. Using the CLaN dataset, we observe one type of structured noise found from extracting labels from prepositional phrases, specifically where images were taken inside vehicles. We hypothesize such structured noise would have significant impact on localization for the vehicle objects. We use CorLoc to estimate the localization ability for vehicles in VOC-07 (“aeroplane", ‘bicycle", “boat", “car", “bus", “motorbike", “train"). We observe a CorLoc of 60.2% and 54.1% for VEIL-SBUCaps and LocalCLIP-E, respectively. This shows structured noise can have strong impact on localization.

Clean Labels	Noisy Labels	WS	Vetting	mAP_{50}
✓			n/a	43.48
✓	✓			42.06
✓	✓		✓	51.31
✓	✓	✓	✓	54.76

Table 7: Mixed supervision from clean (VOC-07 train-val) and noisy labels (SBUCaps). Eval on VOC-07 test.

Naively mixing clean and noisy samples without vetting for WSOD leads to worse performance than only using clean samples. Vetting in-the-wild samples (noisy) with VEIL is essential to improving performance. We study how vetting impacts a setting where labels are drawn from both annotated image-level labels from 5K VOC-07 train-val (Everingham et al., 2010) (clean) and 50K in-the-wild SBUCaps (Ordonez et al., 2011) captions (noisy). In Tab. 7 we observe that naively adding noisy supervision to clean supervision actually hurts performance compared to only using clean supervision. After vetting the labels extracted from SBUCaps (Ordonez et al., 2011) using VEIL-SBUCaps, we observe that the model sees a 17.9% relative improvement (51.31/43.48% mAP) to using only clean supervision from VOC-07. We see further improvements when applying weighted sampling (WS) to the added, class imbalanced data (54.76/51.31% mAP).

VEIL improves WSOD performance even at scale. We sampled the held-out RedCaps dataset in increments of 50K samples up to a total of 200K samples. For each scale, we train two WSOD models with weighted sampling using the unfiltered samples and those vetted with VEIL-SBUCaps,CC. The mAP at 50K, 100K, 150K, and 200K samples is 4.2, 10.7, 12.0, 12.9 with vetting and 1.9, 8.2, 10.6, 10.4 without vetting. The non-vetted model's performance declines after 150K samples. This indicates vetting can adapt to scale better even when VEIL is trained on other datasets. The trend suggests that vetting will continue outperforming no-vetting even when dataset sizes increase.

Conclusion. We showed visually absent extracted labels are common in the wild, VEIL which uses language context to infer if mentioned objects are visually present, and the benefits of its vetting.

References

Hakan Bilen and Andrea Vedaldi. 2016. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854.

Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guo Chun Li. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14064–14073.

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

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yu Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. 2022. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*.

Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. 2022. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*.

Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. 2019. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Joseph Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing*.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

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

Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, Tkianai, YxNONG, Adam Hogan, Lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, , Marc, Yonghye Kwon, , Oleg, Wanghaoyang0106, Yann Defretin, Aditya Lohia, Ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, , Doug, Durgesh, and Francisco Ingham. 2021. ultralytics/yolov5: v5.0 - yolov5-p6 1280 models, aws, supervise.ly and youtube integrations.

Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwook Lee. 2022. Large loss matters in weakly supervised multi-label classification. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14136–14145.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Dhruv Kumar Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *ECCV*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *NIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Shafin Rahman, Salman Hameed Khan, and Nick Barnes. 2020a. Improved visual-semantic alignment for zero-shot object detection. In *AAAI Conference on Artificial Intelligence*.

Shafin Rahman, Salman Hameed Khan, and Fatih Murat Porikli. 2020b. Zero-shot object detection: Joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128:2979 – 2999.

Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. 2020. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Data Centric AI NeurIPS Workshop 2021*, abs/2111.02114.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. 2022. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9601–9610.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. 2020. Don't judge an object by its context: Learning to overcome contextual bias. page 11070–11078.

Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017a. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. 2017b. Multiple instance detection network with online instance classifier refinement. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3059–3067.

Mesut Erhan Unal, Keren Ye, Mingda Zhang, Christopher Thomas, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. 2022. Learning to overcome noise in weak caption supervision for object detection. *IEEE transactions on pattern analysis and machine intelligence*, PP.

Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. 2023. 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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. 2019a. Cap2det: Learning to amplify weak caption supervision for object detection. In *International Conference on Computer Vision (ICCV)*.

Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. 2019b. Cap2det: Learning to amplify weak caption supervision for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, page 9685–9694.

Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. 2022. Regionclip: Region-based language-image pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022,* *New Orleans, LA, USA, June 18-24, 2022*, pages 16772–16782. IEEE.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krahenbuhl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*.

A Appendix

We provide supplemental materials to our main text.

First, we present additional dataset details. Then, we provide a detailed table of the vetting precision and recall of all methods described in the main text, for which we show F1 performance in Table 9 of the main text. Furthermore, we show more comprehensive cross-dataset ablations, such as adding more training datasets and training with a special token.

We discuss our hyperparameter selection for WSOD in further detail and show additional metrics of the WSOD models on the COCO-14 benchmark presented in the main text.

Finally, we showcase the vetting ability of VEIL in comparison to other approaches through qualitative results, along with additional examples from the WSOD models trained using vetted training data.

A.1 Vetting Dataset Details

Dataset	Train	Test
VIST	20339	5086
VIST-DII	12106	3028
VIST-SIS	8233	2060
COCO	216096	94004
SBUCaps	166986	41747
RedCaps	845333	211334
CC	350043	87511

Table 8: The number of samples per split and dataset after filtering captions based on exact match with COCO objects. Note VIST and COCO have multiple captions per image; for the sake of vetting, we evaluate on extracted labels from all captions.

While the overall image-text pairs are 12M pairs for RedCaps, 3M pairs for CC, 1M for SBUCaps, 500K pairs for COCO, 40K and 60K pairs for VIST-DII and VIST-SIS, respectively, after extracting labels using exact match with COCO categories, there are a number of captions which don't have any matches. We filter out those captions. In Table 8 we provide counts after filtering for both vetting train and test splits of each dataset.

	Prepositional Phrase, Co-Occurring Context Took this on a **boat** the other day.	**Different Word Sense, Semantic Similarity:** [...] put a few drops of **orange** oil on it for good	**Non-Literal:** did a wen trim for the first time and it was a piece of **cake**
Extracted Labels from Each Image	{boat}	{orange}	{cake}
No Vetting (Same as Above)	{boat}	{orange}	{cake}
Reject Noun Modifier	{boat}	{}	{cake}
LocalCLIP-E	{boat}	{orange}	{}
VEIL-SBUCapsCC	{}	{}	{}
VEIL-RedCaps	{}	{}	{}

Figure 3: Qualitative examples of extracted labels after vetting on RedCaps-Test. These are additional completely absent VAEL examples from CLaN with their linguistic indicators and similar context annotations, and only VEIL-based methods are able to overcome these three noise types.

A.2 Vetting Precision/Recall

Table 9 in the main text showed the F1 on the extracted label vetting task, from twelve methods. In Table 9 here, we separately show Precision and Recall on the same task.

A.3 Cross-Dataset Ablations

Table 10 is included as reference which shows that precision in the cross dataset setting is always better than no vetting with the exception of COCO.

Combining multiple datasets. We find that VEIL is able to leverage additional datasets to an extent. For example, combining SBUCaps and CC leads to significant improvements (7-16% relative) in F1 as shown in Table 11 and, combining SBUCaps and Redcaps in training improves performance on both validation sets. When combining all datasets, only the non-in the wild datasets see an improved performance.

Using special token. We test VEIL_{ST} which inserts a special token `[EM_LABEL]` before each extracted label in the caption to reduce the model's reliance on category-specific cues and improve generalization to other datasets. We find that using VEIL w/ ST on average improves F1 by 1 pt compared to just VEIL when transferring to other datasets. This comes at a tradeoff with respect to the performance on the same dataset; however CC w/ ST improves performance on all datasets.

A.4 WSOD Implementation Details

We used 4 RTX A5000 GPUs and trained for 50k iterations with a batch size of 8, or 100k iterations on 4 Quadro RTX 5000 GPUs with a batch size of 4 and gradient accumulation (parameters updated

	Method	SBUCaps PREC / REC	SBUCaps F1	RedCaps PREC / REC	RedCaps F1	Conceptual Captions PREC / REC	Conceptual Captions F1
	No Vetting	0.463 / 1.000	0.633	0.596 / 1.000	0.747	0.737 / 1.000	0.849
VL	Global CLIP (Radford et al., 2021)	0.531 / 0.700	0.604	0.618 / 0.551	0.583	0.753 / 0.458	0.569
	Global CLIP - E (Radford et al., 2021)	0.526 / 0.683	0.594	0.625 / 0.522	0.569	0.745 / 0.417	0.534
V	Local CLIP (Radford et al., 2021)	0.588 / 0.246	0.347	0.723 / 0.591	0.651	0.750 / 0.240	0.363
	Local CLIP - E (Radford et al., 2021)	0.708 / 0.820	0.760	0.770 / 0.924	0.840	0.842 / 0.462	0.597
	Reject Large Loss (Kim et al., 2022)	0.530 / **0.898**	0.667	0.700 / 0.908	0.790	0.806 / 0.858	0.831
L	Accept Descriptive	0.449 / 0.542	0.491	0.561 / 0.326	0.413	0.739 / 0.741	0.740
	Reject Noun Mod.	0.517 / 0.769	0.618	0.644 / 0.776	0.703	0.765 / 0.870	0.814
	Cap2Det (Ye et al., 2019a)	0.500 / 0.884	0.639	0.633 / **0.945**	0.758	0.758 / **0.956**	0.846
	VEIL-Same Dataset	**0.828** / 0.791	**0.809**	**0.855** / 0.929	**0.890**	**0.884** / 0.935	**0.909**
	VEIL-Cross Dataset	0.636 / 0.811	0.713	0.747 / 0.847	0.793	0.834 / 0.866	0.850

	Method	VIST PREC / REC	VIST F1	VIST-DII PREC / REC	VIST-DII F1	VIST-SIS PREC / REC	VIST-SIS F1
	No Vetting	0.744 / 1.000	0.853	0.779 / 1.000	0.876	0.695 / 1.000	0.820
VL	Global CLIP (Radford et al., 2021)	0.772 / 0.589	0.668	0.788 / 0.518	0.625	0.754 / 0.624	0.683
	Global CLIP - E (Radford et al., 2021)	0.769 / 0.569	0.654	0.785 / 0.504	0.613	0.741 / 0.595	0.660
V	Local CLIP (Radford et al., 2021)	0.752 / 0.298	0.427	0.787 / 0.341	0.476	0.738 / 0.292	0.418
	Local CLIP - E (Radford et al., 2021)	**0.874** / 0.671	0.759	**0.886** / 0.572	0.695	**0.833** / 0.793	0.812
	Reject Large Loss (Kim et al., 2022)	0.755 / 0.811	0.782	0.792 / 0.796	0.794	0.700 / 0.791	0.743
L	Accept Descriptive	0.755 / 0.631	0.687	0.784 / 0.913	0.844	0.686 / 0.163	0.264
	Reject Noun Mod.	0.775 / 0.879	0.823	0.813 / 0.883	0.847	0.716 / 0.875	0.788
	Cap2Det (Ye et al., 2019a)	0.781 / 0.877	0.826	0.823 / 0.887	0.854	0.704 / 0.859	0.774
	VEIL-Same Dataset	0.789 / **0.971**	0.871	0.819 / **0.992**	**0.892**	0.690 / **0.998**	0.816
	VEIL-Cross Dataset	0.835 / 0.920	**0.875**	0.870 / 0.915	**0.892**	0.765 / 0.920	**0.830**

	Method	COCO PREC / REC	COCO F1
	No Vetting	0.948 / 1.000	**0.973**
VL	Global CLIP (Radford et al., 2021)	0.945 / 0.509	0.662
	Global CLIP - E (Radford et al., 2021)	0.931 / 0.487	0.640
V	Local CLIP (Radford et al., 2021)	0.951 / 0.307	0.464
	Local CLIP - E (Radford et al., 2021)	0.972 / 0.663	0.788
	Reject Large Loss (Kim et al., 2022)	0.963 / 0.837	0.896
L	Accept Descriptive	0.948 / 0.923	0.935
	Accept Narrative	0.942 / 0.077	0.143
	Reject Noun Mod.	0.958 / 0.859	0.906
	Cap2Det (Ye et al., 2019a)	**0.978** / 0.950	0.964
	VEIL-Same Dataset	0.948 / **1.000**	**0.973**
	VEIL-Cross Dataset	0.975 / 0.942	0.958

Table 9: Extracted Label Vetting Evaluation Metrics. Bold indicates best result in column, and in the recall columns No Vetting is excluded as it always has perfect recall.

Train Dataset(s)	ST	DII-VIST	SIS-VIST	COCO	VIST	SBUCaps	RedCaps	CC
No Vetting		0.779 / 1.000	0.695 / 1.000	0.948 / 1.000	0.741 / 1.000	0.463 / 1.000	0.596 / 1.000	0.737 / 1.000
SBUCaps		0.895 / 0.717	0.831 / 0.609	0.979 / 0.647	0.878 / 0.690	0.828 / 0.791	0.808 / 0.684	0.844 / 0.831
RedCaps (R)		0.865 / 0.794	0.787 / 0.752	0.975 / 0.824	0.839 / 0.785	0.668 / 0.759	0.855 / 0.929	0.837 / 0.709
CC		0.863 / 0.902	0.759 / 0.917	0.974 / 0.925	0.824 / 0.914	0.585 / 0.846	0.713 / 0.844	0.884 / 0.935
VIST		0.826 / 0.978	0.729 / 0.949	0.958 / 0.926	0.789 / 0.971	0.518 / 0.939	0.658 / 0.883	0.771 / 0.981
COCO		0.779 / 1.000	0.695 / 1.000	0.948 / 1.000	0.741 / 1.000	0.463 / 1.000	0.599 / 1.000	0.739 / 1.000
SBUCaps,CC		0.885 / 0.840	0.788 / 0.837	0.978 / 0.893	0.847 / 0.838	0.923 / 0.950	0.762 / 0.822	0.965 / 0.978
R,CC		0.876 / 0.888	0.801 / 0.784	0.976 / 0.918	0.855 / 0.852	0.691 / 0.720	0.845 / 0.836	0.892 / 0.914
SBUCaps,R		0.876 / 0.779	0.789 / 0.697	0.976 / 0.791	0.849 / 0.758	0.892 / 0.940	0.923 / 0.958	0.846 / 0.785
SBUCaps	✓	0.885 / 0.798	0.817 / 0.719	0.977 / 0.745	0.866 / 0.768	0.790 / 0.814	0.782 / 0.754	0.834 / 0.866
R	✓	0.880 / 0.744	0.809 / 0.697	0.976 / 0.776	0.856 / 0.721	0.686 / 0.724	0.843 / 0.901	0.831 / 0.526
CC	✓	0.868 / 0.913	0.765 / 0.920	0.975 / 0.942	0.835 / 0.920	0.609 / 0.841	0.721 / 0.862	0.922 / 0.955
SBUCaps,CC	✓	0.870 / 0.915	0.776 / 0.881	0.976 / 0.932	0.830 / 0.905	0.754 / 0.821	0.747 / 0.847	0.891 / 0.943
R,CC	✓	0.862 / 0.922	0.779 / 0.842	0.971 / 0.944	0.837 / 0.894	0.649 / 0.797	0.793 / 0.887	0.868 / 0.931
SBUCaps,R	✓	0.877 / 0.807	0.805 / 0.712	0.973 / 0.856	0.844 / 0.828	0.826 / 0.724	0.804 / 0.905	0.839 / 0.771
ALL		0.860 / 0.969	0.779 / 0.903	0.973 / 0.990	0.832 / 0.947	0.713 / 0.829	0.803 / 0.898	0.874 / 0.941

Table 10: Cross Dataset Vetting Precision and Recall Performance on visual presence validations sets from different sources (DII-VIST...CC). All methods improve precision compared to no vetting.

Train Dataset	ST	DII-VIST	SIS-VIST	COCO	VIST	SBUCaps	RedCaps	CC
No Vetting		0.876	0.820	0.973	0.851	0.633	0.747	0.849
SBUCaps		0.796	0.703	0.779	0.773	**0.809**	0.741	0.837
R		0.828	0.769	0.893	0.811	**0.710**	**0.890**	0.768
CC		**0.882**	**0.830**	0.949	**0.867**	**0.692**	**0.773**	**0.909**
VIST		**0.895**	**0.825**	0.942	**0.871**	**0.668**	**0.754**	**0.863**
COCO		0.876	0.820	0.973	0.851	0.633	**0.749**	**0.850**
SBUCaps,CC		0.862	0.812	0.933	0.843	**0.937**	**0.791**	**0.972**
R,CC		**0.882**	0.793	0.946	**0.854**	**0.705**	**0.841**	**0.903**
SBUCaps,R		0.825	0.741	0.874	0.801	**0.915**	**0.940**	0.810
SBUCaps	✓	0.839	0.765	0.846	0.814	**0.802**	**0.767**	**0.850**
R	✓	0.806	0.749	0.865	0.783	**0.705**	**0.871**	0.644
CC	✓	**0.890**	**0.836**	0.958	**0.875**	**0.707**	**0.785**	**0.938**
SBUCaps,CC	✓	**0.892**	**0.825**	0.954	**0.866**	**0.786**	**0.793**	**0.916**
R,CC	✓	**0.891**	0.809	0.957	**0.865**	**0.716**	**0.837**	**0.899**
SBUCaps,R	✓	0.841	0.756	0.911	0.836	**0.772**	**0.851**	0.803
ALL		**0.911**	**0.836**	**0.981**	**0.886**	**0.767**	**0.848**	**0.906**

Table 11: Cross Dataset Vetting F1 Performance on visual presence validations sets from different sources (DII-VIST...CC). Bold indicates if result is better than no vetting. Train data containing the same source as the validation is highlighted in yellow.

	mAP, IoU			mAP, Area		
	0.5:0.95	0.5	0.75	S	M	L
GT*	4.19	9.17	3.40	1.10	4.34	6.76
No Vetting	3.24	7.70	2.37	1.06	4.00	5.08
Large Loss (Kim et al., 2022)	3.11	7.54	2.15	0.92	3.80	4.88
LocalCLIP-E (Radford et al., 2021)	3.66	7.77	3.08	0.79	3.96	5.96
VEIL$_{ST}$-R,CC	3.90	8.60	3.14	0.93	4.25	6.28
VEIL-SBUCaps	**4.89**	**10.37**	**4.20**	**1.26**	**5.24**	**7.53**

Table 12: COCO-14 benchmark for WSOD models trained with various vetting methods. (GT*) directly vets labels using the pretrained object detectors which were used to train VEIL. Bold indicates best performance in each column and underline indicates second best result in the column.

every two iterations to simulate a batch size of 8). **Learning Rates.** We trained four models without vetting on SBUCaps with learning rates from '1e-5' till '1e-2', for each order of magnitude, and observed that the model trained with a learning rate of '1e-2' had substantially better Pascal VOC-07 detection performance and used this learning rate for all the WSOD models trained on SBUCaps. We applied a similar learning rate selection method for WSOD models trained on RedCaps, except we tested over every half order of magnitude and found that '5e-5' was optimal when training on RedCaps.

Relative Delta. In Large Loss Matters (LLM) (Kim et al., 2022), relative delta controls how fast the rejection rate will increase over training. To find the best relative delta, we tested over three initializations, with $rel\_delta = 0.002$ as the setting recommended in (Kim et al., 2022). We used the best result in Table 13 when reporting results in the main paper.

Relative Delta	Pascal VOC-07 mAP$_{50}$
0.002	28.25
0.01	30.93
0.05	28.11

Table 13: Relative delta hyperparameter ablation

A.5 WSOD Benchmarking on Additional COCO Metrics

In our main text we compared the average precision of the model across all the classes and all the IoU (Intersection over Union) thresholds from 0.5 to 0.95. We show mAP at specific thresholds 0.5

and 0.75 in Table 12. We see that cross dataset VEIL vetting performs relatively 32% better than no vetting in a stricter IoU (0.75). The mAP metric can be further broken down by area sizes of ground truth bounding boxes, which is denoted by S, M, and L. VEIL-based vetting outperforms the rest in Medium (6% better than best non-VEIL vetting) and Large objects (5% better than best non-VEIL vetting); while VEIL-Same Dataset still performs best on small objects, VEIL-Cross Dataset performs slightly worse than no vetting.

A.6 Additional Qualitative Results

Vetting Qualitative Examples. Using annotations from CLaN, we provide qualitative examples comparing the vetting capability of methods on VAELs with common linguistic indicators (prepositional phrase, different word sense, non-literal) found in RedCaps in Figure 3.

WSOD Qualitative Examples. In Figure 4, we present further qualitative evidence on the impact of different vetting methods on weakly supervised object detection. There are varying degrees of part and contextual bias from all methods; however, No Vetting has the most pronounced part domination and context bias as shown by its detection of bicycle wheels and car doors (top two rows), and misidentifying a child as a chair (bottom row) and detections covering both boat and water. Both VEIL methods outperform the rest of the models in detecting smaller objects (see first two rows). LocalCLIP-E misses smaller objects in the background (first two rows) and also has part domination (bicycle).

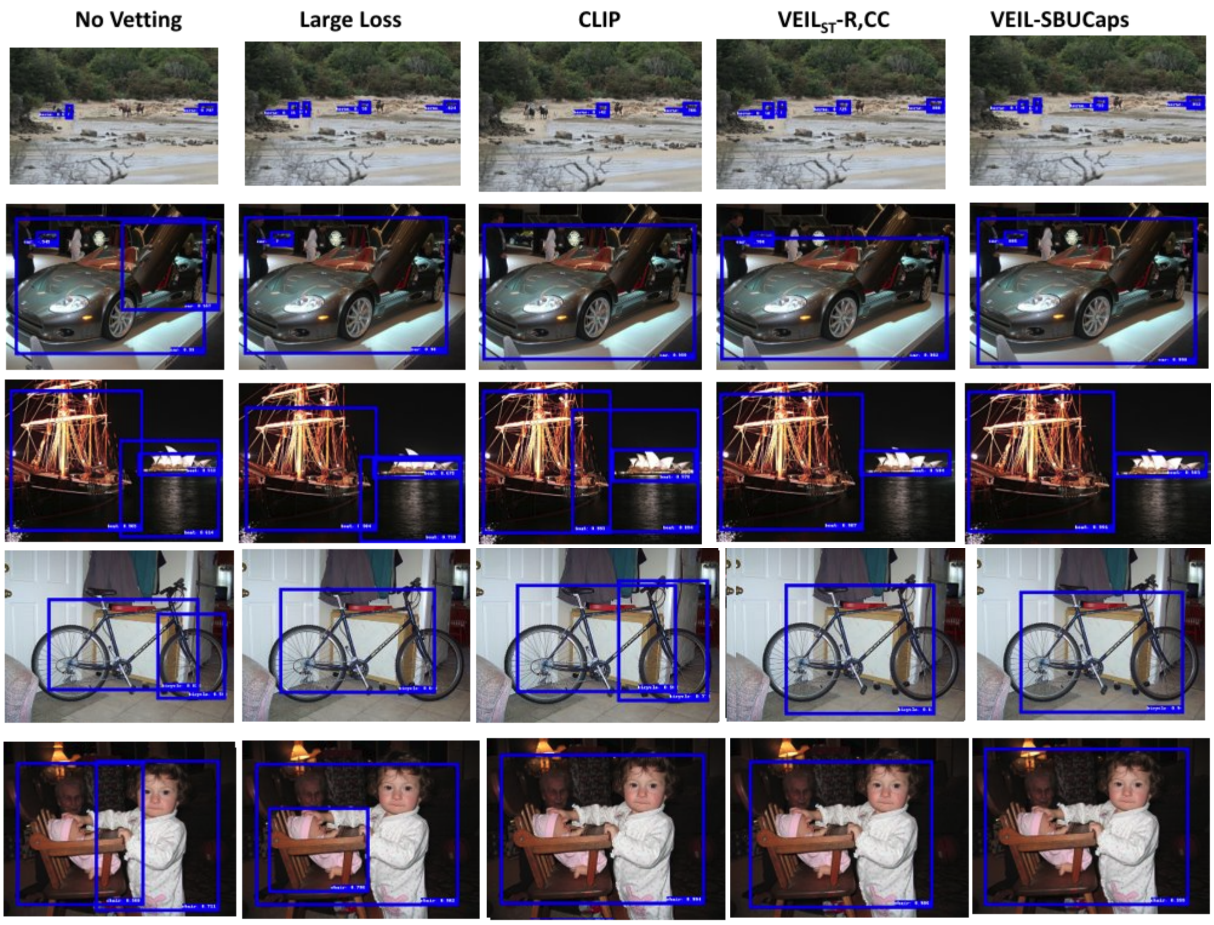

Figure 4: Detections (blue bounding box) from WSOD models trained with various vetting methods (top row) indicate that training with either VEIL-based vetting method (two rightmost columns) leads to similar detection capability on VOC-07 (Everingham et al., 2010). The categories shown by row (from top to bottom) are: horse, car, boat, bicycle, chair.