On Guiding Visual Attention with Language Specification

Suzanne Petryk* Lisa Dunlap*
Keyan Nasseri Joseph Gonzalez Trevor Darrell Anna Rohrbach

UC Berkeley

Abstract

While real world challenges typically define visual categories with language words or phrases, most visual classification methods define categories with numerical indices. However, the language specification of the classes provides an especially useful prior for biased and noisy datasets, where it can help disambiguate what features are task-relevant. Recently, large-scale multimodal models have been shown to recognize a wide variety of high-level concepts from a language specification even without additional image training data, but they are often unable to distinguish classes for more fine-grained tasks. CNNs, in contrast, can extract subtle image features that are required for fine-grained discrimination, but will overfit to any bias or noise in datasets. Our insight is to use high-level language specification as advice *for constraining the classification evidence to task-relevant features, instead of distractors. To do this, we ground task-relevant words or phrases with attention maps from a pretrained large-scale model. We then use this grounding to supervise a classifier's spatial attention away from distracting context. We show that supervising spatial attention in this way improves performance on classification tasks with biased and noisy data, including ∼3−15% worst-group accuracy improvements and ∼41−45% relative improvements on fairness metrics.*

1. Introduction

When trained with limited or biased data, visual models often learn unwanted correlations. For example, consider building a classifier to distinguish two fine-grained categories of birds: "landbird" and "waterbird". The background features from their corresponding habitats such as forests or beaches might be highly or perfectly correlated with the numerical class labels. A baseline model may mistakenly learn the unintended "location" task instead of the actual task, and fail on examples of birds out of their usual habitat (Fig. 1). However, knowledge that the task is *about birds* can disambiguate what the model is meant to learn.

*Equal contribution.

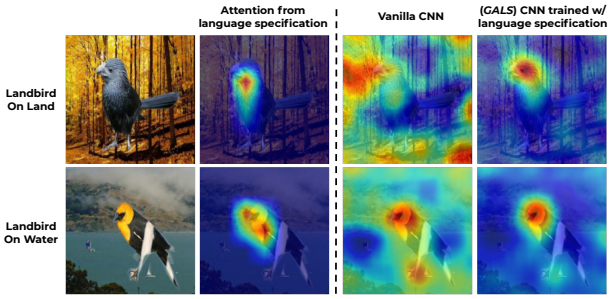


Figure 1. **Guiding attention with language.** Sample attention from the Waterbirds biased dataset. During training, Landbirds mostly appear on land backgrounds and Waterbirds mostly appear on water backgrounds. At testing, each class appears equally on land or on water. A CNN trained on this task learns to look at the background, but if we use a multimodal model to translate the language specification *"a photo of a bird"* into spatial supervision, we can ensure that our CNN learns task-relevant features.

Previous work has considered incorporating knowledge of the task as *language specifications* in the form of class names or class descriptions which can directly serve as a prior over visual model parameters [33, 47]. Several zero-shot methods condition models on attribute labels [17, 26, 49] (e.g., beak shape or wing color) or class descriptions [7, 19, 32, 54] (e.g., from Wikipedia) to enable transfer to unseen classes. However, this relies on the language specification being class-discriminative – an assumption which does not hold for some real-world tasks where only high-level task specification is given (e.g., in Fig. 1, we may only know that this is a "bird" dataset, without the class names being provided or even existing yet). Additionally, simply conditioning on language embeddings may not prevent a model from attending to spurious correlations in biased datasets.

Even when language specifications *are* class-discriminative, such models will perform poorly when there is insufficient image and text data to learn a multimodal model for rare or fine-grained classes (e.g., a large-scale model such as CLIP [33] may not have seen enough examples of the relatively rare "landbird" or "waterbird" classes during pretraining to have good zero-shot performance).

To address these limitations, we propose a new framework called **Guiding visual Attention with Language Specification**, or ***GALS***, in which we translate available language specification provided by the task metadata into spatial attention that is used to supervise a CNN's attention during training. Fig. 1 displays how *GALS* is able to pull the model's attention away from the distractor features while retaining enough flexibility to pick up on fine-grained features which were not captured by the multimodal model.

Specifically, we first leverage an off-the-shelf pretrained vision-language model to *ground* textual information into each given image and obtain a respective saliency map. This is efficient and involves no additional overhead (i.e., no need for training or per-instance manual annotation). Next, we aim to leverage the obtained saliency map to inform the visual classifier. To do this, we *guide* the classifier's attention towards the area highlighted by the saliency from the language specification. Finally, the visual classifier still needs to solve the more fine-grained task, after obtaining the high-level attention guidance. It thus retains some flexibility, e.g. it may even attend to some useful (non-harmful) context. In practice, we use the recent powerful CLIP [33] model to ground textual information into images. We leverage the "Right for the Right Reasons" method [37] to enforce that the classifier indeed attends according to the given guidance. With this approach, we can incorporate language specification via an auxiliary loss during training, and thus the *vision-language model is not needed during inference*.

We show how *GALS* can assist in training on data with explicit and implicit bias. On the synthetic Waterbirds dataset [39] which contains a known, explicit bias (the image backgrounds), our method is able to achieve $\sim$2$-$7% per-group accuracy improvements over baselines, including a model which uses an unsupervised attention mechanism instead of guidance from language. *GALS* also shows a 15% improvement on the worst-group accuracy in the challenging scenario where class labels are perfectly correlated with the distracting backgrounds (Sec. 4.2). For implicit bias, where training and test distributions differ in unknown ways, we see that *GALS* achieves $\sim$41$-$45% relative improvements on fairness metrics for apparent gender recognition (Sec. 4.4). We also show a 2% accuracy improvement on a red-meat classification task from a subset of Food-101 [2], where an implicit bias emerges from noisy training labels (Sec. 4.3). Lastly, we demonstrate that the quality of classifiers' explanations improves with the given advice (12.8% improvement in Pointing Game [50] accuracy, described in Sec. 4.5). Code and datasets can be found at https://github.com/spetryk/GALS.

2. Related Work

Addressing bias with instance annotations. Most prior works that address bias in visual classifiers assume that some form of instance annotation is available. Some rely on expensive spatial annotations, such as object masks [11, 20, 34] or bounding boxes [4]. Hendricks *et al*. [11] address the image captioning task, where they want to reduce bias amplification and ensure a fair outcome for male and female genders by using person masks at training time. Others use slightly less expensive image-level annotations of the biased feature [1, 15, 39, 42, 46]. In contrast, in this work we do not assume that instance-level bias information is available. Instead, we rely on automatically generating attention guidance with readily-available language specification.

Addressing bias without instance annotations. Several works address bias without explicitly relying on instance-level bias annotations [5, 25, 29, 44]. Clark *et al*. [5] train an ensemble of low and high capacity models, forcing them to be conditionally independent, with the hope that the low capacity model will learn bias features, and the high capacity model will then learn the task-relevant features. Nam *et al*. [29] also train two models, one "biased" and the other "unbiased", by amplifying samples "aligned" with the bias for the first model (or easier to learn at the early stages), while amplifying the more difficult samples for the second model (where the first one fails). We view this line of research as complementary to our effort, and envision potentially combining these ideas with ours.

Language as information for visual tasks. Incorporating language in the zero/few-shot setting has been widely explored. Embedding language from class names or descriptions to obtain class "prototypes" is common in zero-shot learning, when no visual samples of the class are available [7, 8, 18, 32, 33]. Several works also aim to learn classes using their semantic attributes for better knowledge transfer [17, 26, 49]. Mu *et al*. [28] use image captions for regularizing few-shot representations to hold semantically meaningful information. Outside of zero/few-shot learning, Kim *et al*. [16] incorporate language advice into an autonomous driving controller, leading to a better performing and more explainable model. Rupprecht *et al*. [38] use language interactively to improve a pretrained CNN during inference time on segmentation tasks. Ling *et al*. [23] use language feedback to improve an image captioning model. To the best of our knowledge, no works have explored using language specification to improve visual attention in biased scenarios.

Information grounding with vision-language models. One of the key components of our approach is to leverage an off-the-shelf vision-and-language model to ground textual information into an image. There is a large body of work on visual grounding, where the models are trained to localize textual expressions in an image with a bounding box [31, 36] or a segmentation mask [12]. Unfortunately, these methods are constrained by the cost of providing these extra labels for the training set. Others can handle more open-ended queries, but the size of the available training

data is small as they require costly localization supervision, limiting the general application of these methods [14, 31]. A recent vision-and-language model CLIP [33] has demonstrated state-of-the-art image-text retrieval capabilities. CLIP is trained on 400M image-caption pairs sourced from the Web, making it a powerful general-purpose representation. We use CLIP and obtain grounding information with the help of salience visualization techniques [40].

Supervising spatial attention in visual classifiers. Another important component of our method is guiding the spatial attention within the visual classifier away from the spurious features. Several prior works have explored supervising spatial attention for, e.g., preventing catastrophic forgetting [6], fine-grained recognition [9], domain transfer [55] and generation of faithful explanations [37]. Specifically, the Right for Right Reasons approach [37] penalizes large input gradients in regions that are not allowed based on the user-defined "right reasons". We leverage this method to guide the classifier's attention towards the evidence pointed out by the language specification.

3. Guiding Attention with Language

In the following we outline *GALS*, our framework for incorporating language specification to guide a visual classifier; Fig. 2 provides an overview of our approach.

Problem Definition. In this work, we consider the learning problem in which we are given an image classification dataset $\{x_i, y_i\}_{i=1}^n$ for a prediction task $\mathcal{T}$ with C classes. Additionally, we assume we have a corresponding natural language specification $\mathcal{T}_s$ of the task or language descriptions of each class within the task $\mathcal{T}_s^c$. We also assume each image $x_i \in \mathbb{R}^{h \times w \times d}$ may contain a region of pixels that is irrelevant to $\mathcal{T}$, yet strongly correlated with y_i.

To model the distinction between the relevant and spuriously correlated pixels, we introduce a latent binary mask $\mathcal{Z}_i^{\mathcal{T}} \in \{0,1\}^{h \times w}$ for each image x_i, which encodes the relevance of each pixel to the task $\mathcal{T}$. That is, if $\mathcal{Z}_{i,(u,v)}^{\mathcal{T}} = 1$, then the value of pixel $x_{i,(u,v)}$ is informative for task $\mathcal{T}$ (and 0 if otherwise). Note that $\mathcal{Z}_i^{\mathcal{T}}$ is dependent on the prediction task. However, for notational convenience, we will omit $\mathcal{T}$ from $\mathcal{Z}_i^{\mathcal{T}}$ in the following.

Next, consider an image classification model f_θ with parameters θ. Our goal is to learn an optimal classifier f_{θ^*}, which outputs predictions $\hat{y}$ that rely only on task-relevant features (where $\mathcal{Z}_i = 1$). As $\mathcal{Z}_i$ is unobserved, we cannot learn f_{θ^*} by simply masking images according to locations of relevant features. Instead, we want to estimate a probability map over $\mathcal{Z}$, where each entry $x_{i,(u,v)}$ corresponds to the probability that pixel $x_{i,(u,v)}$ is relevant to $\mathcal{T}$.

Given this setup, our framework is three-fold: first, we create the high-level natural language specification $\mathcal{T}_s$ describing the semantic concepts relevant to $\mathcal{T}$. This is based on provided class names (e.g. "landbird") or description of the task (e.g. "bird species classification"). We then use a pretrained vision-language model and a spatial attention function to compute an estimate of the task attention $\mathcal{Z}$ for each image w.r.t. $\mathcal{T}_s$. Lastly, we use these estimates to supervise the spatial attention of f_θ, guiding it towards task-relevant features and away from unwanted biases.

Language specification. We assume access to natural language class names or a description of the task $\mathcal{T}$, but not necessarily access to what biases exist in the data. We argue that this is a safe assumption – in most real-world classification tasks, it is expected that a user has knowledge of what the categories are or what the task means. We then use the provided natural language to create $\mathcal{T}_s$ – words or phrases which are compatible with the choice of pretrained vision-language model, described below. For example, we preface task-relevant phrases with "a photo of" or "an image of" for compatibility with CLIP [33]. The language specification can be the same for each instance, or it can be class-specific by using the labels provided during training. Note that $\mathcal{T}_s$ is created once, prior to the training of f_θ, and does not require annotation of each image individually, allowing our framework to easily scale to large datasets.

Generating an estimate of $\mathcal{Z}$ from language specification. Consider a pretrained multimodal vision-and-language model VL, which has a joint understanding of image features and language phrases that correspond to them. For example, VL can be an image captioning or visual grounding model, or a model trained at scale with joint image-text supervision, such as OSCAR [21], VinVL [51], or CLIP [33], the latter of which we use in our experiments.

For every image x_i in the training dataset, we precompute a spatial attention map $\mathcal{A}_i^{VL} = Att^{VL}(\mathcal{T}_s^{y_i}, x_i)$, with $\mathcal{A}_i^{VL} \in \mathbb{R}^{h \times w}$. This serves as a probability map over $\mathcal{Z}_i$, where the attention value at location (u, v) estimates the likelihood that pixel $x_{i,(u,v)}$ is a task-relevant feature. The quality of $\mathcal{A}^{VL}$ as an estimate of $\mathcal{Z}$ depends on the ability of the pretrained vision and language model to ground text phrases in visual features. However, proper grounding within vision-and-language models is a research question on its own [24, 35]. Luckily, recent work on large-scale image-language pretraining has led to promising improvements [21, 33, 51]. Here, we use the saliency method Grad-CAM [40] to obtain reasonable attention maps.

Generating an estimate of the true task attention $\mathcal{Z}$ in this manner provides an automatic method for localizing per-instance, task-relevant features according to user specification $\mathcal{T}_s$. It requires only a high-level description of which semantic concepts are relevant to a task, which we view as a valid assumption for a user of a machine learning system.

Guiding the classifier with spatial attention. Next, for each image x_i, our objective is to guide the spatial attention of the classifier f_θ away from spurious correlations and towards task-relevant features. To do so, we would like to

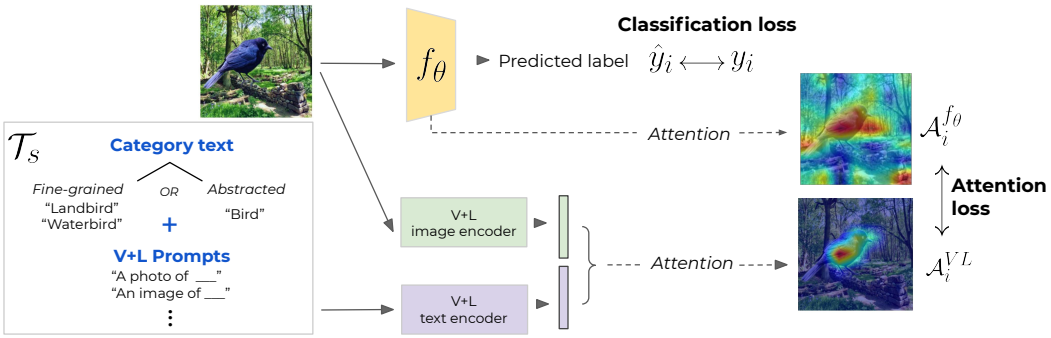


Figure 2. ***GALS* overview.** Our framework consists of three parts. First, we create a language specification $\mathcal{T}_s$ based on provided class names or a description of the task. Next, for every training image x_i, we use a pretrained vision and language model to ground the textual information into an image, in the form of an attention map A_i^{VL}. Finally, when we train the classifier f_θ, we incorporate A_i^{VL} as attention supervision. This encourages f_θ to align its attention $A_i^{f_\theta}$ with task-relevant concepts, and away from distractors.

supervise the spatial attention of f_θ with the $\mathcal{T}_s$ attention maps $\mathcal{A}_i^{VL}$ computed in the previous step of our framework. This requires a function $Att^{f_\theta}(x_i, y_i)$, which computes a differentiable attention map $\mathcal{A}_i^{f_\theta}$. The attention map specifies spatial locations in x_i that were relevant to the prediction $\hat{y}_i$.

We supervise the classifier's attention for each training image x_i by computing a loss $\mathcal{L}_{att}$ between $\mathcal{A}_i^{VL}$ and $\mathcal{A}_i^{f_\theta}$. The final training loss $\mathcal{L}(\theta, X, y, \mathcal{A}^{VL})$ for a batch of training images X with m samples is given as:

$$\mathcal{L}(\theta, X, y, \mathcal{A}^{VL}) = -\frac{1}{m}\sum_{i=1}^{m} y_i \cdot log(\hat{y}_i) + \lambda \mathcal{L}_{att}(\mathcal{A}_i^{VL}, \mathcal{A}_i^{f_\theta}) \tag{1}$$

where λ is a hyperparameter that controls the strength of the attention supervision.

Our proposed framework does not require an architectural change to the classifier f_θ, and only incorporates language-guided spatial attention as an auxiliary loss term in Eq. (1) during training time. Therefore, our framework requires no additional knowledge at test time.

3.1. Model Design Choices

Vision-Language model. We use the CLIP (Constrastive Language-Image Pre-training model) [33] as our multimodal VL model. CLIP is trained on 400M image-caption pairs (x_{text}, x_{image}) sourced from the Web. It consists of two encoders for mapping x_{text} and x_{image} into a shared embedding space. The contrastive objective encourages image and text embeddings from the same pair to be close (as measured by cosine distance), and embeddings from different pairs to be pushed apart. We include ablations with CLIP-based models trained on open source datasets in the Appendix [13, 48].

Generating attention. For the language specification $\mathcal{T}_s$, we define a set of CLIP-style prompts. These are framed as short sentence descriptions, such as "an image of *X*" or "a photo with *X*," for the word or phrase *X* that describes task-relevant concepts. We generate multiple such prompts for each task and later combine (via average or max) the corresponding attention maps for each image, which serves as our estimate for $\mathcal{Z}$. Once the prompts are defined, they are embedded with CLIP's text encoder into the shared image-text latent space. For embedding images, we use the image encoder of CLIP with the ResNet50 backbone provided by Radford *et al.* [33]. For the attention function $Att^{VL}(\mathcal{T}_s, x_i)$, we use the saliency method GradCAM [40] between the image-text similarity score and the feature maps after the last convolutional block in the image encoder.

Attention incorporation. For supervising the classifier's attention, we adapt the framework Right for the Right Reasons, or RRR [37]. The original goal of RRR was to provide the correct explanation for each sample in addition to the correct prediction. First, a user provides per-image binary masks of regions that are *irrelevant* to the task. It then penalizes the input gradients in those regions (the gradient of the output y with respect to the input x). Because our attention maps $\mathcal{A}_i^{VL}$ specify *relevant* regions to the task, we take $1 - \mathcal{A}_i^{VL}$ to specify *irrelevant* regions. We then compute the L1 loss between this and the input gradient. We normalize $\mathcal{A}_i^{VL}$ to contain values between 0 and 1 (instead of using a binary mask), as our intention is to estimate a probability map over the true task attention $\mathcal{Z}$.

Loss function. We apply GradCAM [40] to our chosen VL model (CLIP with a ResNet50 backbone), to provide $\mathcal{A}_i^{VL}$, the input gradients for a ResNet50 model pretrained on ImageNet as $\mathcal{A}_i^{f_\theta} = \frac{dy}{dX_i}$, and the RRR-based loss for $\mathcal{L}_{att}$. Thus, our loss function used in the experiments is:

$$\mathcal{L}(\theta, X, y, \mathcal{A}^{VL}) = \overbrace{-\frac{1}{m}\sum_{i=1}^{m} y_i \cdot log(\hat{y}_i)}^{\text{Classification loss}} + \underbrace{\frac{\lambda}{m}\sum_{i=1}^{m}\left|\frac{dy}{dX_i}(1 - \mathcal{A}_i^{VL})\right|}_{\text{Attention loss}} \tag{2}$$

Our proposed framework is not restricted to a specific choice of pretrained model VL, classifier f_θ, mechanism of generating attention maps $\mathcal{A}^{VL}$ and $\mathcal{A}^{f_\theta}$, and attention loss function $\mathcal{L}_{att}$. *GALS* in the following section refers to the particular choices described above. We include ablations in Sec. 4.6 and Sec. C for other choices of VL, $\mathcal{A}^{f_\theta}$, and $\mathcal{L}_{att}$.

4. Experiments

Training. We use a ResNet50 [10] backbone pretrained on ImageNet for all classification models, with an input image resolution of $(224, 224)$. The GradCAM attention maps from CLIP are of size $(7, 7)$, which is the spatial resolution of the activations from the last convolutional block. We resize them up to the input resolution before computing the L1 loss. All error bars show standard deviation across 10 trials. We report further details on training parameters (such as the loss weight λ) and hyperparameter sweeps in the Appendix.

Baselines. We compare our work with several baselines that do not require per-instance knowledge of bias features. All baselines use the same ResNet50 backbone for consistency. *Vanilla* is trained in the same manner as f_θ in our framework, except without the attention loss $\mathcal{L}_{att}$. *UpWeight* is the same as Vanilla, except it uses class labels to address class imbalance. It computes a weighted average of per-sample cross entropy. The weights are inversely proportional to the frequency of the sample's class in the training data, assigning a weight of 1 to the class with fewest samples. *Attention Branch Network*, or *ABN* [9], learns a feed-forward attention map before the last convolutional block of ResNet50 and element-wise multiples it with the activations, which is added back into the activations before passing to the rest of the model. It also adds an additional cross-entropy loss term based on features in the attention branch, to encourage the spatial attention to be class-specific[1]. We include tabular results of plots in the Appendix.

Visualizations. For all visualizations, the attention from language specification is generated with GradCAM (as in Sec. 3.1), and classifier attentions are generated with the black-box saliency method RISE [30]. More examples of attention for each dataset are in the Appendix.

4.1. Datasets

We evaluate our approach on datasets with explicit and implicit bias. Additional details are in the Appendix, including dataset size and creation. The license, PII, and consent details of each dataset are in the respective papers.

In the *explicit* bias setting, the distractor feature can be clearly defined and (potentially) labeled. We experiment with the synthetic Waterbirds dataset [39], where bias is easy to control. Specifically, the images of birds from the CUB dataset [45] are divided in two classes, landbirds and waterbirds. Next, birds are segmented out and pasted onto random land or water backgrounds from the Places dataset [53]. During training, most waterbirds appear on water backgrounds and landbirds on land backgrounds, while in validation/test sets each class has an equal number of samples on land and water backgrounds. We consider two scenarios, one in which there is a small fraction of samples (5%) in the training data that go against the bias (***Waterbirds-95%***) and a more challenging one, where the bias and labels are perfectly correlated during training (***Waterbirds-100%***).

The Food-101 dataset [2] presents a case of *implicit* bias, as it was intentionally created such that the training images were not cleaned – for example, the images contain noise in the form of incorrect labels, bright colors, and visual confusion. Certain other foods appear more frequently with some classes than the others (e.g. sauce appears more often with baby back ribs than with steak). The evaluation set, on the other hand, was more thoroughly cleaned. We construct a 5-way ***Red Meat*** classification task between baby back ribs, filet mignon, pork chop, prime rib, and steak.

We present a second dataset with implicit bias, ***MSCOCO-ApparentGender***, which is constructed based on MSCOCO Captions [3] and prior work [11, 52]. In this dataset, apparent gender labels are defined based on the people's outward appearance as reflected in image captions. As defined in [11], when discussing people in captions, there are three options: "Man", "Woman" or a gender-neutral term, e.g. "Person". To follow that, we consider a three-way classification task for apparent gender, using the provided captions to generate labels for the classes "Man", "Woman", and "Person" (the latter when the annotators did not use gendered words in the captions). There are different types of spurious correlations in this dataset, e.g. women appearing in some environments more often than men, or a distractor object co-occurring with men but not with women, etc.

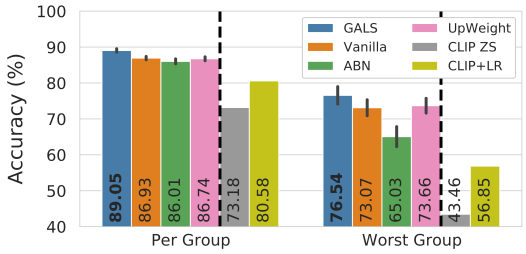
4.2. Explicit bias on Waterbirds

Since the Waterbirds dataset is constructed to encourage the model to pay attention to the background and not the bird, high-level language specification should give direction to attend to the bird, leaving the fine-grained discriminative image features up to the classifier to discover. Specifically, we generate attention from two CLIP prompts, to reduce noise – "an image of a bird" and "a photo of a bird." We average together these per-sample attentions to obtain $\mathcal{A}_i^{VL}$[2].

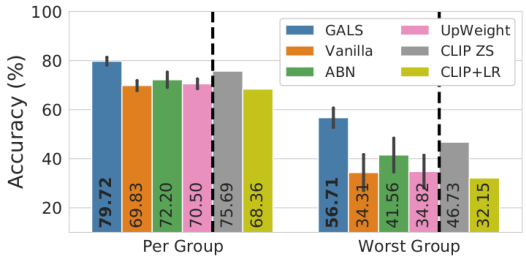
Following [39], we present test accuracy *per-group*, in which accuracy is weighted equally over the groups (specific combinations of class label and background, i.e. landbirds on land, landbirds on water, waterbirds on land, and waterbirds on water), and the *worst-group* accuracy. We are particularly

[1] We also experimented with supervising the attention of ABN with language specification. However, it under-performed the current formulation, and we include it in ablations in Section 4.6.

[2] In rare cases, the attention for a single prompt would be all-zero. Instead of averaging, we use the non-zero attention from the second prompt.



(a) Waterbirds 95%



(b) Waterbirds 100%

Figure 3. ***Waterbirds.*** Test accuracy on *Waterbirds-95%* and *Waterbirds-100%* datasets. Incorporating language specification results in a higher accuracy than all other baselines, including zero-shot CLIP and CLIP finetuned with logistic regression.

interested in the worst-group (usually waterbird on land) performance, which suffers the most when a model makes use of the spurious background correlations.

The concepts "landbird" and "waterbird" are rare with respect to concepts that can be learned from the Web, as would often be the case in new, real-world classification tasks. To illustrate that large-scale models like CLIP may lack fine-grained task-specific knowledge, we compare our method to zero-shot CLIP, as well as logistic regression trained on top of CLIP image-encoder features (following [33]). We find that CLIP often underperforms even the Vanilla baseline, demonstrating the value of taking the "best of both worlds" by combining large-scale multimodal model attention with CNNs on biased datasets with unfamiliar concepts.

Waterbirds-95%: As shown in Fig. 3a, our method outperforms all baselines on both per-group and worst-group accuracy. The strong bias in the data is evident when considering the worst-group accuracy, which drops the Vanilla performance by about 14%. Our model drives up the worst-group performance by 2.88% from the next-closest baseline of class weighting, without sacrificing per-group accuracy.

Waterbirds-100%: Because the class label and background are perfectly correlated, the performance of a classifier without any additional task information depends on whether it is easier to capture the true or bias signal. Surprisingly, the unsupervised attention mechanism in ABN provides ∼7% boost in worst-group performance as compared to upweighting by the class label. Our model improves on this, leading to a 15.15% improvement over ABN.

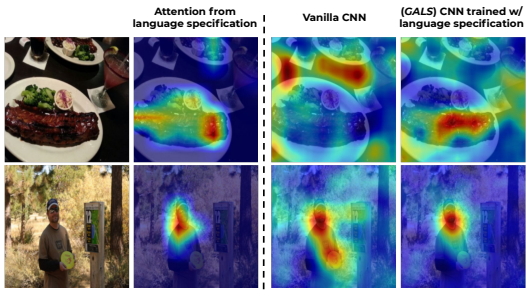


Figure 4. **Qualitative results for implicit bias.** Sample attention on *Red Meat* (top) and *MSCOCO-ApparentGender* (bottom). On these datasets the vanilla classifier may attend to non-task-relevant features due to implicit biases or noise. When we ground relevant features with language specification, we are able to move the classifier's attention away from the distractors.

4.2.1 Using language specification to change the task

Since the "landbird" and "waterbird" labels in the *Waterbirds-100%* training set are perfectly correlated with land and water backgrounds, we can easily redefine the labels to create a background classification task. We investigate whether we can use language specification to choose which hypothesis a model learns during train time: the "bird" or the "background" task. To study this, we keep the training set the same, yet update the validation and test labels to reflect background classification. We use the phrases "nature scene", "outdoor scene", and "landscape", preceded with "a photo of" and "an image of" as in our other experiments. We ensemble the attention maps by taking the max value for each pixel. A vanilla ResNet50 baseline achieves 86.75% per-group accuracy on the test set, with a worst-group accuracy of 72.90%. Impressively, our method outperforms this baseline by **2.22%** and **7.32%** on per-group and worst-group accuracy respectively, demonstrating the flexibility of language specification to select the desired training signal.

4.3. Red Meat Classification with Noisy Data

Along with assisting in removing explicit contextual bias in datasets, in this experiment we will show how our approach can improve the learning process on implicit bias caused by noisy data. We train on 5 balanced classes from the Food-101 dataset pertaining to red meat, as discussed earlier. We generate attention from the CLIP prompts "an image of meat" and "a photo of meat". Our results displayed in Table 1 and visualized in Fig. 4 (top) show that our method

	GALS	**Vanilla**	**ABN** [9]
Accuracy (%)	**71.20 ± 0.84**	67.39 ± 0.88	69.44 ± 1.12

Table 1. ***Red Meat.*** Test accuracy of our method, vanilla, and ABN for *Red Meat* Classification (a subset of the Food-101 dataset).

Method	Man			Woman			Ratio Δ	Outcome Divergence
	Man	*Woman*	*Other*	*Woman*	*Man*	*Other*		
Vanilla	83.60	6.20	10.20	66.80	28.60	4.60	0.349	0.071
ABN [9]	**84.80**	**4.60**	10.60	68.80	25.40	5.80	0.339	0.068
UpWeight	80.20	11.20	8.60	68.00	28.60	3.40	0.272	0.040
GALS	79.80	11.80	8.40	**74.20**	**22.60**	3.20	**0.160**	**0.022**

Table 2. ***MSCOCO-ApparentGender.*** Performance of our approach and the baselines on *MSCOCO-ApparentGender* test set. The best result in each column is **bold**, and second-best is underlined.

is able to outperform the ABN model by $\sim 2\%$ overall.

4.4. Implicit bias on MSCOCO-ApparentGender

Next, we discuss how our approach performs in another implicit bias scenario on the *MSCOCO-ApparentGender* dataset. We follow the evaluation protocol from [11] and generate attention from the CLIP prompts "an image of a person" and "a photo of a person". Table 2 summarizes the quantitative results and Fig. 4 (bottom) displays a qualitative example of the attention maps. For each "Man" / "Woman" sample we separately report the % of the time they have been classified as a *Man*, *Woman*, or *Other*. We penalize gender misclassification, but do not penalize if the "Person" class was predicted. In this task, we care about several aspects. (1) The training data is imbalanced (with more men than women in it), thus we aim to reduce bias amplification at test time [11]. The metric "Ratio Delta" measures how close the predicted men/women ratio is to the true one (which is equal to 1.0), i.e. lower is better. Our approach performs the best in this metric. (2) We also aim to ensure an equal outcome for both men and women. In practice, we see that men tend to be recognized more accurately than women, as seen from the higher **Man**/*Man* values than the **Woman**/*Woman* values (e.g., the Vanilla baseline achieves 83.6% and 66.8% accuracy, respectively). As we see, women often get misclassified as men (22−28% across methods). The "Outcome Divergence" metric measures Jensen-Shannon divergence [22] between the two sets of scores across the two classes, i.e. lower is better [11]. Again, our approach achieves the lowest outcome divergence, demonstrating the most fair behavior across all the compared methods.

4.5. Attention Evaluation

We evaluate the quality of our model explanations to determine if language specification makes the model right for the right reasons in addition to improving accuracy. To do so, we use the Pointing Game [50], a common evaluation for model explanations. For each input x_i, the Pointing Game (PG) requires a corresponding model explanation a_i and binary mask $\mathcal{Z}_i$, both of the same dimensions as x_i. Recall that $\mathcal{Z}_i$ indicates the task-relevant pixels in an image. A model passes the PG on sample i if the maximum value

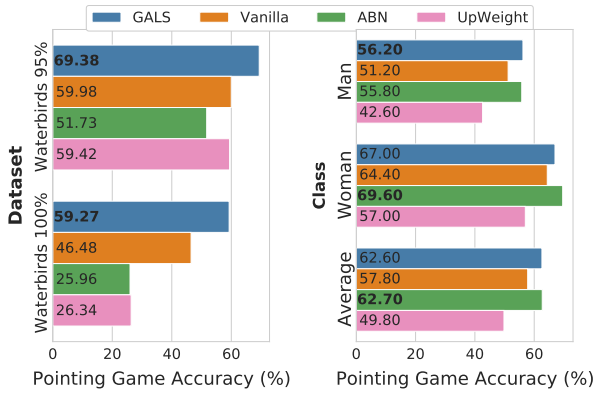


Figure 5. **Pointing game.** Pointing Game experiment [50] on *left.* Variants of Waterbirds datasets and *right.* MSCOCO-ApparentGender. We test whether the peak value of a black-box model explanation, generated with RISE [30], falls inside the segmentation label of the salient object.

of its explanation a_i falls inside $\mathcal{Z}$. In other words, the explanation is "pointing" to the correct region in the image.

For *Waterbirds-95%* and *Waterbirds-100%*, we use segmentation masks of the birds for $\mathcal{Z}$. On *MSCOCO-ApparentGender*, we use the available person segmentation masks, choosing the mask with the largest bounding box if multiple people are present to be consistent with our task. Segmentation masks for red meat in Food-101 are not available. For generating model explanations, we use the black-box saliency method RISE [30]. Fig. 5 presents our results. Our method matches the ABN baseline on *MSCOCO-ApparentGender*. However, we outperform all baselines by 9.4% on *Waterbirds-95%* and 12.8% on *Waterbirds-100%*.

4.6. Model Ablations

We explore several other design choices for the VL model and attention method in Table 3. More VL model ablations can be found in Table 10 in the Appendix. We consider the Attention Branch Network (ABN) [9] as the classification model, while supervising its feed-forward attention map (similar to [27]). We also try supervising the GradCAM from the last convolutional layer of a ResNet50 classification model directly. For generating language specification, we

Classifier Attention Method	Language Attention Source	Validation Accuracy		
		Per Class	Landbird	Waterbird
ABN	CLIP ViT	86.93	90.78	83.08
ABN	CLIP R50	86.10	90.25	81.95
GradCAM	CLIP ViT	87.20	91.32	83.08
GradCAM	CLIP R50	84.44	89.92	78.95
RRR	CLIP ViT	**88.25**	**92.28**	84.21
RRR	CLIP R50	87.26	89.17	**85.34**

(a) *Waterbirds-95%*

Cls. Att. Method	Lang. Att. Source	Man			Woman			RΔ	OD
		Man	*Woman*	*Other*	*Woman*	*Man*	*Other*		
ABN	CLIP ViT	84.40	10.60	5.00	68.40	29.40	2.20	0.306	**0.274**
ABN	CLIP R50	**90.60**	**5.40**	4.00	60.40	37.60	1.80	0.485	0.280
GradCAM	CLIP ViT	85.80	7.60	6.60	70.20	27.00	2.80	0.310	0.331
GradCAM	CLIP R50	83.40	7.40	9.20	66.20	29.80	4.00	0.311	0.298
RRR	CLIP ViT	87.00	8.40	4.60	68.60	29.80	1.60	0.341	0.305
RRR	CLIP R50	82.20	10.60	7.20	**72.20**	**26.00**	1.80	**0.235**	0.309

(b) *MSCOCO-ApparentGender*

Table 3. Comparison of different classifier attention methods and language attention sources on the (a) *Waterbirds 95%* and (b) *MSCOCO-ApparentGender* validation set. In (a), we report class instead of group scores, as we do not assume access to group labels at validation. The method indicated as "*GALS*" in Section 4 is placed at the bottom.

experiment with the CLIP ViT-B/32 (CLIP ViT in the table). The method we denote as "*GALS*" corresponds to the row with RRR as the classifier attention method supervised with CLIP ResNet50 GradCAM attention. For both ABN and GradCAM classifier attention methods, we compute $\mathcal{L}_{att}$ as an L1 loss in a similar style as in RRR — penalizing A^{f_θ} where A^{VL} is low, as opposed to matching A^{f_θ} directly to A^{VL}, finding that this gives slightly better performance. We chose RRR+CLIP R50 since it had the most consistent performance in minority class accuracy and fairness.

5. Limitations and Broader Impacts

In this work, we focus on a scenario where a dataset bias during training time is not present at test time. This is an important issue with serious implications for high-risk domains such as autonomous driving or medical imaging. Generally, as machine learning methods become widespread and impact people's lives, reliance on biases may be harmful to entire populations. Thus, we envision potential positive impact from our work towards mitigating this issue.

One of the datasets used in this work (*MSCOCO-ApparentGender*) is derived from the image captioning MSCOCO-Bias and MSCOCO-Balanced splits introduced in [11]. Following [11], we consider three gender categories: male, female, and gender neutral (e.g., person) based on visual appearance. The gender labels were determined using a previously collected publicly released dataset in which annotators describe images [3]. Importantly, people in the images are not asked to identify their gender. Thus, we emphasize that we are not classifying biological sex or gender identity, but rather outward gender appearance. In particular, we are interested in reducing gender entanglement with contextual features and "equalizing" the outcome across male and female categories.

We also would like to point out that in our experiments, we use the off-the-shelf large-scale vision-language model (CLIP [33]) which may have encoded some internal biases, transferred from the data on which it was trained. Specifically, CLIP was trained on 400M image-caption pairs sourced from the Web, so we can not rule out the presence of biases or harmful (e.g. gender or racial) stereotypes in it. Practitioners who wish to use our approach should be mindful of such sources of bias.

As described in Sec. 3, *GALS* is limited to biases which can be pixel-wise separated from relevant features. As a counterexample, it would not apply to the task of classifying a person's age, with a confounding factor of race. *GALS* also struggles when the vision and language model cannot ground the language specification (Fig. 6). In other scenarios, CLIP may struggle when the prompt is more compositional, such as "the person in the blue shirt sitting next to the table".

6. Acknowledgements

We would like to thank Dr. Sayna Ebrahimi for helpful discussion. This work was supported in part by DoD, including DARPA's LwLL, and/or SemaFor programs, and Berkeley Artificial Intelligence Research (BAIR) industrial alliance programs. In addition to NSF CISE Expeditions Award CCF-1730628, this research is supported by gifts from Amazon Web Services, Ant Group, Ericsson, Facebook, Futurewei, Google, Intel, Microsoft, Scotiabank, and VMware.

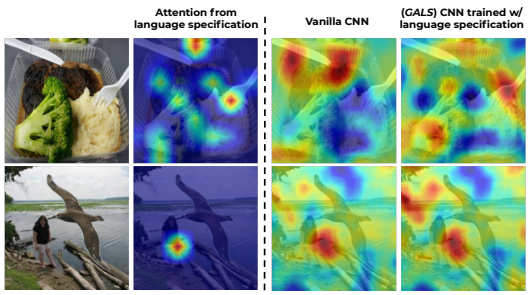


Figure 6. **Limitations.** Example of poor CLIP attention on *Red Meat* (top) and *Waterbirds* (bottom) dataset. Since *GALS* is supervised by the attention from language specification, our classifier's attention fails to ignore distractors when attention generated from language specification does not localize the task-relevant features.

References

[1] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2513–2523, 2021. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 2, 5, 12

[3] 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. 5, 8

[4] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[5] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. 2

[6] Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E Gonzalez, Marcus Rohrbach, and Trevor Darrell. Remembering for the right reasons: Explanations reduce catastrophic forgetting. *arXiv preprint arXiv:2010.01528*, 2020. 3

[7] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591, 2013. 1, 2

[8] Andrea Frome, Greg S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: a deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2121–2129, 2013. 2

[9] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019. 3, 5, 6, 7

[10] 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. 5

[11] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018. 2, 5, 7, 8, 12

[12] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 2

[13] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Open clip, 7 2021. 4, 13, 15

[14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3

[15] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9012–9020, 2019. 2

[16] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2020. 2

[17] 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, 2014. 1, 2

[18] Yannick Le Cacheux, Hervé Le Borgne, and Michel Crucianu. Using sentences as semantic representations in large scale zero-shot learning. In *European Conference on Computer Vision*, pages 641–645. Springer, 2020. 2

[19] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255, 2015. 1

[20] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 2

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 3

[22] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. 7

[23] Huan Ling and Sanja Fidler. Teaching machines to describe images via natural language feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[24] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 3

[25] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 2

[26] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Field-guide-inspired zero-shot learning. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 9546–9555, 2021. 1, 2

[27] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding human knowledge in deep neural network via attention map. *arXiv preprint arXiv:1905.03540*, 5, 2019. 7

[28] Jesse Mu, Percy Liang, and Noah Goodman. Shaping visual representations with language for few-shot classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[29] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[30] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 5, 7

[31] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2, 3

[32] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton Van Den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2249–2257, 2016. 1, 2

[33] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 4, 6, 8, 13, 15

[34] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR, 2020. 2

[35] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 3

[36] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 2

[37] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017. 2, 3, 4

[38] Christian Rupprecht, Iro Laina, Nassir Navab, Gregory D Hager, and Federico Tombari. Guide me: Interacting with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8551–8561, 2018. 2

[39] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 2, 5, 12

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3, 4

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 13, 15

[42] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11070–11078, 2020. 2

[43] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 13, 15

[44] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Towards debiasing nlu models from unknown biases. *arXiv preprint arXiv:2009.12303*, 2020. 2

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 5

[46] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019. 2

[47] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E. Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[48] Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data efficient language-supervised zero-shot recognition with optimal transport distillation. 2021. 4, 13, 15

[49] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *34th Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2020. 1, 2

[50] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 2, 7

[51] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl:

Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021. 3

[52] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017. 5, 12

[53] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 5

[54] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts, 2018. 1

[55] Andrea Zunino, Sarah Adel Bargal, Riccardo Volpi, Mehrnoosh Sameki, Jianming Zhang, Stan Sclaroff, Vittorio Murino, and Kate Saenko. Explainable deep classification models for domain generalization. *arXiv preprint arXiv:2003.06498*, 2020. 3