The contents of the Appendix are as follows:

Section A provides additional details on hyperparameters and model training.

Section B includes details on dataset splits as well as qualitative attention samples.

Section C presents an additional ablation study on the choice of VL model.

Section D provides tabular results for Figures 3 and 5 in the main paper.

A. Training Details

All runs were performed on 1-4 NVIDIA GeForce RTX 2080 GPUs. All models were optimized with stochastic gradient descent with a momentum of 0.9. For simplicity, we do not perform any data augmentations. For hyperparameter tuning, we split each dataset into training, validation, and testing, choosing the final hyperparameters based on which maximize validation accuracy. Our final hyperparameters are summarized in Tab. 5. Language specifications used in the experiments are shown in Tab. 6.

Waterbirds-95%. For the vanilla ResNet50 model, we perform a hyperparameter sweep with batch size 96, and run for 100 epochs. We sweep the backbone learning rate over 0.01, 0.005, 0.001, and 0.0001, and the linear classifier learning rate over 0.1, 0.01, 0.005, 0.001, and 0.0001. We chose a backbone learning rate of 0.01 and classifier learning rate of 0.001. For RRR, using the vanilla model learning rates, we first swept the attention loss weight (λ in Eq. (2)) over 1,000, 10,000, and 100,000, as well as the attention loss function over L1 and L2. From this, we chose a λ of 10,000 and an L1 loss. Next, we ran the same learning rate sweep as for the vanilla model, and chose a backbone learning rate of 0.005 and classifier learning rate of 0.0001.

Waterbirds-100%. We use the same hyperparameters found for *Waterbirds-95%*.

MSCOCO-ApparentGender. For the vanilla ResNet50 model, we run a hyperparameter sweep with a batch size of 96 for 100 epochs, testing backbone learning rates of 0.01, 0.005, and 0.001, and classifier learning rates of 0.1, 0.01, 0.005, and 0.001. We chose a backbone learning rate of 0.01 and classifier learning rate of 0.001. For attention weight λ, we test 1,000, 5,000, and 10,000, and choose 10,000 from validation.

Red Meat For the vanilla ResNet50 model, we run a hyperparameter sweep with a batch size of 96 for 50 epochs, testing backbone learning rates of 0.1, 0.01, 0.001, 0.005, 0.001, 0.0001, 0.0005, and classifier learning rates of 0.1, 0.01, 0.001, 0.005, 0.001, 0.0001, and 0.0005. For attention weight λ, we test 100, 1,000, 10,000, and 100,000, and choose 10,000 from validation.

B. Dataset Details

Waterbirds variants. For *Waterbirds-95%*, we use the same dataset as provided by the authors of [39]. For *Waterbirds-100%*, we follow the code provided by those authors for generating a new synthetic dataset. Land backgrounds are randomly chosen from the "bamboo forest" and "broadleaf forest" categories in the Places dataset, and water background are from the "ocean" and "natural lake" categories. These categories were determined in [39]. Both dataset variants have $4,795$ training images, $1,119$ validation images, and $5,794$ test images. Tables 7 and 8 show the number of samples per class, broken down further by the type of background. However, the validation and test set images themselves differ between *Waterbirds-95%* and *Waterbirds-100%* due to randomization in background selection.

MSCOCO-ApparentGender. For the training set, we begin by using the $22,966$ MSCOCO image ids defined in the Bias split in [52]. We next filter and label these images using a list of "male" words (such as "father", "man", or "groom"), a list of "female" words (such as "daughter", "lady", or "she"), and a list of "person" words which do not have a defined gender (such as "child", "surfer" or "employee") provided by [11]. From these provided lists, we filter out plural words. Next, we filter out images where the annotators do not agree on the gender (at least one caption mentions a male word and at least one caption mentions a female word). We label an image as "Man" if the majority of annotators (3 out of the 5 available captions per image) mention a male word, and "Woman" if the majority mention a female word. We label an image as "Person" if it has not been labeled as "Man" or "Woman" and if the majority of annotators have mentioned a "person" word. We use the same validation and test images for "Man" and "Woman" as in the "Balanced" split defined in [11]. Although these were not labeled in the same manner as our training set, we keep the splits the same to have consistent evaluation with prior work. The number of samples per class is summarized in Table 4.

Food-101. We start by selecting the 5 red meat classes from the Food-101 dataset [2] and split the 750 training samples into 500 training samples and 250 validation samples, keeping the 250 sample test set the same. The number of samples per class is summarized in Table 9.

Split	Man	Woman	Person
Training	10565	4802	2822
Validation	500	500	0
Test	500	500	0

Table 4. Dataset sizes on *MSCOCO-ApparentGender*.

Dataset	Method	Epochs	Batch Size	Base LR	Classifier LR	λ
Waterbirds-95%	Vanilla	200	96	0.01	0.001	-
	ABN	200	96	0.01	0.001	-
	UpWeight	200	96	0.01	0.001	-
	GALS	200	96	0.005	0.0001	10,000
Waterbirds-100%	Vanilla	200	96	0.01	0.001	-
	ABN	200	96	0.01	0.001	-
	UpWeight	200	96	0.01	0.001	-
	GALS	200	96	0.005	0.0001	10,000
Waterbirds-100% Backgrounds	Vanilla	200	96	0.01	0.001	-
	GALS	200	96	0.005	0.0001	1,000
MSCOCO-ApparentGender	Vanilla	200	96	0.01	0.001	-
	ABN	200	96	0.01	0.001	-
	UpWeight	200	96	0.01	0.001	-
	GALS	200	96	0.01	0.001	10,000
Red Meat	Vanilla	150	96	0.01	0.001	-
	ABN	150	96	0.01	0.001	-
	GALS	150	96	0.001	0.001	10,000

Table 5. Hyperparameter details. All models were optimized with SGD using a weight decay of $1e$-5. "Base LR" refers to the learning rate used for the pretrained ResNet50 backbone, and "Classifier LR" refers to the learning rate used for the linear classifier. λ is the attention loss weight from in Eq. (2).

Dataset	Language specifications
Waterbirds-95%	"{a photo/an image} of a bird"
Waterbirds-100%	"{a photo/an image} of a bird"
Waterbirds-100% Backgrounds	"{a photo/an image} of a nature scene", "{a photo/an image} of an outdoor scene", "{a photo/an image} of a landscape"
MSCOCO-ApparentGender	"{a photo/an image} of a person"
Red Meat	"{a photo/an image} of meat"

Table 6. Language specifications used for *GALS* in experiments. "{a photo/an image} of X" indicates that two prompts were used: "a photo of X" and "an image of X".

B.1. Attention Samples

In Figures 7, 8, and 9, we show several qualitative examples of spatial attention. Specifically, for sample images from the *Waterbirds-100%*, *MSCOCO-ApparentGender*, and *Food-101* training sets, we show the CLIP ResNet50 GradCAM A^{VL} guidance, as well as the RISE attention for the vanilla model and ours. We show that in many cases, our model has learned to attend to similar image features as the language-guided attention. However, when the image is difficult for the language-guided attention to ground the object of interest, then our model can have more difficulty in localization as well.

C. Additional VL Model Ablations

Table 10 presents an ablation study of VL models that were trained on open-source and smaller datasets than the original CLIP model [33]. Specifically, we generate GradCAM attention maps from a CLIP-ResNet50 model trained on a subset of 15M samples from YFCC [43], with the trained model provided by OpenCLIP [13]. We also generate attention with OTTER [48], a data-efficient CLIP-style model trained on the 3M samples in ConceptualCaptions [41]. OTTER (Optimal TransporT distillation for Efficient zero-shot Recognition) assigns soft matching labels to image-text pairs in a batch, as opposed to the one-to-one matching in the original CLIP formulation. The soft labels

Split	Landbirds, land	Landbirds, water	Waterbirds, land	Waterbirds, water
Training	3498	184	56	1057
Validation	467	466	133	133
Test	2255	2255	642	642

Table 7. Dataset sizes on *Waterbirds-95%*. The two classes are "Landbird" and "Waterbird." Furthermore, each image can display either a land background or a water background.

Split	Landbirds, land	Landbirds, water	Waterbirds, land	Waterbirds, water
Training	3694	0	0	1101
Validation	467	466	133	133
Test	2255	2255	642	642

Table 8. Dataset sizes on *Waterbirds-100%*. The validation and test splits have the same distribution as validation and test in Table 7 for *Waterbirds-95%*, although the images themselves are different from those in *Waterbirds-95%* due to randomization in background selection.

are based on a similarity matrix between each image-text pair in the batch, reducing noise and thus improving data efficiency by providing a continuous measure of similarity for contrastive learning. Once the attention is generated by the VL model, we use the same model architecture and training settings for *GALS* on the Waterbirds tasks as in Figure 3.

From Table 10, we see that all models perform comparably, with CLIP having the highest mean performance in all metrics except Worst Group on *Waterbirds-95%* (yet all model are within a standard deviation). Additionally, they all outperform the baselines shown in Figure 3. Although CLIP was trained on 400M image-text pairs, compared to the 15M in YFCC and 3M in ConceptualCaptions, the attention maps from all three VL models in Table 10 were similarly adept in guiding the downstream CNN attention away from background bias within the *GALS* framework.

D. Results Tables

We provide the tabular results for Figures 3 and 5 in the main paper. Table 11 presents per-group and worst-group results for *Waterbirds-95%* and *Waterbirds-100%* models. Table 12 shows Pointing Game results on Waterbirds variants and *MSCOCO-ApparentGender*.

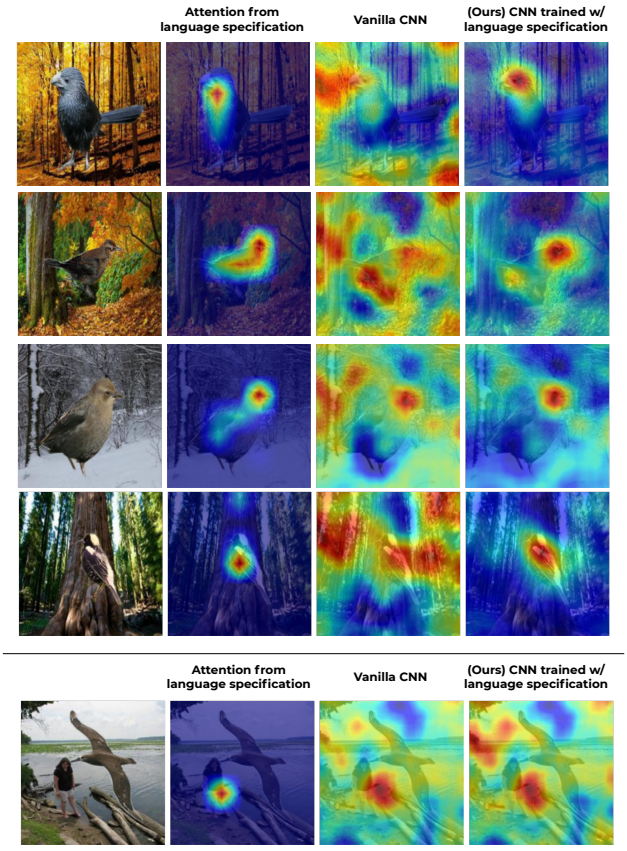

Figure 7. Sample attention visualizations from the *Waterbirds-100%* training set. Our model places considerably less attention on the background features than did the Vanilla baseline. However, it can have difficulty localizing the bird in cases where the language-guided attention also has difficulty in grounding, as shown in the bottom row.

Split	Filet Mignon	Filet Mignon	Pork Chop	Prime Rib	Steak
Training	500	500	500	500	500
Validation	250	250	250	250	250
Test	250	250	250	250	250

Table 9. Dataset sizes on *Food-101*.

	Waterbirds 95%		*Waterbirds 100%*	
VL Model	**Per Group**	**Worst Group**	**Per Group**	**Worst Group**
CLIP [33]	89.20 ± 0.37	75.25 ± 2.88	80.74 ± 1.04	55.30 ± 2.10
CLIP [13, 33] (YFCC [43])	87.87 ± 0.09	75.72 ± 2.97	78.39 ± 2.40	52.96 ± 3.54
OTTER [48] (CC 3M [41])	88.34 ± 1.22	76.63 ± 6.31	78.30 ± 1.99	51.17 ± 2.98

Table 10. Test accuracy with GALS, using attention generated from several different VL models on the *Waterbirds-95%* and *Waterbirds-100%* datasets. All models use a ResNet50 vision backbone. Mean and standard deviation are computed over 3 trials.

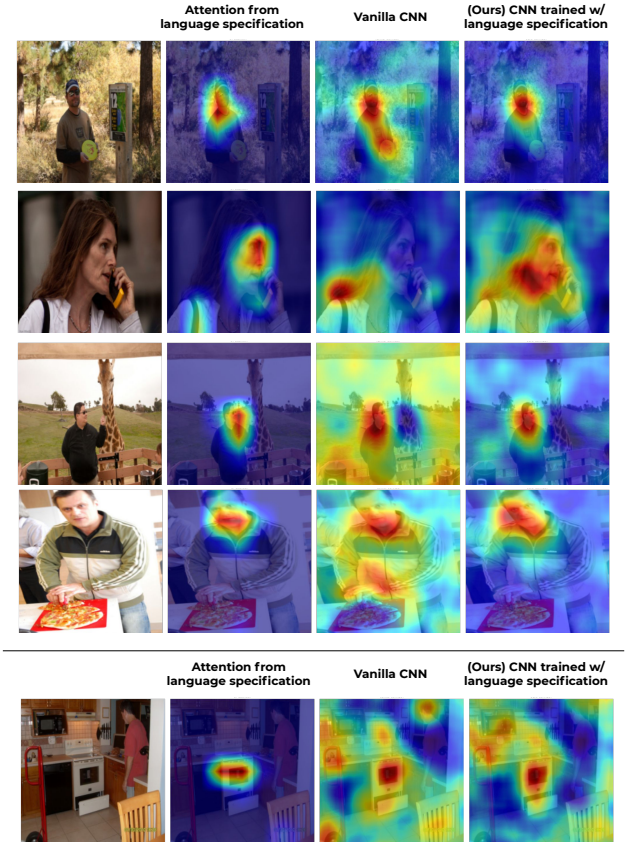

Figure 8. Sample attention visualizations from the *MSCOCO-ApparentGender* training set. Like the attention from language specification, our model is proficient at identifying faces, and placing less attention on potentially biased context. However, the sample shown in the bottom row is an example where the language-guided attention does not localize the person correctly, and our model attends to similar features as the vanilla model.

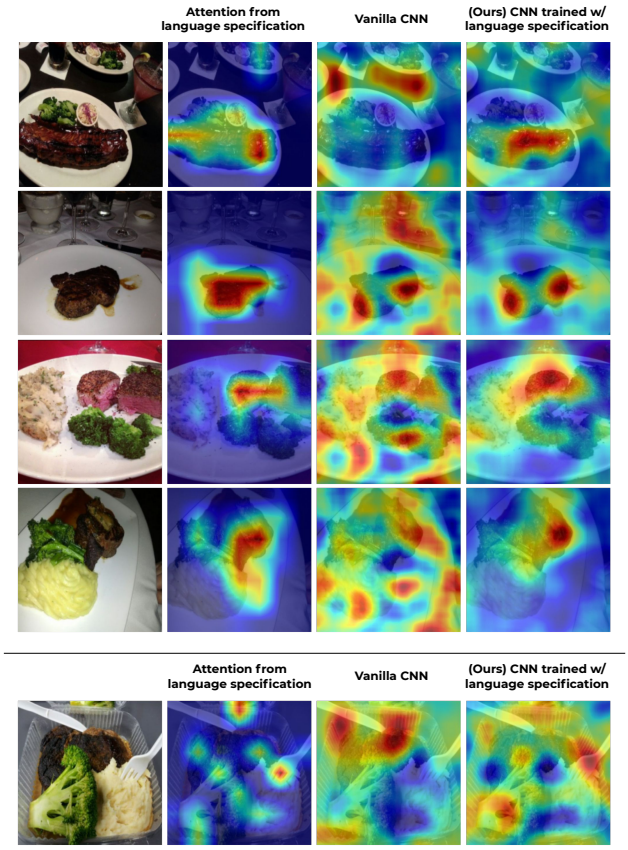

Figure 9. Sample attention visualizations from the *Red Meat* training set. The images tend to show cluttered plates of food, where the meat is often not centered. *GALS* can learn to attend to the meat itself when guidance from the language specification is also able to localize the meat.

Method	*Waterbirds 95%*		*Waterbirds 100%*	
	Per Group	**Worst Group**	**Per Group**	**Worst Group**
CLIP Zero-shot	73.18	43.46	75.69	46.73
CLIP Finetune, LogisticReg.	80.58	56.85	68.36	32.15
Vanilla	86.93 $\pm$ 0.46	73.07 $\pm$ 2.24	69.83 $\pm$ 2.04	34.31$\pm$ 7.31
UpWeight Class	86.74 $\pm$ 0.54	73.66 $\pm$ 2.00	70.50 $\pm$ 2.00	34.82 $\pm$ 6.65
ABN	86.01 $\pm$ 0.70	65.03 $\pm$ 2.77	72.20 $\pm$ 3.02	41.56 $\pm$ 6.76
GALS	**89.05** $\pm$ **0.47**	**76.54** $\pm$ **2.40**	**79.72** $\pm$ **1.60**	**56.71** $\pm$ **3.92**

Table 11. Test accuracy of approaches on the *Waterbirds-95%* and *Waterbirds-100%* datasets. The percentage indicates the proportion of training samples in each class which have a spurious correlation between the class label and the background. Note that the CLIP zero-shot accuracy differs in Waterbirds 100% and Waterbirds 95% because test set backgrounds differ. Tabular version of results in Figure 3.

Method	*Waterbirds-95%*	*Waterbirds-100%*	*MSCOCO-ApparentGender*		
			Man	**Woman**	**Average**
Vanilla	<u>59.98</u>	<u>46.48</u>	51.20	64.40	57.80
ABN	51.73	25.96	<u>55.80</u>	**69.60**	**62.70**
UpWeight	59.42	26.34	42.60	57.00	49.80
GALS	**69.38**	**59.27**	**56.20**	<u>67.00</u>	<u>62.60</u>

Table 12. Pointing game accuracy on *Waterbirds* dataset variants and *MSCOCO-ApparentGender*. Tabular version of results in Figure 5.