A. Unsupervised Dataset Bias Analysis

We first present additional figures and examples from Sec. 4.6. Sec 4.6. describes how LaViSE can be used for unsupervised bias analysis and shows detected concepts which are found to be associated with genders. Prior studies [1] have shown these biases using human annotations. We show that our method can uncover hidden biases without using any annotations except genders. To validate our argument, we quantitatively measure how effectively our method can find gender-skewed concepts and predict the degree of skewness by comparing the gender ratio for a concept from annotations in a dataset and the equivalent ratio automatically found by our method.

MS COCO is gender imbalanced. We use the same gender annotations for images in MS COCO as in [1]. Following the metric presented in the same work, we compute the gender ratio of each object class c in MS COCO as its bias toward males:

$$r_{gender,c} = \frac{N_{male,c}}{N_{male,c} + N_{female,c}} \tag{1}$$

where $N_{male,c}, N_{female,c}$ are the numbers of training images of class c that also contain either males or females respectively, excluding samples that are related to both genders. Note that LaViSE does not use this quantity or use class information, but we will show it can discover the same information. The model is just trained for a single binary category and the annotations (gender). In Figure 1, along the x-axis, we show how much different classes in MS COCO have their training set imbalanced toward males. A value above 0.5 indicates that there is a bias toward males for that class. As we can see from Figure 1, MS COCO has more training samples that involves males than females for most of its object classes. As the training data MS COCO itself is imbalanced with a bias toward the male group, we expect the concepts learned by models that are trained with MS COCO are biased as well.

Our gender bias analysis is consistent with the gender ratios of the training data. For each concept that we discover from the target layer with LaViSE, we have at least one filter associated with this concept. To validate the explanations given by LaViSE and the analysis supported by these explanations, we expect the gender ratios for the discovered concepts to have a positive correlation with the gender ratios of the corresponding classes in the dataset. For example, if we discover the "tennis" concept from a trained model, we expect it to have a similar gender ratio as the "tennis racket" class. Fig 2 shows more concepts and examples with an activated regions.

To compute the gender ratio for each discovered concept that corresponds to a training class, we first calculate the gender ratio for each associated filter u as the ratio of male images in the qualified images (see the definition of qualified images in Section 4.6). Formally,

$$r_{gender,u} = \frac{N_{male,u}}{N_{male,u} + N_{female,u}} \tag{2}$$

$N_{male,u}, N_{female,u}$ are the numbers of qualified images for filter u that include either males or females respectively. Then we simply take the average of the gender ratios of all associated filters to get the gender ratio for each concept.

As shown in Figure 1, our bias analysis results yield a strong positive correlation with the gender ratios from the original dataset. This quantitatively demonstrates the effectiveness of our framework as a tool for unsupervised bias detection, which can apply to a dataset or a model trained from unknown arbitrary dataset.

B. Details of the Evaluation Protocol

We want to evaluate how accurate the explanations given by our framework and its variants are. For any convolutional filter u ($u \in \{1, ..., d\}$) in the target layer, our framework gathers $s \times p$ words collected from top p most activated images and rank the words based on their frequencies. (s and p are tunable parameters.) It then composes explanations for filter u using top-α ranked words W_u ($|W_u| = \alpha$), where α is user-defined.

The challenge for our evaluation is that there is no ground-truth labels for the concepts learned by each filter. We only have the annotation masks $M_{x_i} = \{M_{x_i,j} \in \{0,1\}^{h_{x_i} \times w_{x_i}}\}_{j=1,...,k_i}$ for each image x_i, where k_i is the number of annotated concepts this image contains, h_{x_i}, w_{x_i} are the height and width of the input image, and $t_{x_i,j}$ is the concept corresponding to $M_{x_i,j}$,

We consider each annotated image individually. For the u-th filter, it will have p most activated images $\{x_{u,q}\}_{q=1,...,p}$. For each image $x_{u,q}$, we have its activated region

$$R_{x_{u,q}u} = (\theta_{ext}(x_{u,q})_u > T_u) \tag{3}$$

on filter u. T_u is a per-filter activation threshold and is determined in the same way as in [2] such that $P(\theta_{ext}(x_i) > T_u) = 0.005$ for all $x_i \in X$.

Suppose $x_{u,q}$ is the only image that we will have to explain filter u. For each annotation mask $M_{x_{u,q},j}$ of $x_{u,q}$, if the intersection-over-union score of $M_{x_{u,q},j}$ and $R_{x_{u,q}j}$ exceeds an universal threshold (0.04 [2]), we consider the corresponding concept $t_{x_{u,q},j}$ as a ground-truth concept for filter u. Suppose there are $r_{u,q}$ ground-truth concepts $G_{u,q}$ ($|G_{u,q}| = r_{u,q}$) for filter u.

Then for different choices of α, our framework outputs a sequence of words $W_{u,q}$ ($|W_{u,q}| = \alpha$) to explain filter

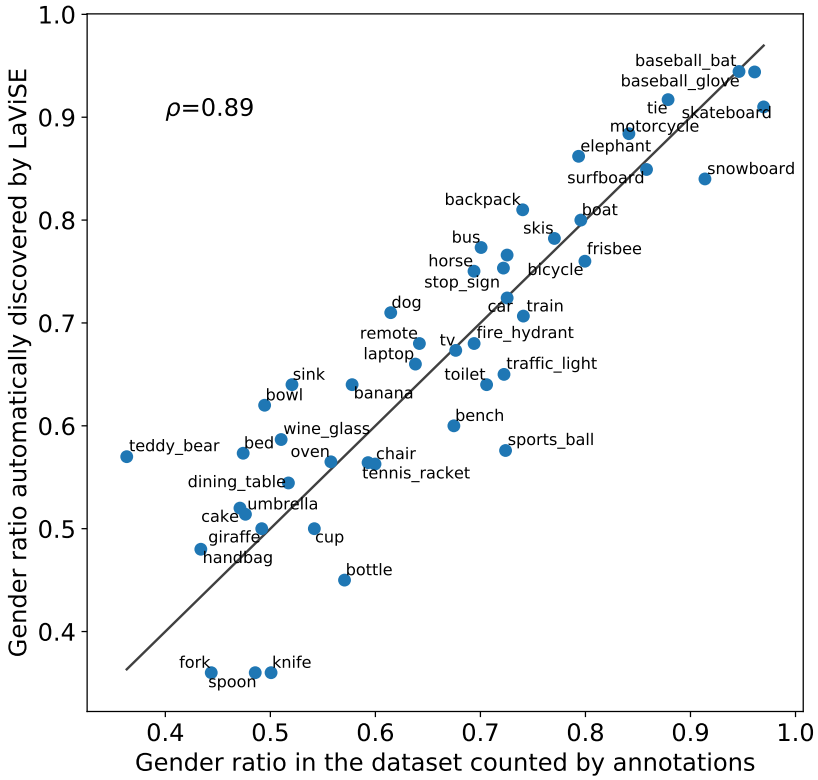


Figure 1. The validation of our unsupervised bias detection using MS-COCO annotations. The average gender biases discovered by our method are highly correlated with the actual gender ratios for the same concepts in the annotations of the dataset. ρ is the Pearson correlation coefficient.

Trained on	Recall (Top-5)	Recall (Top-10)	Recall (Top-20)
ImageNet	0.226	0.302	0.373
Random init.	0.009	0.025	0.034

Table 1. Interpretability on random-initialized feature extractor.

u. We compute the recall of using this single image $x_{u,q}$ to explain filter u as

$$Recall_{u,q} = \frac{|W_{u,q} \bigcap G_{u,q}|}{|G_{u,q}|} \tag{4}$$

To compute the overall recall of using LaViSE to explain a target convolutional layer in a model, we take the average of the recalls given by different images and filters:

$$Recall = \frac{1}{d \times p} \sum_{u=1}^{d} \sum_{q=1}^{p} Recall_{u,q} \tag{5}$$

C. Interpreting Random Features

The objective of our method is to interpret any existing black-box models, not to learn more interpretable models or to interpret an uninterpretable model beyond what it actually captures. We design an experiment to test if our method can only interpret what a model learned and does not generate irrelevant explanations. Foe example, a model that does not have any meaningful (interpretable) features may still yield interpretations by our method (*e.g.* due to the explainer). To ensure that our method does not interpret uninterpretable features, we apply our method to a randomly initialized feature extrator. Table 1 shows the result. As expected, the interpretability is very low, validating that the explainer indeed explains only the features captured in the extractor.

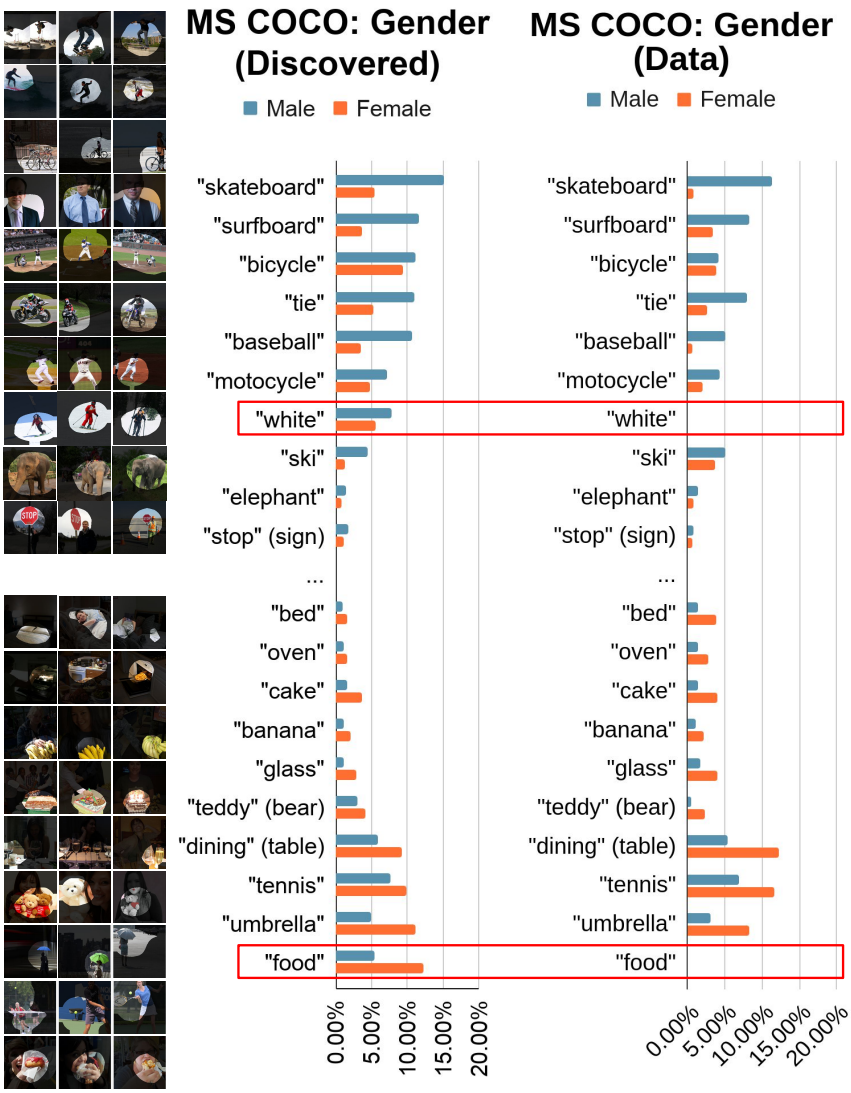


Figure 2. Concepts that best distinguish gender groups discovered by LaViSE from the last convolutional layer of a ResNet-18 model trained with the MS COCO dataset. "White" and "food" are not defined in the dataset but our method was able to find the concepts. For the "white" filter, our method also generates "baseball" and "sleeve" but "white" was the top word. This filter is also activated on the images of non-baseball players, which explains a smaller gender gap than the "baseball" filter.

D. Effect of the Number of Words

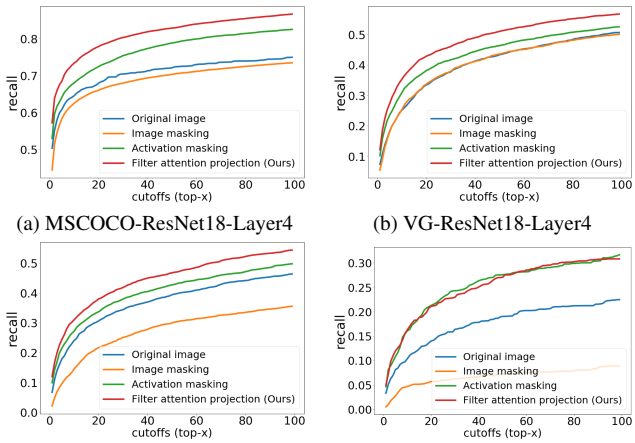
Figure 3 shows the quantitative results based on alignment between explanations and annotations as an approximation of the baselines' performance when the problem scales (i.e., the number of words increases). Our method performs the best in all settings upon different numbers of words asked for the explanations. In the fourth case, although the activation masking baseline has a comparative performance with ours, it may not perform as well in reality as in this approximation while our method has stably good performance according to the results of human evaluation.

E. Effect of Pre-training on Interpretability

Following the discussion in Section 4.5.5, Figure 4 shows the full lists of concepts LaViSE used to explain a pretrained and a randomly-initialized model. Both models are ResNet-18 and trained with the MS COCO dataset. See Section 4.5.5 for details.

References

[1] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Pro-*



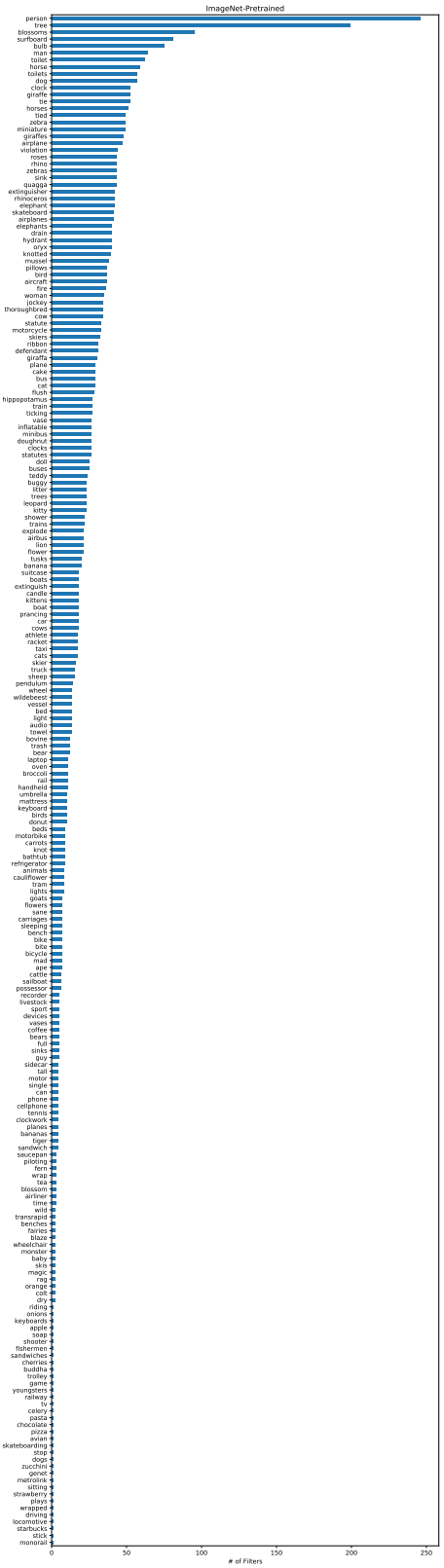
(a) MSCOCO-ResNet18-Layer4 (b) VG-ResNet18-Layer4

(c) VG-ResNet50-Layer4 (d) VG-ResNet50-Layer3

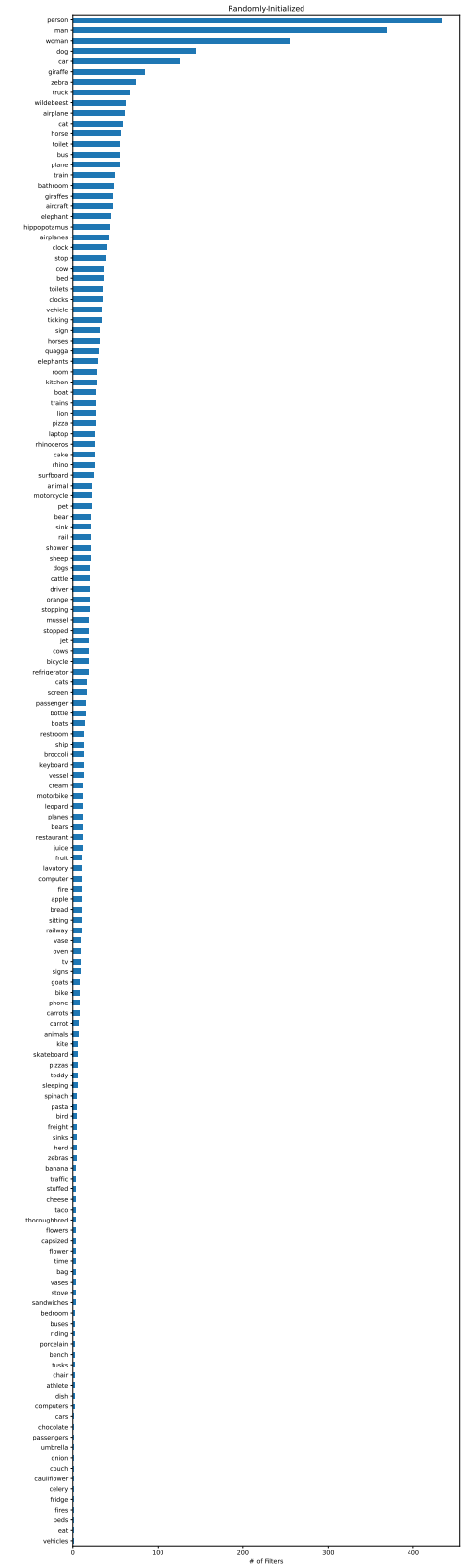
Figure 3. Recall rates of three baseline methods and our framework when x words are allowed for each explanation. These results correspond to the experiments shown in Table 2 in the main paper.

ceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2941–2951, 2017. 1

[2] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145, 2018. 1



(a) ImageNet-Pretrained



(b) Randomly-Initialized

Figure 4. Comparative analysis of pretrained and randomly-initialized models trained with the MS COCO dataset.