# Calibrating Multi-modal Representations: A Pursuit of Group Robustness without Annotations

Supplementary Material

# Appendix

## Table of Contents



## A. Datasets

We describe the benchmarks details used in our study. We benchmark CFR on the following sources.

- **Waterbirds** [61]: is a widely-used binary classification dataset focusing on spurious correlations. This benchmark combines the Caltech-UCSD Birds-200-2011 (CUB) dataset [61] with backgrounds from the Places dataset [76]. The classification goal discerns between landbirds and waterbirds, influenced by the spurious background attribute (either land or water). We adhere to the standard train/val/test splits as in [15].
- **CelebA** [34] is a binary classification image dataset consisting of over 200,000 celebrity portraits. This dataset's pivotal task – frequently cited in spurious correlation studies – is to ascertain hair color (specifically distinguishing blond from non-blond). Intriguingly, gender emerges as the spurious attribute. We adhere to the standard dataset splits as in [15]. This dataset is compliant with the *Creative Commons Attribution 4.0 International* license..
- **CheXpert** [16] is a comprehensive chest X-ray dataset from Stanford University Medical center, including over 200,000 images. The primary label is "No Finding", where a positive classification suggest the absence of illness. Drawing inspiration from [53], we factor in both race (*White, Black, Other*) and gender as intertwined attributes. We adhere to the train/val/test splits as in [67].
- **MetaShift** [31] stands as a versatile approach to crafting image datasets leveraging the Visual Genome project [24]. In this work, we follow [67] to adopt the pre-processed Cat *vs.* Dog dataset, aiming to accurately identify these two distinct animal species. It is important to note the presence of a spurious attribute in this dataset – image background, which tends to depict cats indoors and dogs outdoors. Further, we have utilized the "unmixed" version of the dataset, derived directly from the authors' original codebase, ensuring reliability and integrity in our analysis.

Table 3. **Experimental Settings**.

| Condition | Parameter | Value |
|---|---|---|
| *Model Architecture*: | | |
| CLIP-RN50 [48] | Input size | 256×256 |
| | Size of anchor/pos/neg features | $2048 \times 7 \times 7$ |
| | Output size of projection layer | 1024 |
| CLIP-ViT [48] | Input size | $336 \times 336$ |
| | Size of anchor/pos/neg features | 1024 |
| | Output size of projection layer | 768 |
| *Training*: | | |
| Optimizer | Type | SGD |
| | Learning rate | 1e-5 |
| | Momentum | 0.9 |
| | L2 weight decay | 1e-4 |
| | Metric to pick best model | WGA |
| Batch size | Anchor | 128 |
| | Cosine Similarity Loss | 128 |
| *Algorithm-specific*: | | |
| ○ CFR (ours) | Number of positive points | 16 |
| | Number of negative points | 16 |
| | EMA-coefficient (centroid) | 0.9 |
| CnC [73] | Number of positive points | 16 |
| | Number of negative points | 16 |
| JTT [33] | $\lambda_{\text{up}}$ | 10 |
| GroupDRO [51] | $\eta$ | 0.01 |
| *Dataset-specific*: | | |
| Waterbirds [61] | Raw input size | $224 \times 224$ |
| CelebA [34] | Raw input size | $178 \times 218$ |
| CheXpert [16] | Raw input size | $390 \times 320$ |
| MetaShift [31] | Raw input size | $256 \times 256$ |

## B. Implementation Details

Here we give the full details of the implementation for all evaluated methods. A collections of hyper-parameters is given in Table 3.

**Model Architecture** Our study utilizes CLIP [48] as the visual-language model. CLIP comprises two different components: a visual and a language branch. Within the visual domain, we utilize two popular architectures, ResNets (RN) and Visual Transformers (ViT), with a specific focus on

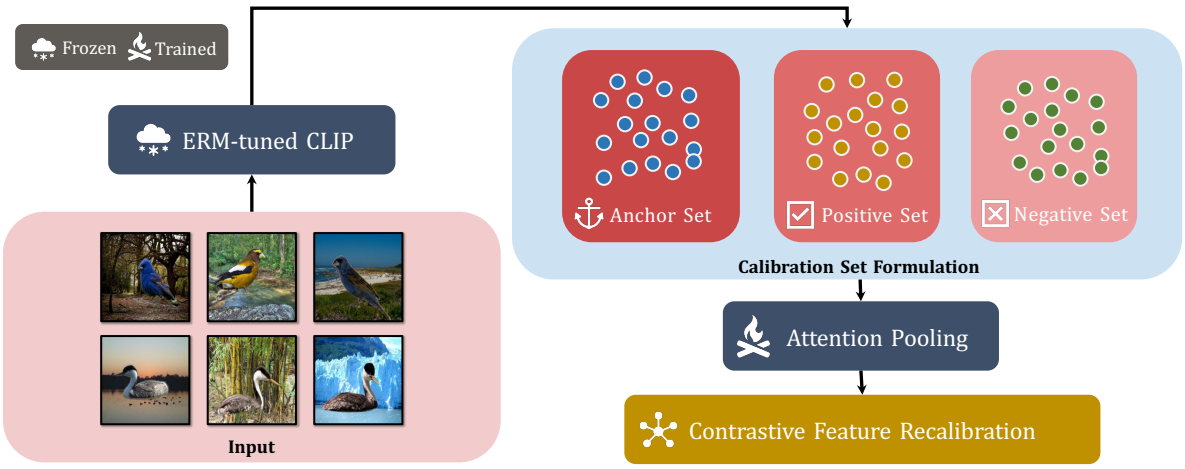


Figure 5. **The illustration of our proposed method CFR.** CFR decompose a lightweight representation calibration into two steps. (1) **Calibration Set Formation**: When a pre-trained CLIP is fine-tuned using ERM, this ERM-tuned CLIP with the frozen weights constructs a calibration set from the training data, as introduced in Sec. 5.1 (Main Context). This set comprises pivotal anchor points, with each sample selected based on its misclassification by the ERM-tuned CLIP. These anchors play a crucial role in refining the robustness across the dataset. (2) **Contrastive Feature Recalibration**: Utilizing the curated calibration set, CFR focuses on refining sample representations. This process involves aligning them more closely with the centroid of their respective class in the feature space while simultaneously distancing them from centroids of opposing classes. Such a recalibration is efficiently performed via a contrastive loss. Details about the positive and negative sample selection strategies used in CFR are discussed in Sec. 5.2 (Main Context).

ResNet-50 (RN50) and ViT-L/14@336px, aligning with the setting in [66]. Here 'ViT-L/14@336px' denotes the ViT-L/14 model that is fine-tuned on image inputs of 336 by 336 pixels. Concurrently, for the language branch, we incorporate the pre-trained mask language model, BERT [20].

In adherence to established protocols from prior work [66], our experiments, along with baseline methodologies, consistently freeze the language and vision encoders's weights. This deliberate choice, aimed at advancing methods that are not only efficacious but also resource-efficient, allows for training solely on the projection layer. Such an approach preserves the intrinsic knowledge within the model and safeguards against model collapse (*i.e.*, mitigating potential overfitting and yielding performance drops).

**Metrics.** Our study utilizes 'Worst-Group Accuracy' (WGA) and 'Average Accuracy' (Avg) as key performance indicators. Specifically, WGA denotes the lowest model accuracy observed across diverse groups within the test dataset. These groups are determined by considering the product space of the spurious attributes and the classes within the test dataset. WGA is a widely adopted metric in the spurious correlation literature, providing insights into the model's robustness across different groupings. Meanwhile, Average Accuracy represents the classification accuracy averaged over all classes within the test set. It offers a holistic view of the model's overall performance across all class categories.

**Experimental Setup.** In all of our experiments, we maintained consistent experimental setup using a single NVIDIA GeForce RTX 3090 GPU, and utilized fixed random seeds. We conduct our experiments using PyTorch 1.10.2+cu113 and Python 3.8.11, to ensure reproducibility.

**Calibration Set Generation.** Here we describe how our methods generate the calibration set, an essential component of our research. The calibration set comprises tuples, each consisting of an anchor, its corresponding positive batch, and its negative batch. The process starts by selecting anchors from the training dataset. For each data point in the training set, we employ an ERM-tuned CLIP model to obtain an initial prediction. If the prediction is incorrect, the data point is added to the calibration set as an anchor.

To sample the positive batch for a given anchor, we pre-compute positive sets for each class. The positive set of a class comprises all training data that is correctly predicted by the ERM-tuned CLIP. This pre-computation of positive sets is performed only once at the beginning. To select the positive batch for an anchor, we sample a small positive batch of size 16 uniformly from the positive set corresponding to the anchor's class.

For the negative batch, we also pre-compute negative sets, one for each class. For a specific class, the data points from all other classes are chosen to form the negative set. The key distinction is that the negative set includes data points regardless of the correctness of their initial predictions by the ERM-tuned CLIP. We have two options for building the negative batch from the negative set:

- Option 1 (RNS): Randomly sample a small batch of size 16 from the negative set, matching the size of the positive batch. This option is referred to as the RNS option in the

main text.

- Option 2 (NNS): Utilize cosine similarity to identify the nearest points among all the data points in the pre-computed negative set. Select the top 16 points to construct the negative batch, and this is referred to as the NNS option in the main text.

**Training Details.** For all methods evaluated in our experiments, including both baselines and our approach, we employ an SGD optimizer with weight decay set to $10^{-4}$ and a momentum of 0.9. The learning rate is fixed at $10^{-5}$, and the models are trained for 100 epochs. For the Calibration Loss ($\mathcal{L}_{\text{cal}}$), the batch size is set to 128, implying that each batch consists of 128 anchors along with their corresponding positive and negative batches. For the Cosine Similarity Loss ($\mathcal{L}_{\text{CS}}$), the batch size is also set to 128.

The model selection process remains consistent across all methods. We evaluate the model at the end of each epoch on the validation set and select the one with the best worst-group accuracy for the final testing. All accuracy metrics reported in this paper are based on the test set.

**Dataset Preprocessing.** Our dataset preprocessing steps are same across all four datasets and all evaluated methods. Initially, we resize the raw images, maintaining a fixed height-to-width ratio. This ensures that the shorter edge of the image has dimensions of 256 for ResNet-50 and 336 for ViT-L/14@336px. Subsequently, the resized image is cropped to 256×256 for ResNet-50 and 336×336 for ViT-L/14@336px. Following this, the image is normalized by subtracting the average pixel value and dividing by the standard deviation, a procedure consistent with CLIP [48]. No further data augmentation is applied after these steps, as our methods primarily focus on lightweight fine-tuning, which involves updating the projection layer of the vision branch of CLIP. Employing data augmentation in this scenario could lead to under-fitting due to the small parameter size of the projection layer.

## C. Additional Results

Owing to space limit in the main text, we have included Figure 6 in our supplementary materials. This figure provides a performance comparison of various semi-supervised methods utilizing CLIP-ViT, complementing Figure 3 from the main text, which specifically presents results for CLIP-ResNet50.

We report additional experiments on the Colored-MNIST (CMNIST) dataset for our methods and all the baselines. For CMNIST, we follow the same setup as in [73]. The result is shown in Table 4. We also report the result on pretrained CLIP without any further training or fine-tuning. We observe that DPS+RNS still performs the best.

## D. Additional Ablations

**Ablation on Loss Function Weights.** Our study delves into the balance between the two loss terms, $\mathcal{L}_{\text{cal}}$ and $\mathcal{L}_{\text{CS}}$, with a particular focus on the ratio $\lambda$ as defined in Eq. (5.4). The findings, as shown in Table 5, indicate that a $\lambda$ value of 1.0 is optimal.

**Ablation on Batch Sizes for Sample Selection.** As detailed in Sec. 5.2, our sample selection process involves selecting both a positive and a negative batch for each anchor within the calibration set. In this study, we delve into exploring the optimal batch sizes for these positive and negative samples. The result is shown in Table 6. We observe that our choice of $(16, 16)$ is the optimal for both the ResNet50 and ViT architectures.

**Ablation on Centroid-only Positive Batch.** For the DPS sample selection strategy of the positive batch in Sec. 5.2, we incorporate the positive subset $P(\mathbf{x})$ into the calibration loss together with the estimated optimal centroid $c_{\mathbf{y}}$ to intensify the recalibration effect, as detailed by Eq. 5.2. To demonstrate the necessity of adding $P(\mathbf{x})$, we evaluate CFR with $P(\mathbf{x})$ removed from the calibration loss. This results in the following variant of the calibration loss, which simply removes $P(\mathbf{x})$ from the summation in Eq. 5.2:

$$\mathcal{L}_{\text{cal}}(\mathbf{x}) = -\log \frac{e^{z_+}}{e^{z_+} + \sum_{\mathbf{v}^- \in N(\mathbf{x})} e^{z_-}},$$

where $z_+ = \langle f_\theta(\mathbf{v}), c_{\mathbf{y}} \rangle / \tau$, and $z_- = \langle f_\theta(\mathbf{v}), f_\theta(\mathbf{v}^-) \rangle / \tau$. The ablation result is shown in Table 7. Compared to semi-supervised baselines (*i.e.*, AFR, JTT, CnC), our proposed feature recalibration method, utilizing only $c_{\mathbf{y}}$, outperforms or performs on par with the three semi-supervised baselines. However, using $P(\mathbf{x})$ in the calibration loss (*i.e.* our default choice {DPS+RNS}) significantly boost the performance compared to $c_{\mathbf{y}}$-only, on both ResNet-50 and ViT.

## E. Additional Dataset Details

In this section we provide the details of the group information for all the datasets used in our experiments. The groups and the number of samples in each group of Waterbirds, CelebA, CheXpert and MetaShift are summarized by Table 8, 9, 10 and 11, respectively. Images samples from the datasets are shown in Figure 7, 8, 9 and 10.

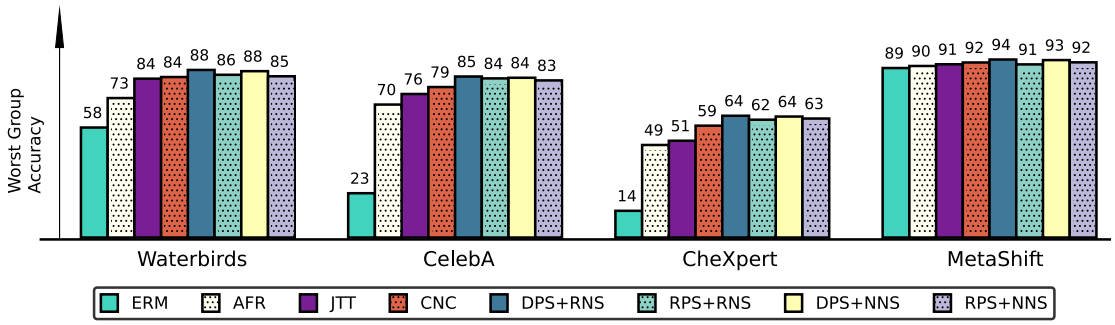


Figure 6. **Comparison of methods using the CLIP-ViT architecture on four benchmark datasets.** We use Worst Group Accuracy to evaluate the performance for various methods, including ERM, semi-supervised baselines (*i.e.*, AFR [47], CnC [73], JTT [33]), and our proposed methods. We observe that CFR combined with the sample selection strategies (*i.e.*, {DPS, RPS}×{RNS,NNS}) outperforms all semi-supervised baselines across all benchmarks.

Table 4. **Comparison results across various supervised methods, semi-supervised methods and our proposed four methods across the Waterbirds, CelebA, CheXpert, MetaShift and CMNIST (0.99) benchmarks.** Best results within the *semi-supervised* group are in **bold**. Please refer to the text for discussion.

| | | ResNet-50 | | | | | | | | | | ViT | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Waterbirds | | CelebA | | CheXpert | | MetaShift | | CMNIST | | Waterbirds | | CelebA | | CheXpert | | MetaShift | | CMNIST | |
| | Method | WGA | Avg | WGA | Avg | WGA | Avg | WGA | Avg | WGA | Avg | WGA | Avg | WGA | Avg | WGA | Avg | WGA | Avg | WGA | Avg |
| **supervised** | ERM [60] | 45.64 | 94.08 | 52.78 | 93.88 | 18.46 | 90.01 | 73.85 | 90.05 | 0.00 | 20.55 | 57.91 | 97.60 | 23.33 | 94.30 | 14.07 | 90.48 | 89.23 | 97.37 | 52.43 | 97.78 |
| | Pretrained CLIP [48] | 44.87 | 90.81 | 65.80 | 80.74 | 0.00 | 80.44 | 79.54 | 91.74 | 6.01 | 52.06 | 54.44 | 93.46 | 70.56 | 85.04 | 0.00 | 89.96 | 89.31 | 97.14 | 39.06 | 71.31 |
| | GroupDRO [51] | 75.08 | 83.84 | 84.09 | 89.54 | 68.29 | 75.04 | 83.19 | 87.30 | 54.87 | 69.19 | 90.82 | 96.37 | 88.33 | 91.24 | 67.02 | 73.53 | 93.85 | 97.37 | 87.34 | 94.17 |
| | S-CS [66] | 77.51 | 83.16 | 75.24 | 80.38 | 67.34 | 74.74 | 81.15 | 89.82 | 46.27 | 59.62 | 89.09 | 95.69 | 86.11 | 89.29 | 65.26 | 74.48 | 92.31 | 97.14 | 86.60 | 95.54 |
| | S-CL [66] | 75.23 | 85.96 | 75.56 | 80.56 | 64.49 | 75.89 | 81.54 | 88.79 | 48.97 | 58.75 | 89.93 | 96.04 | 87.78 | 90.51 | 66.26 | 74.19 | 93.14 | 96.89 | 85.86 | 95.29 |
| | DFR [23] | 73.22 | 83.82 | 82.22 | 91.57 | 60.64 | 74.96 | 83.08 | 88.33 | 48.97 | 69.18 | 89.69 | 97.80 | 85.56 | 90.80 | 68.09 | 76.59 | 92.31 | 97.03 | 83.87 | 94.71 |
| **semi-sup** | AFR [47] | 48.38 | 89.31 | 53.44 | 94.25 | 45.21 | 59.41 | 76.92 | 86.84 | 0.00 | 21.39 | 73.42 | 88.17 | 70.00 | 85.17 | 48.72 | 74.99 | 90.31 | 97.14 | 69.06 | 71.34 |
| | JTT [33] | 61.68 | 90.63 | 60.16 | 79.93 | 45.89 | 59.01 | 78.46 | 89.36 | 24.90 | 41.81 | 83.64 | 97.29 | 75.56 | 93.25 | 50.95 | 73.96 | 91.21 | 94.16 | 66.32 | 82.27 |
| | CnC [73] | 61.21 | 87.14 | 63.89 | 90.34 | 45.10 | 57.52 | 78.31 | 87.07 | 24.80 | 66.50 | 84.49 | 97.51 | 79.22 | 89.33 | 58.89 | 74.46 | 92.15 | 94.74 | 61.89 | 82.48 |
| | Con-Adapter [72] | 69.89 | 70.51 | 63.98 | 90.19 | 42.78 | 59.12 | 77.92 | 85.47 | 30.77 | 41.89 | 86.14 | 95.54 | 76.11 | 93.06 | 49.59 | 71.98 | 91.29 | 93.36 | 69.59 | 84.48 |
| | ○ DPS+RNS | **76.93** | 77.61 | **73.66** | 81.07 | **54.44** | 62.76 | **81.54** | 89.52 | **45.86** | 71.56 | **88.23** | 96.79 | **84.77** | 87.81 | **64.11** | 73.48 | **93.72** | 95.54 | **76.92** | 90.44 |
| | ○ RPS+RNS | 73.08 | 76.66 | 72.78 | 80.52 | 49.50 | 61.02 | 80.13 | 84.55 | 29.03 | 63.61 | 85.67 | 94.74 | 83.78 | 88.17 | 62.03 | 74.22 | 91.15 | 94.05 | 69.60 | 85.73 |
| | • DPS+NNS | 76.63 | 78.93 | 73.21 | 77.52 | 52.67 | 62.67 | **81.54** | 89.59 | 35.73 | 69.60 | 87.58 | 96.40 | 84.11 | 86.67 | 63.69 | 71.62 | 93.41 | 95.31 | 71.64 | 79.63 |
| | • RPS+NNS | 72.43 | 77.26 | 68.44 | 70.99 | 46.25 | 60.96 | 81.15 | 89.24 | 38.19 | 65.54 | 84.89 | 96.23 | 82.72 | 88.05 | 62.64 | 73.97 | 92.23 | 94.98 | 72.68 | 85.15 |

Table 5. **Ablation on Different Loss Components.** In this analysis, we adopt the {DPS+RNS} sampling strategy.

| | WGA | | | | | |
|---|---|---|---|---|---|---|
| Model | 0.1 | 0.2 | 0.5 | ○ 1.0 (**ours**) | 2.0 | 5.0 |
| CLIP-ResNet50 | 75.23 | 75.30 | 74.28 | 76.93 | 73.83 | 74.72 |
| CLIP-ViT | 85.83 | 86.78 | 86.87 | 88.23 | 86.85 | 87.07 |

Table 6. **Ablation on Batch Size in Sample Selection.** In this analysis, we adopt the {DPS+RNS} sample strategy.

| | | CLIP-ResNet50 | | | CLIP-ViT | | |
|---|---|---|---|---|---|---|---|
| | | Negative | | | Negative | | |
| | Size | 8 | 16 | 32 | 8 | 16 | 32 |
| Positive | 8 | 73.79 | 73.83 | 73.05 | 86.76 | 86.89 | 87.07 |
| | 16 | 73.39 | 76.93 | 73.68 | 86.21 | 88.23 | 86.30 |
| | 32 | 72.99 | 72.43 | 73.36 | 86.60 | 86.43 | 86.45 |

Table 7. **Ablation on the positive subset $P(\mathbf{x})$ on Waterbirds.** The configuration labeled as '$c_{\mathbf{y}}$-only+RNS' represents a scenario where $P(\mathbf{x})$ is excluded from the calibration loss, and RNS is employed as the method for selecting negative samples. For a more comprehensive comparison, we have also incorporated three semi-supervised baseline methods (*i.e.* AFR, JTT, CnC). Notably, integrating $P(\mathbf{x})$ into the calibration loss results in a significant performance improvement for both ResNet-50 and ViT models.

| | ResNet-50 | | ViT | |
|---|---|---|---|---|
| Method | WGA | Avg | WGA | Avg |
| $c_{\mathbf{y}}$-only+RNS | 60.22 | 66.11 | 76.60 | 85.00 |
| ○ DPS+RNS (ours) | 76.93 | 77.61 | 88.23 | 96.79 |
| JTT [33] | 61.68 | 90.63 | 83.64 | 97.29 |
| CnC [73] | 61.21 | 87.14 | 84.49 | 97.51 |
| AFR [47] | 48.38 | 89.31 | 73.42 | 88.17 |

Table 8. **Group information of Waterbirds.** The data pertains to the distribution of samples across all groups in the training and testing splits of the Waterbirds dataset. The dataset is categorized based on spurious attributes, which include {Water Background (BG), Land BG}, and the classes, which are {Waterbird, Landbird}.

| | Train | | Test | |
|---|---|---|---|---|
| | Water BG | Land BG | Water BG | Land BG |
| Waterbird | 1057 | 56 | 642 | 642 |
| Landbird | 184 | 3498 | 2255 | 2255 |


Figure 7. **Sample images from Waterbirds.**

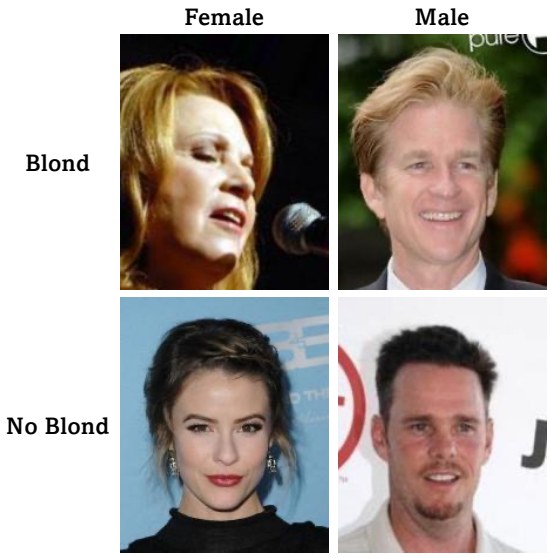


Figure 8. **Sample images from CelebA.**

Table 9. **Group information of CelebA.** The data pertains to the distribution of samples across all groups in the training and testing splits of the CelebA dataset. The dataset is categorized based on spurious attributes, which include {Female, Male}, and the classes, which are {Blond, Non-blond}.

| | Train | | Test | |
|---|---|---|---|---|
| | Female | Male | Female | Male |
| Blond | 22880 | 1387 | 2480 | 180 |
| Non-blond | 71629 | 66874 | 9767 | 7535 |

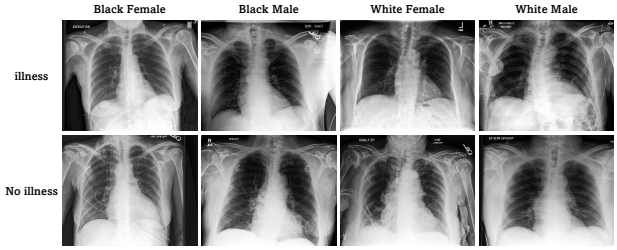


Figure 9. **Sample images from CheXpert.**

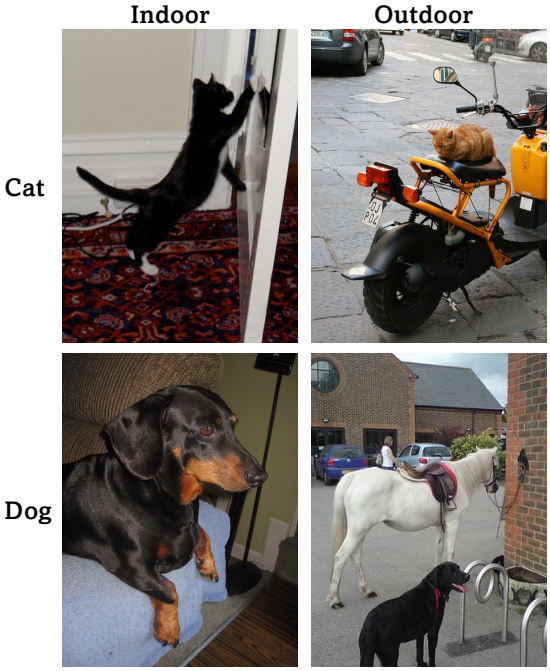


Figure 10. **Sample images from MetaShift.**

Table 10. **Group information of CheXpert.** The data pertains to the distribution of samples across all groups in the training and testing splits of the CheXpert dataset. The dataset is categorized based on spurious attributes, which include {Black female, Black male, White female, White male, Other female, Other male}, and the classes, which are {Illness, No illness}.

| | **Train** | | | | | | **Test** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Black female | Black male | White female | White male | Other female | Other male | Black female | Black male | White female | White male | Other female | Other male |
| Illness | 3877 | 4051 | 33676 | 51606 | 23407 | 33604 | 783 | 796 | 6772 | 10420 | 4696 | 6763 |
| No illness | 506 | 543 | 3490 | 5446 | 2975 | 3912 | 94 | 123 | 661 | 990 | 581 | 740 |

Table 11. **Group information of MetaShift.** This data pertains to the distribution of samples across all groups in the training and testing splits of the MetaShift dataset. The dataset is categorized based on spurious attributes, which include {Indoor Background (BG), Outdoor BG}, and the classes, which are {Cat, Dog}.

| | **Train** | | **Test** | |
|---|---|---|---|---|
| | Indoor BG | Outdoor BG | Indoor BG | Outdoor BG |
| Cat | 630 | 153 | 345 | 65 |
| Dog | 402 | 635 | 191 | 273 |