A. Appendix

A.1. Further Discussion on Related Work

Machine learning does not prescribe what features a model learns, further models may not learn the same features as a human, e.g., models may use texture over shape to classify objects unlike humans [21]. **Spurious correlations highlight the importance of understanding what features are learned and how they are represented**. Hermann et Lampinen [29] study the relevant question of what shapes feature representations, including an analysis of correlated features, with experiments on a synthetic "trifeature" dataset of a similar type to our pilot experimental data. By measuring the linear decodability of visual features from intermediate model layers, they find that when multiple features redundantly predict class labels, models preferentially represent the feature that is most linearly decodable from the untrained model, hypothesising that the "decodability of features from an untrained model reflects the model's inductive biases, and might predict the extent to which a feature would be preserved after training the model on a different task." There are some theoretical works also backing up their empirical findings [74, 75, 78].

The authors further find that across training, task-relevant features are enhanced, and irrelevant features for the task are partially suppressed. Their findings suggest that the features models represent depend on both the predictivity of features and their easiness of learning. This may explain why more reliable features can be suppressed by a less reliable, but easier-to-learn one. With regards to harmful spurious correlations [60], this would indicate that a model learns to rely on a less predictive feature since its easiness-to-learn dominates over its lack of reliability for prediction. Methods like MID, or other methods that fine-tune a model relate to these findings as they aim to make the model less sensitive to spurious features. They make the easier-to-learn, spurious attribute less predictive by training the model on examples that break the spurious trend. They may not (in particular, retraining only the final layer *will* not) reduce the linear decodability of the spurious attribute, but will adjust the feature contributions to the model output prediction.

Kirichenko *et al*. [39] advocate for simple last layer retraining, demonstrating that it can match or outperform state-of-the-art approaches on spurious correlation benchmarks, with far less expense. They further show that they could reduce reliance on background and texture information on large ImageNet-trained models using this technique. Izmailov *et al*. [36] study feature learning in the setting of spurious correlations. They show features learned by simple ERM are highly competitive, and that retraining the last layer beats specialised group robustness methods for reducing the effect of spurious correlations. Following this work, in our experiments, we retrained only the last layer. However, other methods of fine-tuning are also possible. Final layer retraining alters the weights of how features contribute to the output classification in enforcing the utilisation of semantically consistent features. As mentioned above, this will not alter the decodability of relevant or irrelevant features. However, further layer retraining may or may not suppress the spurious and now less predictive features [29] or induce a kind of catastrophic forgetting [53, 92].

Latent L	n(L)	Values
Shapes	3	S, O, H
Scales	6	0.5 - 1
Orientations	40	0 - 2
Positions in X	32	0 - 1
Positions in Y	32	0 - 1

Table 2. DSprites space of latent dimension values.

The work of Hendrycks and Gimpel [27] is also relevant to discuss in relation to our work. The authors observe that accurately classified examples usually have a greater maximum softmax probability than erroneously classified and out-of-distribution (OOD) examples. They use this finding to develop a simple baseline that utilises probabilities from the model's softmax distribution for detecting if an example is misclassified or out-of-distribution. Our work has a different underlying motivation to this work, as we examine logits (model outputs before any activation function is applied), and frame our work as a flavour of interpretability since we are motivated to comprehend representations with a richer level of detail than previous works examining maximally activating examples. However, our experiments on utilising the mid-range activations on spurious benchmark problems have a similar focus to the goal of Hendrycks and Gimpel as we locate low spurious images (an aspect OOD), and counterexamples to the spurious trend. A further overlap with this work is where we found that mid-level activations were useful for finding misclassified examples, as these are examples where the model has a near zero output logit value due to exhibiting uncertainty on the example or memorisation of the label. An interesting future research direction could be to measure how well mid-level activations can locate out-of-distribution data in these same benchmarks used by Hendrycks and Gimpel.

A.2. Synthetic Data Training and Experiments

The available combinations of the DSprites latent dimensions to sample values are described in Tab. 2.

DSpritesUnfair Details: The DSprites data consists of 3 shapes (the attribute to be classified), which are homogeneously dispersed in data that is randomly sampled, includ-

ing by x position. Therefore, we would expect each shape to be located in the left, middle, and right segments of the images one-third of the time. For each level of bias, b, we replace that proportion of the data with data containing only squares to the left, ovals in the middle and hearts on the right, the remaining 1-b proportion of the data is randomly sampled. For example, if 10% of the data is intentionally biased, the remaining 90% of the data is the original fair data, so the shortcut holds in 40% instead of 33.33% of the data.

Training Details: All models were trained with the Pytorch library [67] with a learning rate of 1e-3, Adam optimiser and a batch size 1000. The loss is the typical cross-entropy loss. DSprites experiments were performed locally.

Further Demonstrations of Representational Similarities: In Fig. 5, we plot the representation similarities for encoders trained on varying levels of biased training data. Cosine similarity is used as the measure of similarity. For a test sample of 1000 images, the similarity of the encoder's representation for each pair of images is calculated. The images are then sorted by (a) shape, and (b) x position to reveal patterns related to those attributes in the representational space. We see that for an encoder trained without bias (level 0.0), the encoder representations for elements of each shape are more similar to other images containing that same shape (high similarity along the diagonals) and the similarities ordered by position show no pattern. For bias level 0.5 (the best-performing model on the test set), the similarity pattern sorted by shape is even clearer. However, some pattern is also present for pairs ordered by position. For bias level 0.7 (where the model's performance on a fair test set is compromised), the pattern in the similarity of representations sorted by shape is not as clean. The pattern in the similarities sorted by x position shows more structure. By bias level 0.9 (here, the model's test accuracy is near random as the model relies mostly on the position shortcut), the similarities sorted by shape lack the clear pattern shown for lower bias levels, and a clear pattern can be seen when ordering images by x position. Fig. 6 shows sample images in the maximal and middle logit range for the class shape "square" for a model trained on data containing a harmful bias level of 0.9.

A.3. Harmful and Seemingly "Helpful" Spurious Correlations

Murali *et al*. [60] found that when the spurious feature is easier to learn than the core (which they term a *harmful* spurious feature), the model learns to leverage them and that in most cases, the core-only test accuracy drops to nearly random chance. However, results on one dataset (KMNIST with a patch shortcut) showed when a harmful spurious feature was removed in testing, the accuracy dropped, although it remained considerably higher than random accuracy, and the model performance had a wider variance across random seeds. In our setting, we found that some relatively small amounts of bias in the train set positively impacted model performance on a fair test set. We conjecture that this could be caused by the spurious signal interacting with the loss landscape in a way that makes it easier to find a lower loss. The easier-to-learn spurious pattern provides further signal for the model to find a better local optimum with gradient descent. The bias becomes a "harmful" spurious correlation (as defined by Murali *et al*. [60]) when the spurious signal overpowers the core features.

A.4. Further details on Spurious Correlation Datasets and Training

Tab. 3 gives specific details on the compositions of the spurious benchmark datasets used for our experiments. As captured by the "Group Counts" column, the datasets contain large group imbalances. Further, by conditioning on spurious attribute value, we see that the group imbalances are highly correlated with the labels. In the Waterbirds dataset, it is unlikely to find land birds on a water background. In the CelebA dataset, there are not many images of blonde celebrities who are male.

Following other works, for CelebA we finetuned the ImageNet pretrained model with stochastic gradient descent using an initial learning rate of 1e-3, a cosine learning rate scheduler, and a weight decay of 1e-4 for 50 epochs. For Waterbirds, we set an initial learning rate to 3e-3 and trained for 100 epochs, all else the same.

A.5. MID Filtering

For CelebA, we took 3,839 points out of 162,770 (2,000 logit intercept points for each class, meaning 3,839 unique points as there were overlapping points). Fig. 7 shows the logits sorted in descending order for the non-blonde class. Of these, the ERM model misclassified 1,137 points (out of a total of 2063 misclassified points. That is, the selected data contains 55.1% of all the ERM model's errors). Of the 3,839 logit intercept points, many of these points may appear ambiguous with respect to the label, or may have a wrong label. To handle mislabels and ambiguously labelled data, we apply BLIP in a VQA setting by asking "Is this person blonde?" for each image to obtain a "yes" or "no" answer. Fig. 8 shows a sample of randomly selected images with disputed labels. The first four images on the left are clear mislabels. The two rightmost images contain celebrities with a debatable hair colour. In both scenarios, it is apparent why the model would be confused by such images. In total, BLIP disputed the label for 1679 points, leaving 2160 points remaining after step 2 in MID.

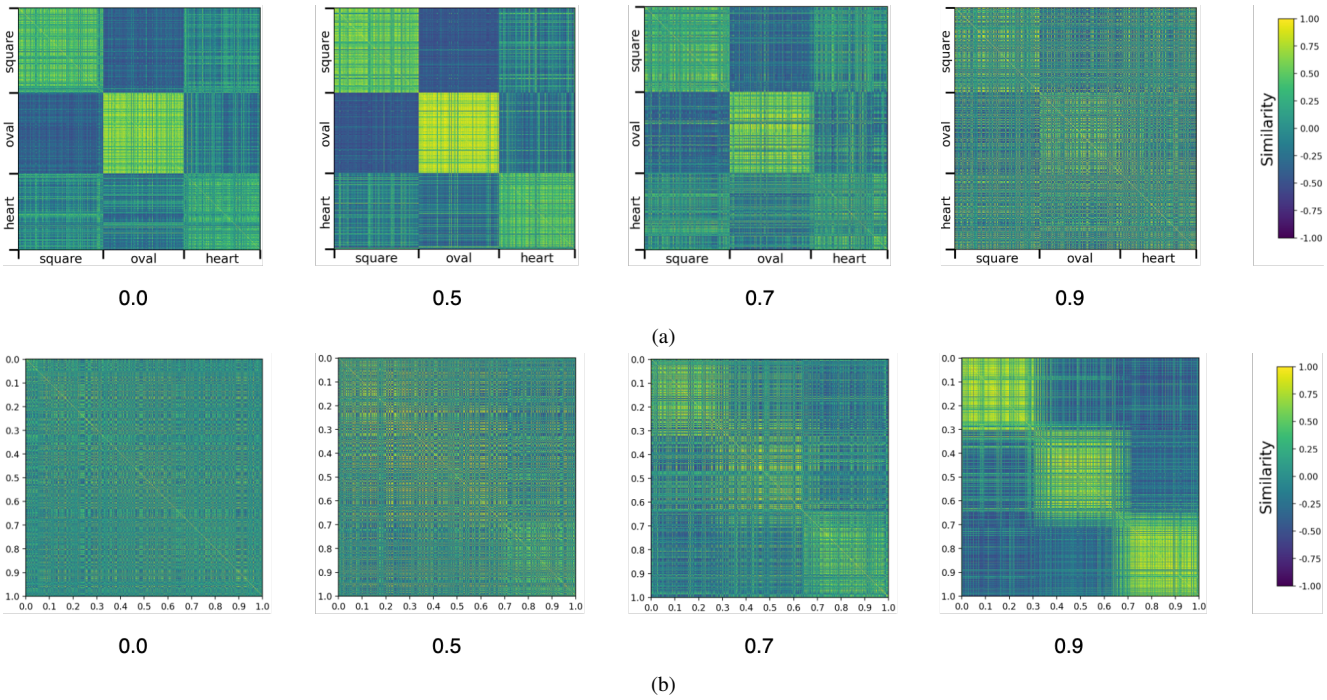
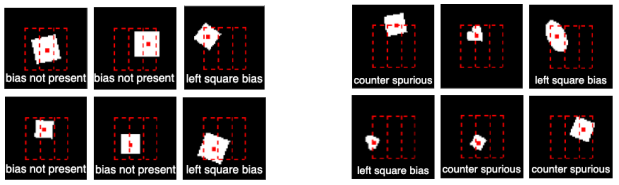


Figure 5. Representational similarity matrices plotted for encoders trained with varying levels of bias (below each image). The order of encoder embeddings is sorted by (a) shape, and (b) the x position of the shape.

Dataset	Labels	Group Counts		Class Totals	P($Y = y \mid S = s$)	
	$\downarrow y/s \rightarrow$	Water	Land	4795	Water	Land
Waterbirds	Water	3498 (73.0%)	184 (3.8%)	3682 (76.8%)	98.4%	14.8%
	Land	56 (1.2%)	1057 (22.0%)	1113 (23.2%)	1.6%	85.2%
		Female	Male	162770	Female	Male
CelebA	Non-blonde	71629 (44.0%)	66874 (41.1%)	138503 (85.1%)	75.8%	98.0%
	Blonde	22880 (14.1%)	1387 (0.8%)	24267 (14.9%)	24.2%	2.0%

Table 3. Label and group details for worst-group accuracy benchmarks. These datasets have both label and group imbalances. The final columns calculate how the class probabilities shift when conditioning on the spurious attribute s. The datasets exhibit a class imbalance which contributes to a large group imbalance. For example, less than 15% of the dataset has the label "blonde". A spurious correlation is created as 24% of females are blonde while only 2% of males are blonde.



(a) (b)

Figure 6. (a) Maximally activating examples for the neuron corresponding to the square class for training bias level 0.9. (b) Mid-level activating images in the same setting.

A.6. MID Clusters and Retraining Details

K-means cluster analysis: K-means was applied using the scikit learn package [68] starting with $k = 2$ as described in Sec. 5.1 with default settings besides a fixed random seed 0, 5 initial centroids, and a maximum number of iterations set to 1000. A manual inspection of 50 images from each cluster was allowed to establish if the clusters consisted of data that was reasonable for the model to exhibit uncertainty about. Fig. 9 shows examples of images from each of the three clusters. Cluster 1 appears to consist of unusual images, such as strange lighting, unusual artefacts and hair colours that the model is likely not

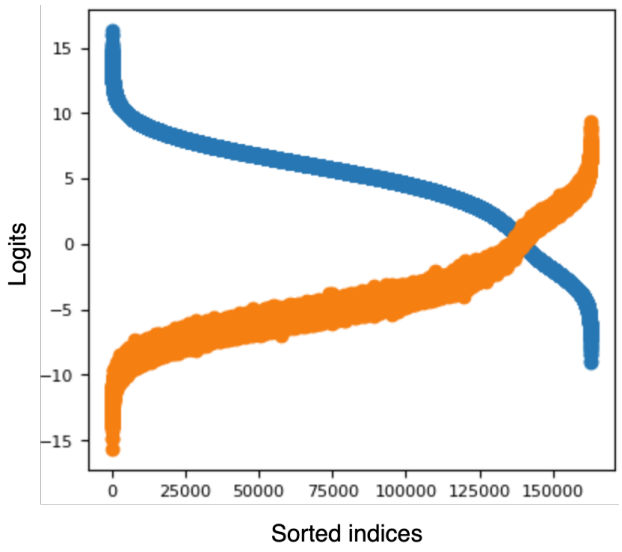


Figure 7. Logits in descending order for the class non-blonde (in blue), and blonde logits (in orange).

too familiar with. For these reasons, it is not unreasonable that the model might struggle with classifying these images. Cluster 2 contains many blonde males, but also some blonde females. Studying the pattern led us to summarise this cluster by containing masculine features such as a larger chin (male or female), short hair, or tightly tied back hair. Since the images inspected were all clearly blonde, the model's poor performance on this cluster is less understandable. Further, as discussed above, an apparent pattern is visible. Therefore, we marked this cluster for retraining. Cluster 3 is composed of images of female celebrities with hair colour that is between blonde and brunette, or auburn coloured hair. We believe this cluster reflects an understandable level of uncertainty expressed in the logits. Fig. 10 shows the group compositions of the cluster, which our method does not use, but we include for demonstration purposes. The groups represent the combination of label and spurious attribute. Group 1 is the combination of non-blonde and female attributes, group 2 non-blonde males, group 3 blonde females, and group 4 is the minority group consisting of blonde males. As discussed above, cluster 2 contains mainly images of blonde celebrities. The corresponding middle chart in Fig. 10 shows that this cluster contains many blonde males and blonde females. We note that we began clustering with 2 clusters and found good results for 3 clusters. However, we repeated analysis for 4 clusters to check if a pattern emerged. We found similar results, with the first cluster described above roughly splitting in two.

Selecting data for retraining and manual inspection: The data used for retraining is simply taken from the clusters where poor model performance on cluster labels is deemed inappropriate (e.g., it is unacceptable to perform poorly in classifying the hair colour of people with masculine type features) and not clusters where poor performance is deemed reasonable (e.g., a cluster containing celebrities with in between hair colours). Regarding manual inspection and automating our method, the cluster labelling component can be done without human input using an appropriate VLM if available for the application. Automating the inspection of what is unusual about the clusters or if poor performance is acceptable is left to future work. This is because the method allows for the spurious correlation to not be preconceived (however, if the spurious correlation is preconceived, a VLM can again do this step), and reasoning about whether poor performance on a cluster is appropriate or not requires common sense, and domain knowledge in some cases.

We suggest that for many applications an LLM may be able to assist with this component of the method (e.g., for the cluster of dark blonde and light brown-haired females, one could ask an LLM if it is reasonable for a model to have uncertainty classifying this type of hair as blonde or non-blonde?). However, a full study of the applicability of LLMs to automate this type of common sense human analysis should be undertaken before trusting a model with this task for many applications. Not requiring preconceived ideas about spurious attributes is a key feature of our work which distinguishes it from methods requiring group label information, which implicitly assume a preconceived spurious attribute. In many cases, we may not know which concepts are entangled with each other. For now, having a human in the loop to some extent may be wise to avoid failures for many applications where the spurious correlation is not preconceived as is allowed for in our MID experiments. Fortunately, given that MID significantly reduces the amount of data to be inspected, the human workload should be greatly reduced with inspection required for just a few images from each cluster.

MID retraining: We train the regularisation parameter for the logistic regression reweighting on one-half of the data from poorly performing selected clusters (Step 4). We test for values in the range $\{1.0, 0.7, 0.3, 0.1, 0.07, 0.03, 0.01\}$ and find a strength of 1.0 is the best-performing value for both CelebA and Waterbirds. We then select this optimal value that leads to the best accuracy and train on the available validation data with additional randomly selected training set data so that the model does not "forget" the majority groups leading to a poorer overall accuracy in favour of a higher worst-group accuracy.

Figure 8. Middle logit images with y labels on top disputed by BLIP labels shown below each image.

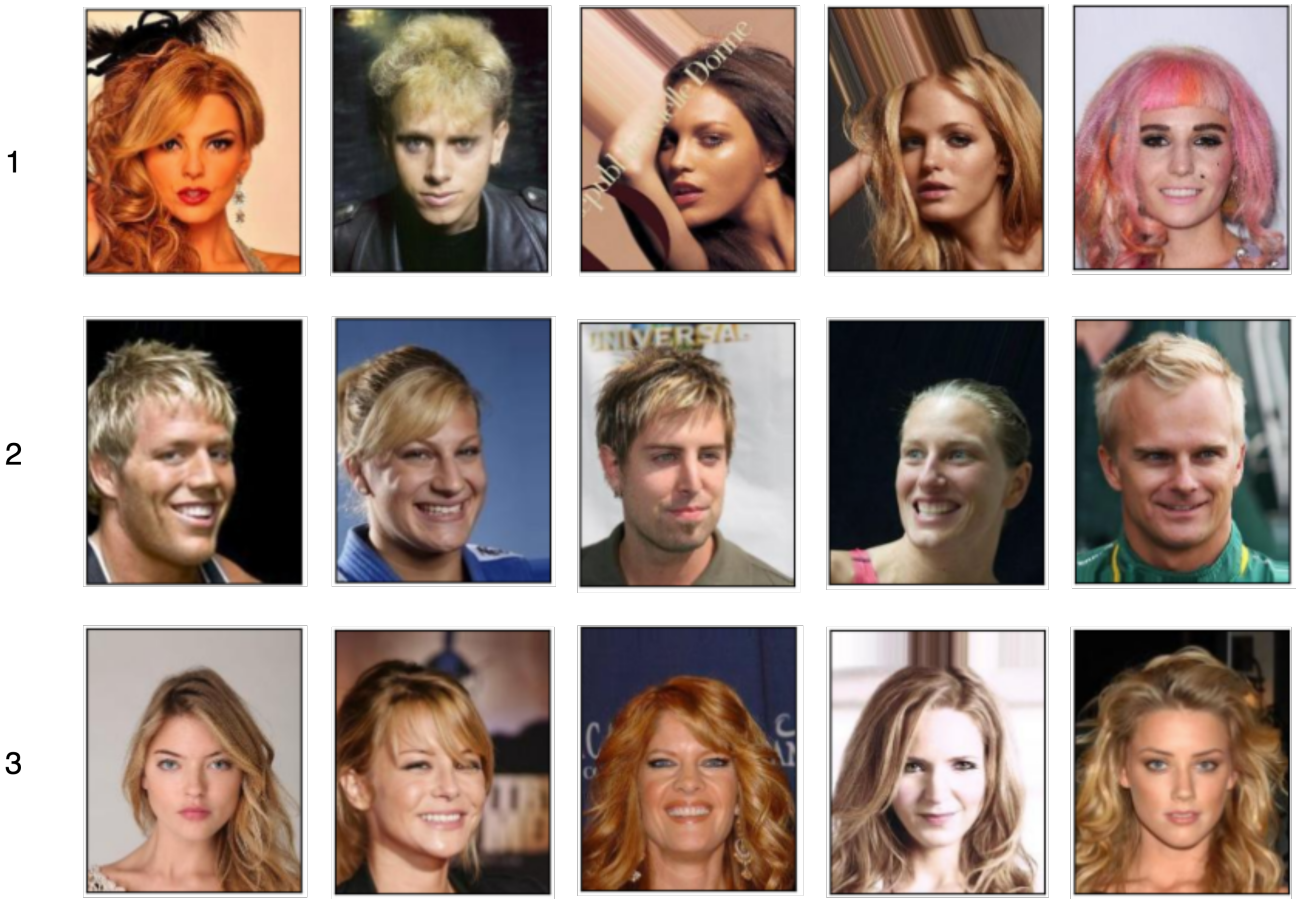


Figure 9. Group composition of k-means clusters for $k = 3$ on the filtered CelebA dataset.

A.7. MID Ablation Experiments

In this section, we demonstrate the importance of taking only mid-level logits through the absence of Step 2 in our method, MID, described in Sec. 5.1 for the CelebA dataset. We investigate the need to narrow our focus to the middle logit intercept data, we repeat k-means clustering (step 3 of our method) without the preceding filtering step (Step 2). Fig. 11 shows the resulting cluster compositions by label and spurious attribute. The pattern that emerges from analysing the clusters is roughly group 1 contains non-blonde celebrities with various backgrounds, group 2 consists of blonde and light-haired non-blondes, group 3 contains brunette celebrities with long hair, and group 4 contains non-blonde celebrities with short hairstyles. None of the clusters were found to reveal a spurious pattern in the data. This finding is in line with Sohoni *et al*. [80]. We further investigate the need to select specific clusters that the model should perform well on, rather than just a ran-

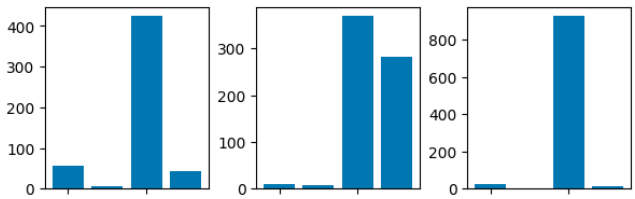


Figure 10. Group composition of k-means clustering on the filtered CelebA dataset.

dom cluster. We find that applying MID to random clusters does not necessarily improve model performance, and can negatively impact model performance in some cases. E.g., retraining the model on the cluster of photos of celebrities with hair that is between blonde and brown, does not help with performance and certainly not the WGA. This cluster reflects a reasonable model uncertainty as a human could easily spot that classifying these samples is difficult and that this cluster may also reflect some label inconsistency.

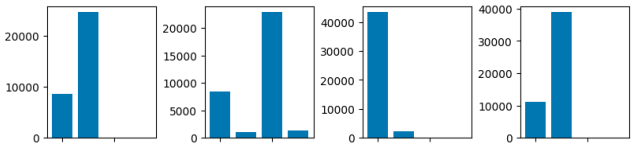


Figure 11. Group composition of k-means clustering on the entire CelebA dataset.