A. Diabetic Retinopathy Concept Creation

For creating images representing diagnostically relevant concepts, we use the pixel-wise annotations of the FGADR dataset to construct positive and negative concepts. An illustration of concept overlap in the fundus images, motivating the need for a controlled creation pipeline, is provided in Figure 8.

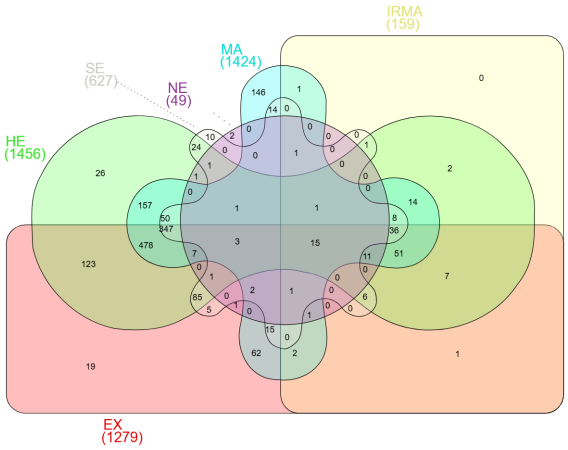

Figure 8. Concept entanglement diagram for FGADR dataset, showcasing how often all concepts coincide in a single image.

First, we horizontally flip all left-eye images to ensure consistent orientation across the dataset, so that all eyes appear as right-eye views. This will not introduce bias, as all training images can be randomly flipped.

For creating backgrounds, we start with defining the set of healthy images. We take all images that do not have any lesion annotations. Then, a subjective visual assessment is made to exclude healthy images with irregular patterns. Although possibly introducing bias, this is done to ensure the remaining candidates can serve as a background that does not introduce much noise.

After these steps, we apply the concept creation pipeline to create 128 positive and negative concept image pairs. This was a trade-off decision: while more concepts would always be preferable for creating stable LCRs, increasing the number would require reusing source images more often, reducing variability across concept examples. Note that the number of 128 could be sufficient, as natural images only need a couple dozen images [15]. An example of such a pair is shown in Figure 9.

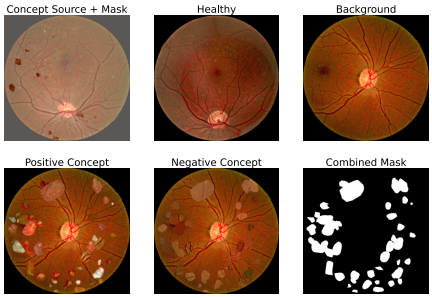

Figure 9. Example of concept creation pipeline on the fundus images from the FGADR dataset for concept: *Soft Exudates*. Done on non-preprocessed data for visualisation, as patches can be seen more easily without correcting for colour.

Internal Consistency and Domain Robustness. We conducted experiments to evaluate whether using multiple source images per concept pair improves the quality of the learnt LCRs. Results of these experiments are shown in Figure 10. To this end, we measured the accuracy of the internal DB of filter-CAVs, computed across all layers of a ResNet50 trained on the APTOS dataset.

While high DB scores indicate that positive and negative concept examples are well-separated in the latent space, this alone does not confirm that the resulting LCRs are meaningful. Rather, it provides a necessary, but not sufficient, condition for their utility in LCRReg. If the concepts were not consistently separable, the resulting DBs and CAVs would be no more informative than random directions in the representation space.

In addition, we evaluated the CAVs on a model trained on the FGADR dataset, the same dataset from which the concepts were derived. This experiment was conducted to assess whether concept performance improves significantly when evaluated in-domain. If performance on the FGADR-trained model is not substantially higher than on a model trained on APTOS, this suggests that the concepts are robust to data shift, as their utility does not strongly depend on the training domain of the model.

Results. We found that using only a single source image per concept pair resulted in near-random DB scores, indicating poor separability in the latent space. Increasing the number of source images improved performance: using five images led to substantially better results, and using ten yielded further improvements. We did not extend beyond ten images due to limited data availability and the need to preserve variability across concept examples. DB accuracies were comparable between models trained on APTOS

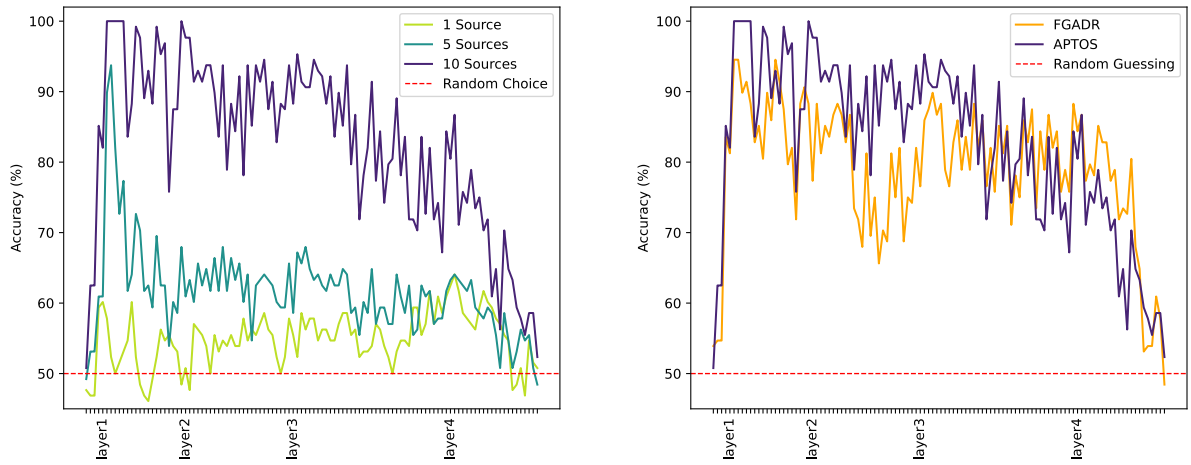

Figure 10. Illustrations of the internal coherence score per layer of a ResNet50 trained on APTOS. Filter-CAV was calculated using 128 images, and evaluated using 64 other images, using accuracy of the internal decision boundary. The train-test split was created using different source images. (a) shows the relation between number of source images per concept example, while (b) depicts the effect of changing the models' training data.

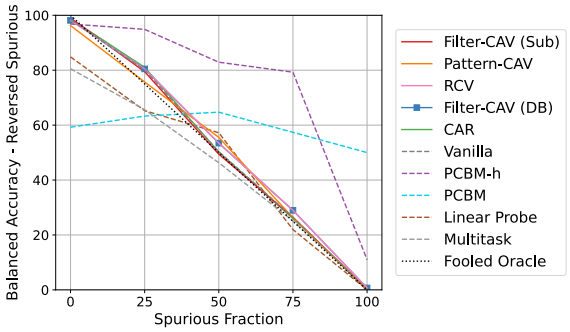

Figure 11. Performance averaged over 5 runs of various LCRReg models, PCBM-h, and the vanilla model, assessed across multiple values of p_{SC}. Models are using hyperparameters that are tuned using Optuna, with the validation set being of similar distribution as the training set.

and FGADR, suggesting that the learnt concepts are robust to data shift and do not rely heavily on the training domain.

B. Ablations

In this Appendix, we report further results of ablations studies. In particular, Figure 12 and Figure 13 show additional ablations on hyperparameters and training strategies on the Elements dataset, discussed in Section 3.1. Table 2 shows performance of LCRReg on different model architectures, which is addressed in Section 3.2.3.

C. Robustness to Spurious Correlations

Figure 11 reports the results of the different models on the APTOS dataset with the hyperparameters finetuned with Optuna, commented in Section 3.2.1. Note that we also provide the accuracy of PCBM, which is equal to PCBM-h without residual fitting. However, this is not discussed in the main section due to it not achieving performance significantly different than random performance, and it is thus not reported in Figure 6 in the Main Text.

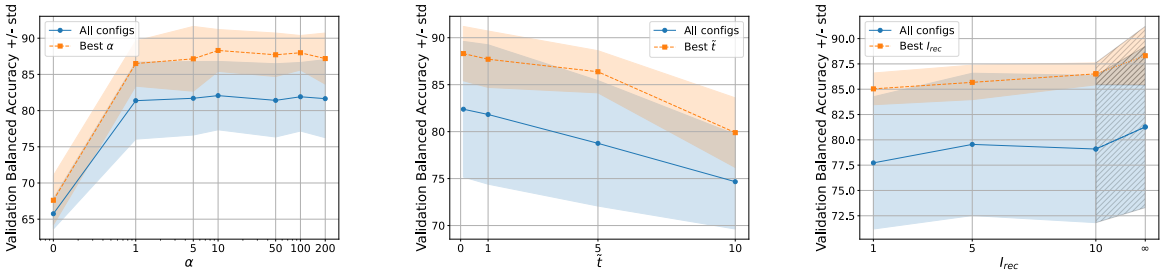

Figure 12. Results of hyperparameter grid search on the binary classification task (Elements dataset) using filter-CAV with cosine loss. Each configuration was run five times with different train/validation splits. Blue lines: mean ± std over all combinations of remaining parameters. Orange lines: mean ± std with best setting of other parameters per fixed value. (a) Weight of the $\mathcal{L}_{LCRReg}$: α_t. (b) Starting epoch: $\tilde{t}$. (c) Recomputation interval: I_{rec}.

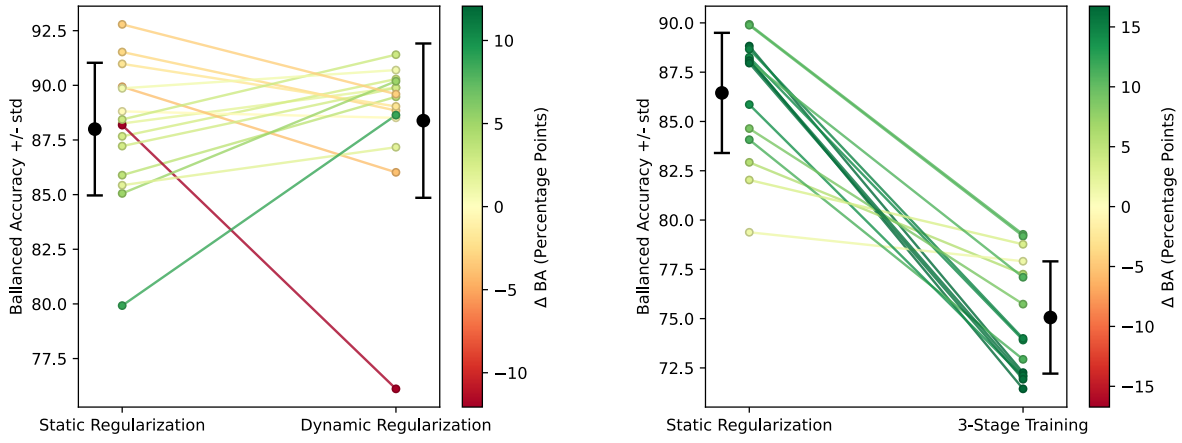

Figure 13. Comparison of different Static regularisation to both Dynamic regularisation, and 3–Stage Training, evaluating all approaches on 15 distinct train–test dataset pairs. The figure shows both the Balanced Accuracy (BA) of individual runs, and the total average. Statistical comparison was done via pairwise t-testing. Comparison of Static regularisation vs. (a) Dynamic regularisation and vs. (b) 3–Stage Training.

Table 2. Performance of Latent Concept Representation-based regularisation across different datasets and model architectures, trained and evaluated on APTOS. Trained on a Spurious Dataset with $p_{SC} = 1$. Performance reported in terms of balanced accuracy. Parameter counts (in millions) are shown in parentheses below model names.

Dataset	**Model**	**ResNet18** (11.6 M)	**ResNet152** (60.2 M)	**DenseNet121** (8.0 M)	**InceptionV3** (23.9 M)
Reverse Spurious	Vanilla	14.81 (9.87)	14.60 (9.76)	43.71 (13.10)	17.96 (4.32)
	Regularised	31.47 (15.55)	16.82 (16.07)	65.74 (11.83)	49.12 (6.27)
Base	Vanilla	75.32 (5.70)	84.81 (6.16)	88.91 (4.15)	79.66 (6.16)
	Regularised	80.14 (5.72)	82.32 (1.56)	85.26 (2.42)	66.25 (9.80)
Spurious	Vanilla	99.69 (0.23)	99.66 (0.26)	99.06 (0.50)	97.22 (1.47)
	Regularised	99.49 (0.30)	99.32 (0.86)	93.78 (3.58)	80.49 (12.0)