A. Details of Experimental Setup

A.1. Datasets

We evaluate our methods on four real-world image tasks: *CUB*, *AwA2*, *WBCatt* and *7-point*.

- **CUB [14]**: the Caltech-UCSD Birds-200-2011 (CUB) dataset, which includes a total of 11,788 avian images including 4,796 training images 1,198 validation images and 5,794 testing images. The objective is to accurately categorize these birds into one of 200 distinct species. Following [9], we use k = 112 binary bird attributes representing wing color, beak shape, etc.
- **AwA2 [33]**: Animals with Attributes 2 consists of in total 37,322 images distributed in 50 animal categories. We use 80% for training, 10% for validation and 10% for testing. The AwA2 also provides a category-attribute matrix, which contains an 85-dim attribute vector (e.g., color, stripe, furry, size, and habitat) for each category.
- **WBCatt [42]**: the White Blood Cell Attributes dataset includes a total of 10,298 microscopic images from the PBC dataset [1] with class label, include 6,169 images for training, 1,030 images for validation and 3,099 images for testing. Each image is annotated with 11 morphological attributes (e.g., cell shape, chromatin density and granule color).
- **7-point [22]**: the Seven-Point Checklist Dermatology Dataset is designed for diagnosing and classifying skin lesions. It includes 1011 lesion case distribute in 5 diagnostic categories. There are 413 samples for training, 203 samples for validation and 395 samples for testing.

	CUB	AwA2	WBCatt	7-point
Images	11,788	37,322	10,298	1011
Classes	200	50	5	5
Concepts	112	85	31	19

Table 4. Statistics of the datasets used in our experiments.

A.2. Implementation Details

First, we resize the images to an input size of 299 x 299. The use of different symbols in Section 4 and Figure 2 is to indicate that the two components can employ different backbones, highlighting the method's generality. In our experiments, Ω and Ψ share parameters, and we employ ResNet34 [13] as the backbone to transform the input into latent code, followed by a fully connected layer to convert it into concept embeddings of size 16 (32 for CUB). During pseudo-labeling, we also utilize ResNet34 with the KNN algorithm with k = 2. Additionally, for obtaining concept labels using a threshold, we set the threshold to 0.6. We set $\lambda_1 = 1$ and $\lambda_2 = 0.1$ and utilize the SGD optimizer with a learning rate of 0.05 and a regularization coefficient of 5e-6. We train SSCBM for 100 epochs with a batch size of 256 (for AwA2, the batch size is 32 due to the large size of individual images). We repeat each experiment 5 times and report the average results.

To construct the concept saliency map, we first upsample the heatmaps $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k\}$ calculated in Section 4.2 to the size $H \times W$ (the original image size). Then, we create a mask based on the value intensities, with higher values corresponding to darker colors.

A.3. Impact of Different Backbones

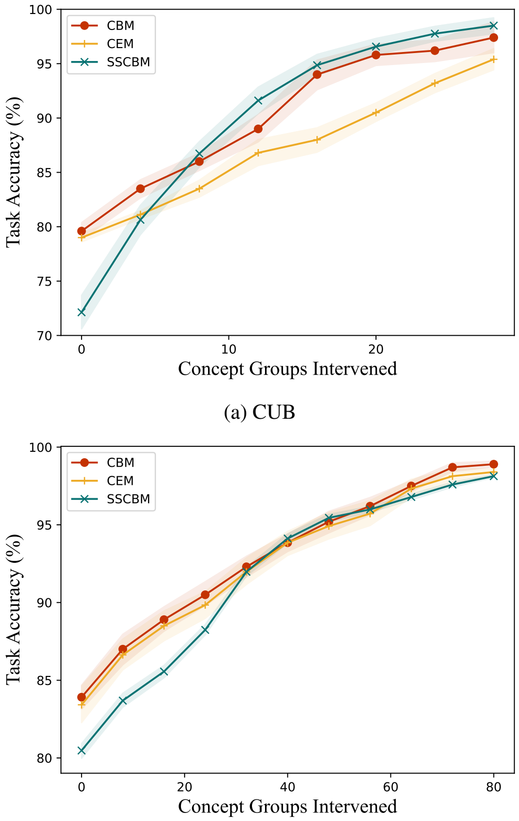
Here, we evaluate the performance of SSCBM and the baseline using different backbones (ResNet18, ResNet34, and ResNet50). We present the results in Table 5. We observe that using ResNet34 as the backbone achieves a significant performance improvement compared to ResNet18. However, for ResNet50, its performance is almost on par with ResNet34, and in some cases, it even performs worse. We analyze that a possible reason could be that ResNet50 has a significantly larger number of parameters relative to the rest of the model, making it difficult to converge simultaneously with other parts during training, which leads to suboptimal results. We also note that [9] encounters a similar situation, where ResNet34 is used as the backbone.

B. Test-time Intervention

For AwA2, there are no grouped concepts, so we adopt individual intervention. In the CUB, we do the group intervention, i.e., intervene in the concepts with associated attribution. For example, the breast color::yellow, breast color::black, and breast color::white are the same concept group. So, we only need to correct the concept label in the group. We expect that the model performance will steadily increase along with the ratio of concept intervention, indicating that the model learned such correct label information and automatically corrected other labels.

Results in Figure 5 demonstrate our model's robustness and an increasing trend to learn the information of concept information, indicating our interpretability and model prediction performance. Here, we train SSCBM with a label ratio of 0.1 and compare its performance with CEM and CBM trained on the full dataset. It can be observed that without any intervention, the task performance of SSCBM is lower than the accuracy of the supervised model. However, as the number of intervened concept groups increases, the prediction accuracy of SSCBM gradually improves, eventually achieving comparable performance to CBM and CEM when interventions are applied to all concept groups. This lies in our loss of alignment in effectively learning the correct information pairs in unlabeled and labeled data.

Here, we present some successful examples of Test-time Intervention illustrated in Figure 19. The first two on the left show examples from the CUB dataset. In the top left



(a) CUB

(b) AwA2

Figure 5. Test-time Intervention on CUB and AwA2 dataset.

image, by changing the wing color to brown, we successfully caused the model to predict the Great Crested Flycatcher instead of the Swainson Warbler. In the bottom left, because the model initially failed to notice that the upper part of the bird was black, it misclassified it as Vesper Sparrow. Through test-time intervention, we successfully made it predicted the bird was a Grasshopper Sparrow. The results on the right side of the image are from the AwA2 dataset. We successfully made the model predict correctly by modifying concepts at test time. For example, in the top right image, by modifying the concept of 'fierce' for the orca, we prevented it from being predicted as a horse. In the bottom right, we successfully made the model recognize the bat through the color of the bat.

C. Additional Interpretability Evaluation

We provide our additional interpretability evaluation in Figure 6 - 13 for CUB dataset, Figure 14 - 16 for WBCatt dataset, and Figure 17 - 18 for 7-point dataset as follows. Image regions that are highly relevant to the concept are highlighted.

D. Limitations

While we solve a small portion of annotation problems by semi-supervised learning, semi-supervised models may not be suitable for all types of tasks or datasets. It is more effective that the data distribution is smooth. However, this is the limitation of semi-supervised learning, not our methods.

E. Broader Impact

The training of current CBMs heavily relies on the accuracy and richness of annotated concepts in the dataset. These concept labels are typically provided by experts, which can be costly and require significant resources and effort. Additionally, concept saliency maps frequently misalign with input saliency maps, causing concept predictions to correspond to irrelevant input features - an issue related to annotation alignment. In this problem, we propose SSCBM, a strategy to generate pseudo labels and an alignment loss to solve these two problems. Results show our effectiveness. This method has practical use in the real world.

Table 5. Performance of different backbones under different ratios of labeled data.

Dataset	**Backbone**	**Ratio**	**CBM+SSL**		**CEM+SSL**		**SSCBM**	
			Concept	**Task**	**Concept**	**Task**	**Concept**	**Task**
CUB	ResNet18	K=1	81.69%	7.92%	80.44%	54.90%	87.50%	60.77%
		0.05 (K=2)	84.67%	7.46%	83.16%	60.53%	89.16%	64.31%
		0.1 (K=3)	85.01%	8.54%	82.46%	59.56%	90.28%	66.76%
		0.15 (K=4)	85.17%	10.20%	83.81%	64.27%	90.55%	66.40%
		0.2 (K=5)	85.42%	9.70%	84.35%	64.22%	91.02%	68.31%
	ResNet34	K=1	83.11%	5.51%	82.36%	59.35%	88.99%	66.72%
		0.05 (K=2)	84.51%	8.35%	83.72%	62.20%	90.04%	67.43%
		0.1 (K=3)	84.96%	9.84%	84.03%	63.12%	90.88%	67.67%
		0.15 (K=4)	85.47%	9.96%	84.30%	64.14%	91.47%	68.36%
		0.2 (K=5)	86.67%	16.43%	86.83%	67.64%	92.09 %	70.07%
	ResNet50	K=1	81.43%	9.75%	78.99%	57.39%	89.03%	69.59%
		0.05 (K=2)	84.49%	8.42%	83.00%	62.12%	90.91%	71.73%
		0.1 (K=3)	85.09%	7.97%	83.15%	63.27%	92.10%	74.09%
		0.15 (K=4)	85.14%	9.56%	83.32%	64.43%	91.75%	68.50%
		0.2 (K=5)	85.39%	10.67%	83.71%	66.09%	93.21%	75.77%
WBCatt	ResNet18	K=1	72.21%	99.45%	73.06%	99.42%	89.07%	99.58%
		0.05 (K=62)	80.83%	99.65%	76.75%	99.19%	93.76%	99.55%
		0.1 (K=124)	83.33%	99.81%	78.74%	98.87%	93.54%	99.55%
		0.15 (K=186)	85.10%	99.84%	82.96%	99.52%	93.83%	99.55%
		0.2 (K=247)	84.87%	99.74%	84.11%	99.42%	94.18%	99.58%
	ResNet34	K=1	79.06%	99.39%	70.27%	98.64%	91.48%	99.13%
		0.05 (K=62)	81.08%	99.48%	73.82%	99.52%	93.53%	99.61%
		0.1 (K=124)	85.48%	99.32%	72.25%	99.29%	93.98%	99.68%
		0.15 (K=186)	85.39%	99.68%	72.68%	99.58%	94.42%	99.71%
		0.2 (K=247)	87.07%	99.74%	74.14%	99.52%	94.42%	99.71%
	ResNet50	K=1	71.30%	99.71%	72.99%	99.48%	88.46%	99.71%
		0.05 (K=62)	82.69%	99.45%	69.74%	99.35%	93.73%	99.61%
		0.1 (K=124)	81.23%	99.71%	72.22%	98.32%	93.99%	99.55%
		0.15 (K=186)	82.97%	99.74%	87.17%	99.23%	94.32%	99.71%
		0.2 (K=247)	84.40%	99.84%	85.03%	99.77%	94.47%	99.61%
7-point	ResNet18	K=1	56.08%	56.46%	56.30%	61.52%	58.19%	64.56%
		0.05 (K=5)	63.58%	59.75%	64.32%	62.28%	67.24%	66.08%
		0.1 (K=9)	65.78%	63.29%	66.65%	64.30%	71.03%	68.61%
		0.15 (K=13)	67.71%	63.54%	67.02%	67.34%	73.34%	69.87%
		0.2 (K=17)	69.54%	57.22%	68.79%	63.80%	75.02%	70.63%
	ResNet34	K=1	59.91%	55.95%	62.78%	66.09%	66.58%	66.84%
		0.05 (K=5)	65.36%	57.47%	67.85%	67.09%	70.98%	68.77%
		0.1 (K=9)	68.82%	55.70%	72.23%	66.33%	73.67%	70.09%
		0.15 (K=13)	66.14%	59.75%	66.54%	67.09%	73.94%	72.56%
		0.2 (K=17)	70.29%	55.44%	73.04%	66.84%	76.52%	74.56%
	ResNet50	K=1	55.95%	64.81%	56.46%	64.56%	59.89%	64.30%
		0.05 (K=5)	62.80%	60.00%	64.66%	66.08%	69.29%	66.84%
		0.1 (K=9)	65.14%	60.25%	66.48%	67.34%	68.54%	64.56%
		0.15 (K=13)	69.15%	57.47%	67.94%	66.33%	73.58%	67.85%
		0.2 (K=17)	71.03%	58.73%	67.97%	67.09%	74.27%	64.05%

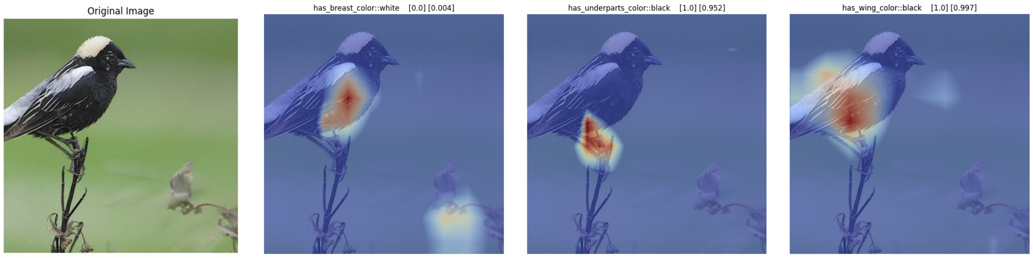

Original Image
has_breast_color::white [0.0] [0.004]
has_underparts_color::black [1.0] [0.952]
has_wing_color::black [1.0] [0.997]

Figure 6. Concept saliency map on CUB dataset (bobolink) shows reasonable localization of the ground truth concept regions.

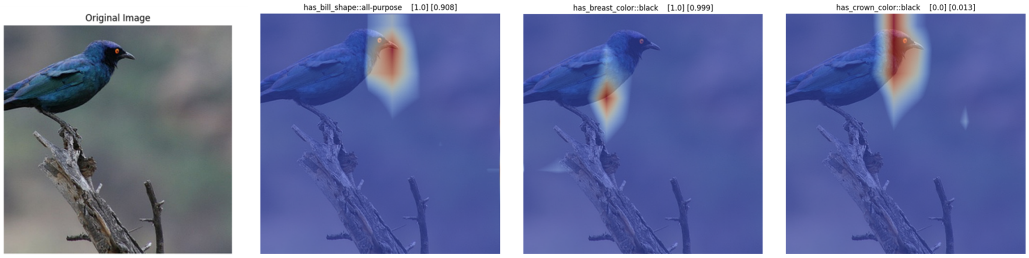

Original Image
has_bill_shape::all-purpose [1.0] [0.908]
has_breast_color::black [1.0] [0.999]
has_crown_color::black [0.0] [0.013]

Figure 7. Concept saliency map on CUB dataset (cape glossy starling) shows reasonable localization of the ground truth concept regions.

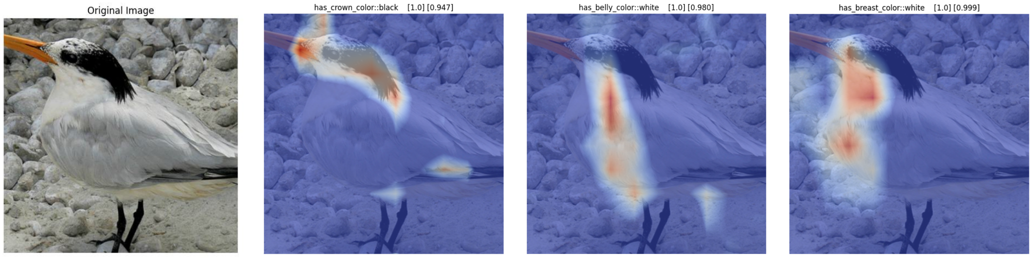

Original Image
has_crown_color::black [1.0] [0.947]
has_belly_color::white [1.0] [0.980]
has_breast_color::white [1.0] [0.999]

Figure 8. Concept saliency map on CUB dataset (elegant tern) shows reasonable localization of the ground truth concept regions.

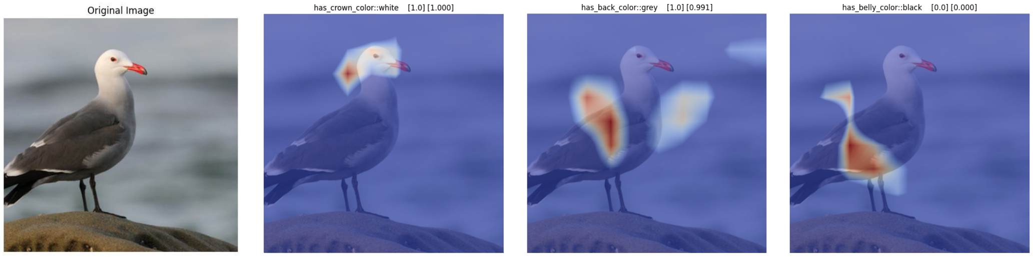

Original Image
has_crown_color::white [1.0] [1.000]
has_back_color::grey [1.0] [0.991]
has_belly_color::black [0.0] [0.000]

Figure 9. Concept saliency map on CUB dataset (heermann gull) shows reasonable localization of the ground truth concept regions.

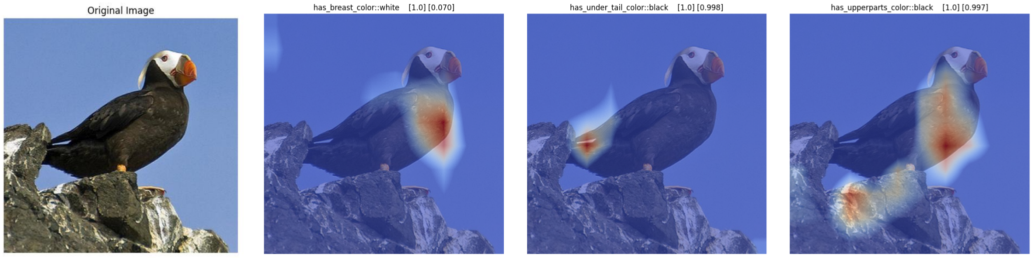

Original Image
has_breast_color::white [1.0] [0.070]
has_under_tail_color::black [1.0] [0.998]
has_upperparts_color::black [1.0] [0.997]

Figure 10. Concept saliency map on CUB dataset (horned puffin) shows reasonable localization of the ground truth concept regions.

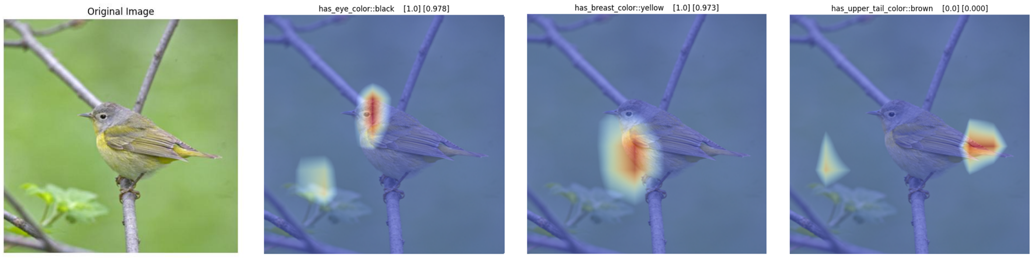

Original Image
has_eye_color::black [1.0] [0.978]
has_breast_color::yellow [1.0] [0.973]
has_upper_tail_color::brown [0.0] [0.000]

Figure 11. Concept saliency map on CUB dataset (nashville warbler) shows reasonable localization of the ground truth concept regions.

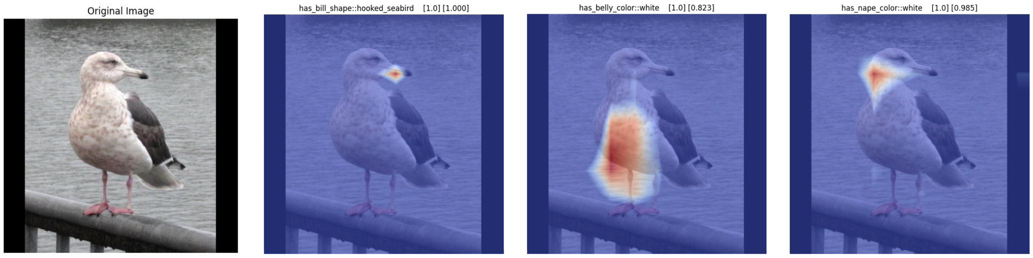

Original Image
has_bill_shape::hooked_seabird [1.0] [1.000]
has_belly_color::white [1.0] [0.823]
has_nape_color::white [1.0] [0.985]

Figure 12. Concept saliency map on CUB dataset (slaty backed gull) shows reasonable localization of the ground truth concept regions.

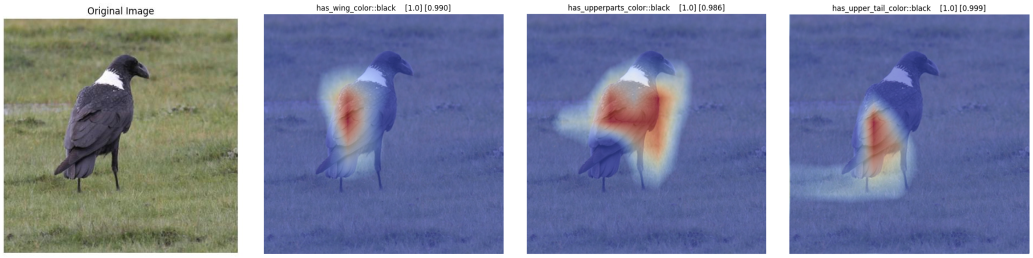

Original Image
has_wing_color::black [1.0] [0.990]
has_upperparts_color::black [1.0] [0.986]
has_upper_tail_color::black [1.0] [0.999]

Figure 13. Concept saliency map on CUB dataset (white necked raven) shows reasonable localization of the ground truth concept regions.

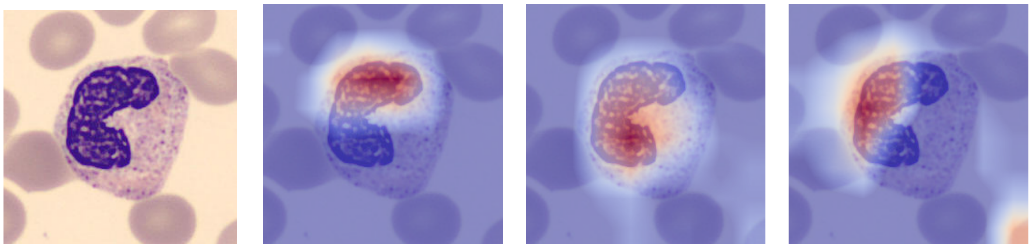


Figure 14. Concept saliency map on WBCatt dataset (neutrophil) shows reasonable localization of the ground truth concept regions.

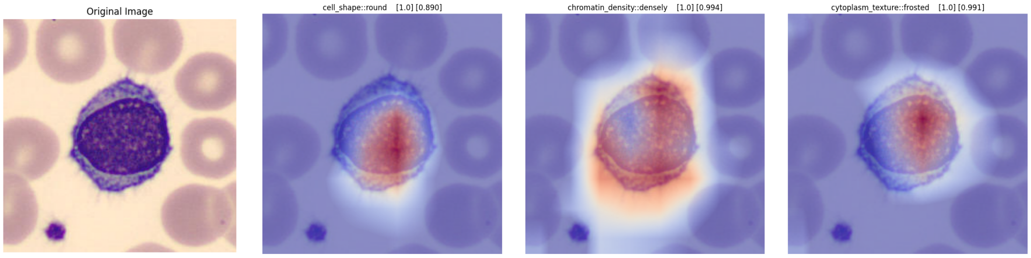

Original Image
cell_shape::round [1.0] [0.890]
chromatin_density::densely [1.0] [0.994]
cytoplasm_texture::frosted [1.0] [0.991]

Figure 15. Concept saliency map on WBCatt dataset (lymphocyte) shows reasonable localization of the ground truth concept regions.

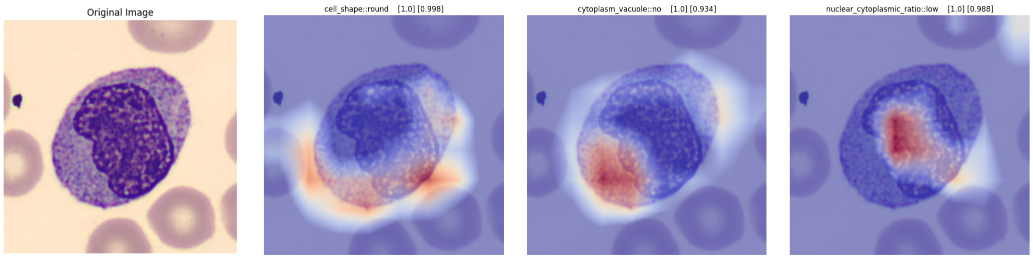

Original Image
cell_shape::round [1.0] [0.998]
cytoplasm_vacuole::no [1.0] [0.934]
nuclear_cytoplasmic_ratio::low [1.0] [0.988]

Figure 16. Concept saliency map on WBCatt dataset (monocyte) shows reasonable localization of the ground truth concept regions.

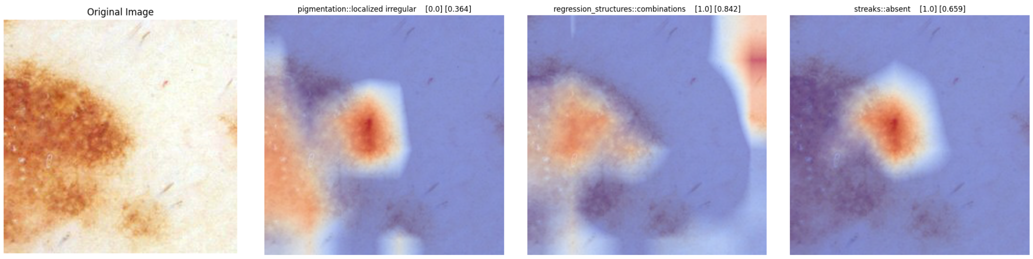

Original Image
pigmentation::localized irregular [0.0] [0.364]
regression_structures::combinations [1.0] [0.842]
streaks::absent [1.0] [0.659]

Figure 17. Concept saliency map on 7-point dataset (congenital nevus) shows reasonable localization of the ground truth concept regions.

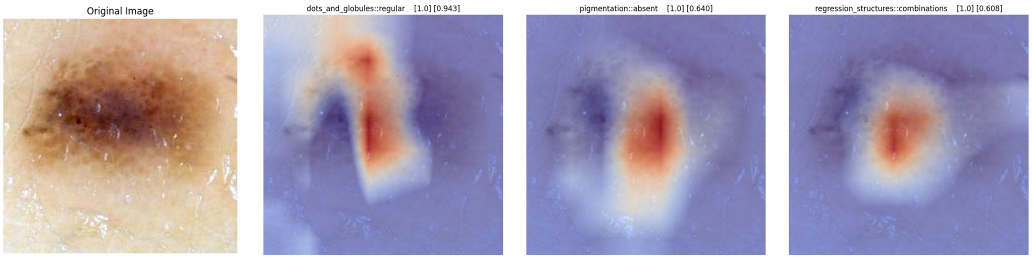

Original Image
dots_and_globules::regular [1.0] [0.943]
pigmentation::absent [1.0] [0.640]
regression_structures::combinations [1.0] [0.608]

Figure 18. Concept saliency map on 7-point dataset (melanoma female) shows reasonable localization of the ground truth concept regions.

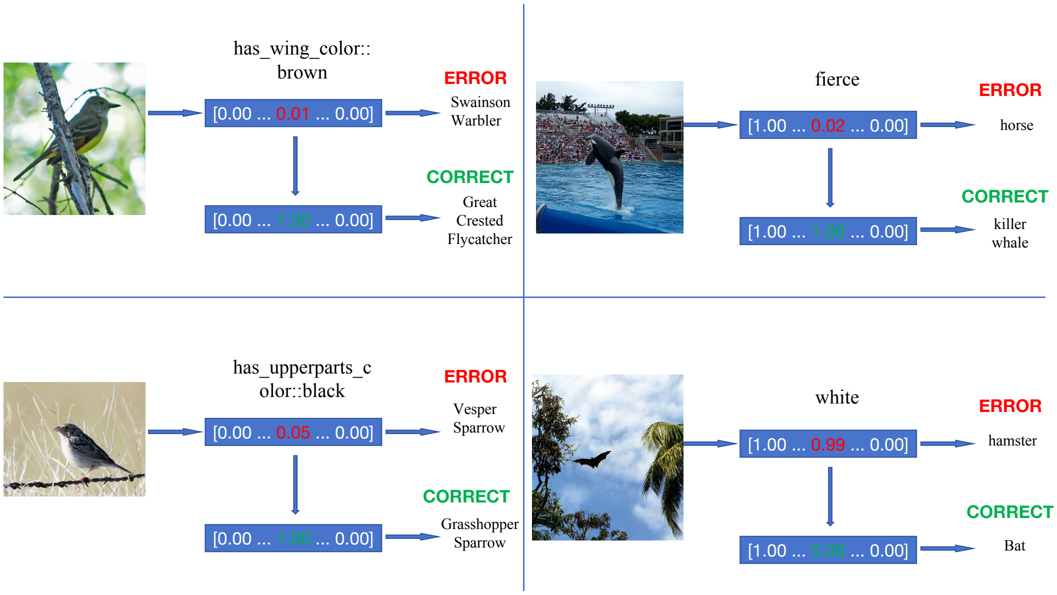


Figure 19. Examples of Test-time Intervention.