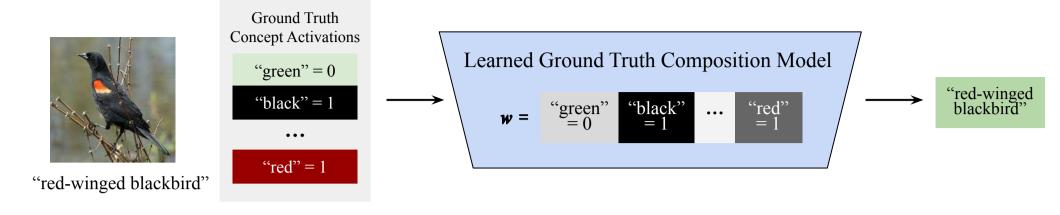
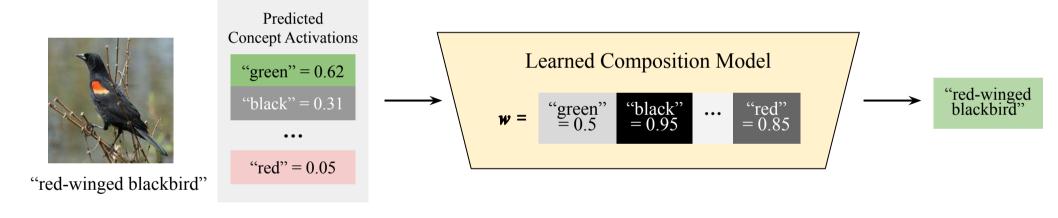
(a) A learned ground truth composition model trained with ground truth concept activations. During inference time, the model takes in ground truth concept activations and predict the correct bird class.



(b) A composition model learned from concept activations predicted by a vision-and-language (VL) model. The model is affected by the green background, and thus learns a weight of 0.5 for concept "green".



(c) After training this learned composition model with predicted concept activations. We *intervene* this model with ground truth concept activations during inference time. The model predicts the wrong bird class because the learned weights reflect a positive correlation between the green background and *red-winged blackbird*.

