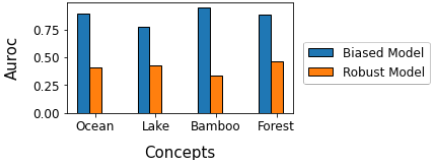
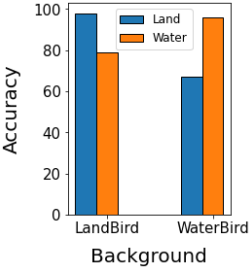


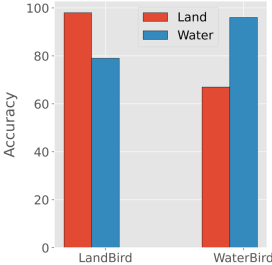
Auroc scores of the background concepts



Accuracy scores the biased Blackbox

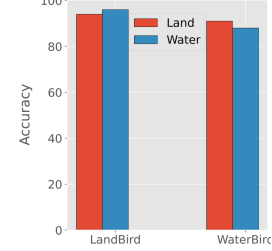


Accuracy scores the biased Blackbox



(a)

Accuracy scores MoE from the robust Blackbox



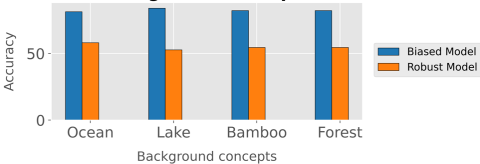
(b)



**Biased Blackbox**  
Groundtruth: WaterBird  
Prediction : LandBird  
Explanation : LandBird ↔ WingShapeRoundedwings ∧ Forest

**Robust Blackbox**  
Groundtruth: WaterBird  
Prediction : WaterBird  
Explanation : WaterBird ↔ BillLengthAboutTheSameAsHead  
∧ ¬BillLengthShorterThanHead ∧ ¬SizeSmall5...9in  
∧ ¬ShapePerchingLike ∧ CrownColorWhite

Accuracy scores of the background concepts



(d)

(e)