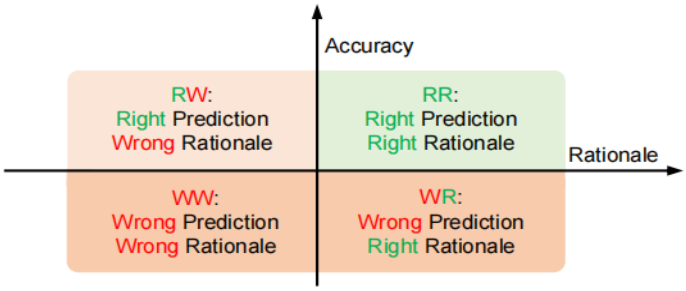
FLCP, and FT drop 6.4%, 5.65%, and 4.07% respectively, on ImageNet-1K (Russakovsky et al. 2015) dataset. Notably, existing work (Goyal et al. 2023) highlights the effectiveness of fine-tuning for VLMs, asserting that FLCP consistently improves prediction accuracy and should be considered the "standard" method for fine-tuning CLIP. However, our findings suggest that this conclusion does not hold when evaluating the rationality of VLM predictions. This discrepancy underscores the importance of considering different possibilities when evaluating prediction rationality.



Lastly, we conducted ablation studies to verify the consistency of our findings, which remain consistent across various experimental settings, including different training optimizers, learning rates, explanation heatmap methods, and fine-tuning techniques such as prompt tuning (Zhou et al. 2022) and adapter tuning (Zhang et al. 2022).

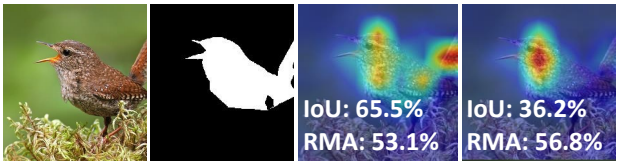


Figure 1: Both (a) and (b) have low responses to the background while (a) pays more attention to the whole body of the bird and (b) pays more attention to the discriminative feature of the bird (head). Compared with the IoU score between (a) and (b), the difference between them is negligible. Moreover, both achieve correct predictions. Input is from CUB-200-2011 (Wah et al. 2011) dataset. "GT" denotes abbreviation of "Ground Truth" and "Explan" denotes abbreviation of "Explanation".

Preliminaries

There has been a surge of people exploring VLMs for their downstream tasks. A typical way is to use them for image classification (Goyal et al. 2023). In our prediction evaluations, we study the image classification task and measure model performances using the top-1 accuracy metric.

We evaluate whether the model provides valid evidence for its predictions by examining whether the explanation heatmap generated by VLMs focuses on the target objects. Specifically, a heatmap that strongly highlights key object regions while showing minimal responsiveness to background pixels indicates valid evidence. Therefore, we rely on the "Relevant Mass Accuracy (RMA)" score (Arras, Osman, and Samek 2022; Brandt, Raatjens, and Gaydadjiev 2023), which satisfies this criterion by measuring how much "mass" one method assigns to pixels within the region of target objects (ground truth). RMA score is calculated by determining the ratio of the total heatmap pixel values within the target object regions, to the sum of all pixel values across the entire heatmap. It requires both the generated explanation heatmap (H) from VLMs and the ground truth explanation mask (M), whose pixels on the target objects are marked as 1 otherwise marked as 0. RMA score is defined as:

$$\text{RMA}(H, M) = \frac{\sum H \odot M}{\sum H}, \tag{1}$$

where $\odot$ represents Hadamard product. Note that the evaluations from many studies (Selvaraju et al. 2020; Arras, Osman, and Samek 2022) require the presence of ground-truth mask for heatmap localization.

We emphasize that the RMA metric provides a more reasonable evaluation for classification tasks compared to metrics like "Intersection over Union (IoU)" used in other works. For instance, Grad-CAM (Selvaraju et al. 2020) relies on the IoU score to measure the overlap between the explanation heatmap and the ground truth mask. However, the IoU score fails to reasonably evaluate two vastly different yet valid pieces of evidence. In Figure 1, we show two explanation heatmaps, (a) and (b), that are from different models.