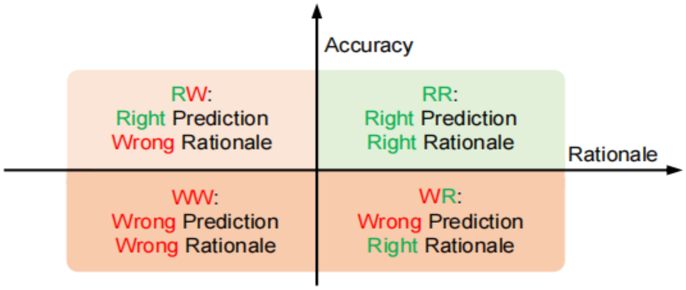
FLCP, and FT drop 6.4%, 5.65%, and 4.07% respectively, on ImageNet-1K (Russakovsky et al. 2015) dataset. Notably, existing work (Goyal et al. 2023) highlights the effectiveness of fine-tuning for VLMs, asserting that FLCP consistently improves prediction accuracy and should be considered the "standard" 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.



[redacted]

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).

[redacted]

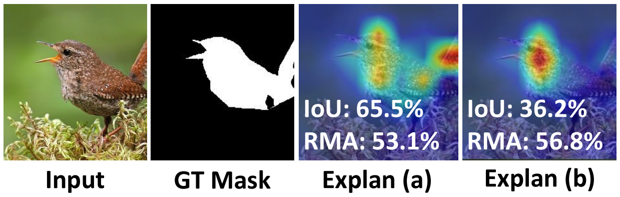


Figure 1: Both (a) and (b) have low responses to the background while (a) pays more attention to the whole body of the bird and (b) pays more attention to the discriminative feature of the bird (head). Compared with the IoU score between (a) and (b), the difference between them is negligible. Moreover, 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:

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

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.

Evaluations	Methods	VLMs	Datasets				Avg.
			ImageNet-1K	CalTech-101	Stanford-Dogs	CUB-200-2011	
Prediction Trustworthiness (PT, %) ↑	ZS	ALBEF-ViT-B/16	90.61	76.28	95.02	49.31	**71.26**
		BLIP-ViT-B/16	89.01	61.72	93.95	23.93	
		CLIP-ViT-B/16	87.05	62.99	92.96	29.38	
		CLIP-ViT-B/32	89.39	73.44	94.58	30.57	
	LP	ALBEF-ViT-B/16	82.37	59.08	90.30	19.91	64.78
		BLIP-ViT-B/16	80.36	52.57	92.63	12.98	
		CLIP-ViT-B/16	80.65	56.40	92.19	36.17	
		CLIP-ViT-B/32	84.05	68.22	92.76	35.89	
	FLCP	ALBEF-ViT-B/16	87.07	62.43	92.68	64.33	67.95
		BLIP-ViT-B/16	82.57	59.52	91.46	36.17	
		CLIP-ViT-B/16	81.40	64.32	76.44	16.56	
		CLIP-ViT-B/32	85.48	71.29	91.59	23.84	
	FT	ALBEF-ViT-B/16	86.28	48.97	92.22	24.98	67.01
		BLIP-ViT-B/16	85.54	39.96	93.13	25.85	
		CLIP-ViT-B/16	82.98	56.86	91.60	27.98	
		CLIP-ViT-B/32	86.29	80.01	94.17	55.43	
Inference Reliability (IR, %) ↑	ZS	ALBEF-ViT-B/16	48.95	76.74	30.56	16.43	56.65
		BLIP-ViT-B/16	49.65	90.05	33.87	18.92	
		CLIP-ViT-B/16	66.33	85.58	61.96	68.05	
		CLIP-ViT-B/32	61.09	85.23	56.12	56.80	
	LP	ALBEF-ViT-B/16	74.76	92.56	66.21	55.46	75.67
		BLIP-ViT-B/16	74.78	90.89	65.11	59.91	
		CLIP-ViT-B/16	78.93	95.05	75.41	77.08	
		CLIP-ViT-B/32	74.91	93.76	68.53	67.37	
	FLCP	ALBEF-ViT-B/16	78.54	96.81	78.36	80.92	81.71
		BLIP-ViT-B/16	80.04	94.93	78.73	73.40	
		CLIP-ViT-B/16	75.00	94.84	80.21	77.97	
		CLIP-ViT-B/32	71.84	95.46	76.43	73.87	
	FT	ALBEF-ViT-B/16	82.95	94.41	81.93	81.87	**83.52**
		BLIP-ViT-B/16	82.86	91.55	79.18	85.26	
		CLIP-ViT-B/16	83.25	90.66	82.06	81.12	
		CLIP-ViT-B/32	79.34	93.86	73.22	72.72	

Table 2: Comparisons of four methods with proposed “PT” and “IR” metrics. Here we observe that mainstream fine-tuning methods come with both strengths and weaknesses. We show that fine-tuning mostly leads to a worse capability of prediction trustworthiness but enhances the inference reliability of VLMs than the ZS method. The best-averaged score among the four methods is **bolded**, while the second-place averaged score is underlined.

Goyal et al. 2023) are limited to the impact of mainstream VLM fine-tuning methods regarding predictive accuracies, ignoring their positive impacts on VLM prediction rationality. In this paper, we have analyzed and explored the benefits of fine-tuning VLMs from a new perspective. Our experimental results show that fine-tuning has its merits and is not completely worthless for the prediction rationality of VLMs.

In summary, we conducted extensive experiments to validate the existing mainstream VLM fine-tuning methods in terms of both their strengths and weaknesses from a prediction rationality perspective. On the one hand, fine-tuning leads to good inference reliability: when provided with valid evidence of target objects, fine-tuned VLMs are more likely to generate accurate predictions. On the other hand, we also confirm that mainstream fine-tuning methods tend to hurt the inherent capabilities of VLMs, specifically in terms of prediction trustworthiness. These are aspects that merit attention from the community of machine learning.

Analysis on Out-of-Distribution Data

Question: Will out-of-distribution data change our observations?

Answer: No, all findings remain consistent.

Distributional shifts has garnered significant attention in the field of machine learning (Qiao, Zhao, and Peng 2020; Qiao and Peng 2023). During the fine-tuning, the distributional discrepancy between the fine-tuning and testing data is worth considering. Real-world data distributions can change due to factors such as time, location, and environment. Testing on out-of-distribution data helps simulate these changes, ensuring the model performs well in diverse scenarios. For example, in autonomous driving, the models need to remain stable in multiple weather conditions.

In this section, we study the fine-tuning methods when testing on out-of-distribution data. Here we use the ImageNet-C dataset, which includes multiple corruption categories and levels of severity. As shown in Figure 4, our key findings are

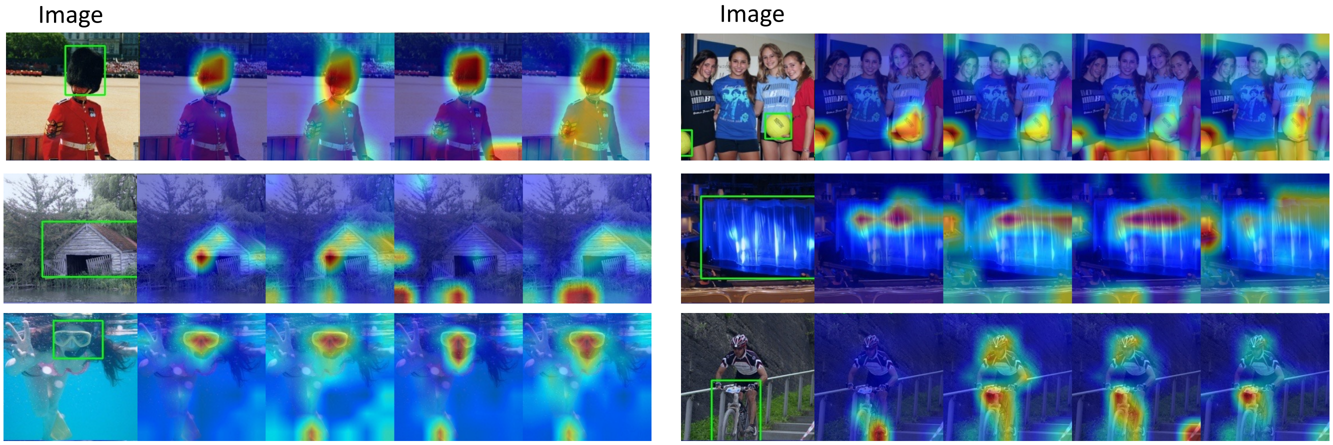
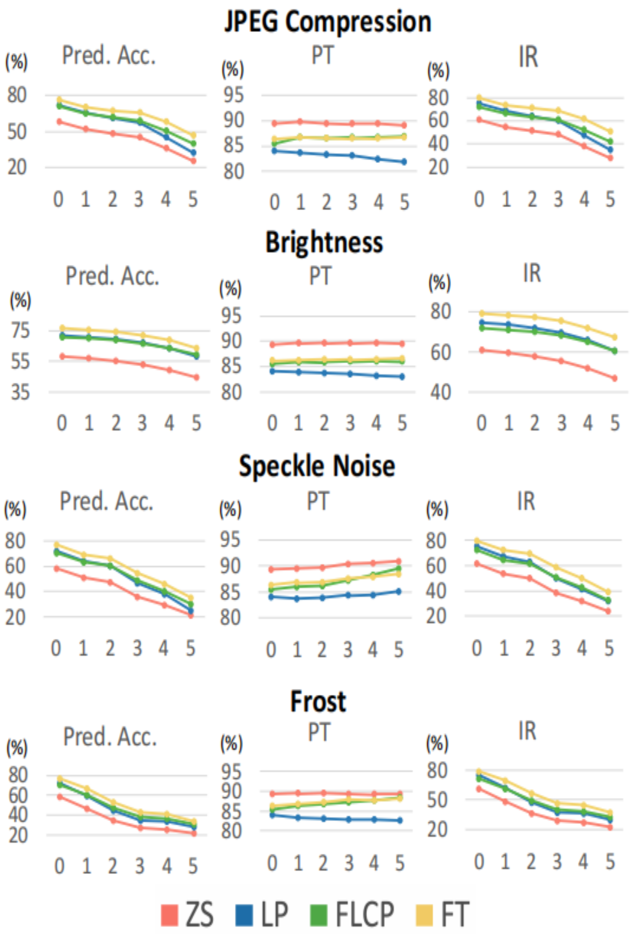


Figure 3: Visualization comparisons among different methods. Compared with zero-shot (ZS), current mainstream fine-tuning methods (LP, FLCP, and FT) for VLMs tend to show enhanced responses in background pixels that are irrelevant to predictions. Here we select the samples for which all four methods make correct predictions. Here we display bounding box annotations indicating the positions of the predicted target.

as follows:

1. Fine-tuning on in-distribution data [illegible]



and 1e−5 for FT. For the original learning rate settings

regarding other models and datasets please refer to our supplementary material. Note that the aforementioned three experiments are conducted with the CLIP-ViT-B/32 model on the ImageNet-1K.

As shown in Table 3, *our findings remain unaffected* with multiple setups. On the one hand, prevalent fine-tuning approaches tend to increase the instances with correct predictions based on invalid evidence, despite the enhancement in prediction accuracy. On the other hand, fine-tuning typically demonstrates strong inference reliability.

Recently, there have been other fine-tuning techniques proposed by the community including prompt tuning (Zhou et al. 2022), and adapter tuning (Zhang et al. 2022). We find that *our findings are also consistent under these fine-tuning methods.* Due to the space limits please refer to our supplementary material for the related experimental results and introduction of these methods.

Related Works

Multimodal Foundation Models

In recent years, there has been a surge of interest in research regarding Vision-Language Models (VLMs). These VLMs (Radford et al. 2021; Li et al. 2021, 2022c; Singh et al. 2022; Jia et al. 2021; Li et al. 2022a,e; Yuan et al. 2021; Li et al. 2022d, 2023; Chen and Wang 2022; Zhong et al. 2022; Kim, Son, and Kim 2021; Chen et al. 2020), have attracted substantial attention due to their remarkable capacity to achieve robust performance, both in zero-shot and fine-tuned scenarios, across a diverse spectrum of vision-language-related tasks (Antol et al. 2015; Vinyals and Le 2015; Xie et al. 2019; Suhr et al. 2017). Notably, CLIP (Radford et al. 2021), as a prominent exemplar in this domain, has also demonstrated exceptional zero-shot performance in image classification. The contrastive learning approach it employs has also found applications in fields such as mul-