Beyond Accuracy: On the Effects of Fine-tuning Towards Vision-Language Model's Prediction Rationality

Qitong Wang, Tang Li, Kien X. Nguyen, Xi Peng

DeepREAL Lab, Department of Computer & Information Sciences, University of Delaware
{wqtwjt, tangli, kxnguyen, xipeng}@udel.edu

Abstract

Vision-Language Models (VLMs), such as CLIP, have already seen widespread applications. Researchers actively engage in further fine-tuning VLMs in safety-critical domains. In these domains, prediction rationality is crucial: *the prediction should be correct and based on valid evidence*. Yet, for VLMs, the impact of fine-tuning on prediction rationality is seldomly investigated. To study this problem, we proposed two new metrics called *Prediction Trustworthiness* and *Inference Reliability*. We conducted extensive experiments on various settings and observed some interesting phenomena. On the one hand, we found that the well-adopted fine-tuning methods led to more correct predictions based on invalid evidence. This potentially undermines the trustworthiness of correct predictions from fine-tuned VLMs. On the other hand, having identified valid evidence of target objects, fine-tuned VLMs were more likely to make correct predictions. Moreover, the findings are also consistent under distributional shifts and across various experimental settings. We hope our research offer fresh insights to VLM fine-tuning.

Code — https://github.com/deep-real/vlm-pred-rationality

Introduction

Vision-Language Models (VLMs), such as CLIP (Radford et al. 2021), have recently begun to see widespread adoption in high-stakes applications, such as healthcare (Wang et al. 2022b) and autonomous driving (Chen et al. 2023). A common practice in utilizing VLMs involves undertaking further fine-tuning (Goyal et al. 2023; Wortsman et al. 2022a; Wang et al. 2022b) in these models to their specific tasks rather than training deep models from scratch. While existing studies have evaluated mainstream fine-tuning methods, they have primarily focused on prediction accuracy (Kumar et al. 2022; Wortsman et al. 2022b; Goyal et al. 2023), overlooking an essential aspect: *prediction rationality*, where model predictions should not only be accurate but also grounded in valid evidence. Besides, the current academic community widely accepts that "clearly explaining a rationale for a clas-

sification decision to an end-user can be as important as the decision itself." (Hendricks et al. 2016) One significant reason is that neglecting the model's prediction rationality will cause severe consequences in safety-critical domains. For example, doctors employ a fine-tuned VLM which can accurately predict the presence of cancer tumors from X-ray images, to help decision-making. If its predictions are based on erroneous reasons: the input's background instead of tumor region, doctors will lack trust in fine-tuned VLM, leading them to disregard the usage of the model. Therefore, in this paper, we study a crucial yet seldom investigated question: *how do mainstream fine-tuning methods affect the rationality of VLM predictions?*

To systematically study this question, we propose two new metrics to evaluate the rationality of VLM predictions after fine-tuning: (1) Prediction Trustworthiness (PT): the ratio of correct predictions with valid evidence overall correct predictions. (2) Inference Reliability (IR): the percentage of correct predictions given that the model has identified valid evidence of target objects. To assess whether the model focuses on valid evidence for the image classification task, we measure if the generated explanation heatmap from VLMs focuses on the target objects, based on the "Relevant Mass Accuracy (RMA)" score (Brandt, Raatjens, and Gaydadjiev 2023). We study the mainstream methods including "Zero-Shot" (ZS), "Linear-Probing" (LP), "Finetune Like CLIP Pretrain" (FLCP), and standard "Fine-tuning" (FT). We conducted extensive experiments and have obtained novel and consistent findings. Our results reveal that widely used fine-tuning methods exhibit significant limitations, yet they also possess certain advantages. Our key findings are summarized as follows:

Will mainstream fine-tuning methods hurt the rationality of VLM predictions? Surprisingly yes! With our proposed "Prediction Trustworthiness" metric, fine-tuning results in more appearance of samples with correct predictions based on invalid evidence than zero-shot, making the correct predictions untrustworthy. For instance, with the ALBEF-ViT-B/16 model, compared with ZS, the PT scores of LP, FLCP, and FT drop 17.2%, 13.85% and 27.31% respectively, on CalTech-101 (Li et al. 2022b) dataset, despite improving prediction accuracies. And with the CLIP-ViT-B/16 model, compared with ZS, the PT scores of LP,

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.

Will valid evidence help enhance predictions made by fine-tuned VLMs? Yes. Using our "Inference Reliability" metric, we find that when VLMs focus on valid evidence of target objects, the prediction accuracy of fine-tuned VLMs improves. For example, in the ImageNet-1K dataset, with the CLIP-ViT-B/16 model, LP, FLCP, and FT outperform ZS in IR scores by 12.6%, 8.67%, and 16.92% respectively. Existing works (Kumar et al. 2022; Wortsman et al. 2022b; Goyal et al. 2023), which study the positive impacts of VLM fine-tuning, are limited to the prediction accuracies. Our research provides insights into the impact of fine-tuning VLMs from a novel perspective, highlighting the benefits of fine-tuning in terms of enhancing prediction rationality.

Will out-of-distribution data change our observations? No. There is a critical need to make sure that models work reliably in real-world situations, where the data distribution they encounter might be different from what they were trained on. For instance, the model must maintain stability and effectiveness in autonomous driving applications across various weather conditions. In parallel, previous work (Radford et al. 2021) has demonstrated the remarkable predictive performance of CLIP in both in-distribution and out-of-distribution data. Therefore, we discuss how our observations might change in the context of out-of-distribution data. We find that all our findings remain consistent across various types and magnitudes of distributional shifts, as demonstrated through experiments in ImageNet-C (Hendrycks and Dietterich 2018).

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).

Our contribution lies in discovering new findings through extensive experiments across various benchmarks including ImageNet (Russakovsky et al. 2015), which are typical and widely used in the community. We provide novel insights about both the strengths and weaknesses of widely adopted fine-tuning strategies for VLMs, from the perspective of the rationality of VLM predictions. Moreover, our findings remain consistent across evaluation scenarios involving both in-distribution data and out-of-distribution data, as well as under various experimental settings. This paper provides new insights for people to rethink the effects of mainstream fine-tuning methods for VLMs.

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:

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

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.

Figure 2: Overview of the four quadrants (RR, RW, WR, WW) of Accuracy and Rationale that are utilized to evaluate prediction rationality.

Even though the IoU metric treats them differently, both of them achieve correct predictions with valid evidence. They both exhibit a low response to background pixels. (a) pays attention to the whole body of the bird. (b) is also reasonable because it effectively identifies the distinguishing features of the bird, despite not highlighting the more complete bird region as in (a). This indicates that compared to IoU, RMA evaluation can fairly treat two distinct but valid evidence.

Explanation Heatmap Generation. The method we use is directly from "Generic Attention Attribution" (Chefer, Gur, and Wolf 2021). In this case, the heatmaps are generated from attention maps of the transformer-based model, which is one of the most well-adopted methods, used in recent works including (Mao et al. 2023). It has been demonstrated in existing work (Liu et al. 2022) that it achieves the best faithfulness performance among all well-known explanation methods when applied to transformer-based models. The main idea is Hadamard's product between attention maps and their gradient to the output. It is defined as:

$$\overline{\mathbf{A}} = \mathbb{E}_h((\nabla \mathbf{A} \odot \mathbf{A})^+), \qquad (2)$$

where $\odot$ is the Hadamard product, $\nabla \mathbf{A} := \frac{\partial y_t}{\partial \mathbf{A}}$ for y_t which is the model's output for the class t that we wish to visualize. $\mathbb{E}_h$ is the mean across the heads dimension. The $+$ indicates that the negative contributions are removed before averaging. Note that the class we explain are based on the index given by the annotations instead of predictions.

Our Proposed Evaluations

We present our evaluation protocols with two criteria in mind. (1) A trustworthy VLM should not produce instances of invalid evidence among samples with correct predictions. (2) When focusing on the correct predicted objects, a reliable VLM should leverage such valid evidence to achieve correct predictions. To determine whether the evidence (or rationale) of the model is correct, we use a threshold of 0.5 on the RMA measure. Specifically, an RMA score of 0.5 or above is considered valid evidence and vice versa. As a result, we achieve four scenarios: RR, RW, WR, and WW (Figure 2) that are used to formalize our two novel metrics:

1. **Prediction Trustworthiness (PT)**. A dependable and trustworthy model should generate valid evidence that cor-responds to accurate predictions. Hence, we introduce the "PT" metric, which calculates the proportion of samples where the prediction is "right" and its evidence is also valid or "right" (RR) among all samples with right predictions, defined as:

$$PT = \frac{RR}{RR + RW}, \qquad (3)$$

where "RW" denotes data with the "right" classifications based on invalid or "wrong" evidence. It is evident that an increase in the number of RW samples, i.e. irrational predictions, results in a decrease in PT scores.

2. **Inference Reliability (IR)**. Given that the model could pinpoint the regions of target objects, a reliable model should make correct predictions. Consequently, we introduce the "IR" metric, which calculates the proportion of samples with correct prediction and valid evidence among all samples with valid evidence of target objects, defined as:

$$IR = \frac{RR}{RR + WR}, \qquad (4)$$

where "WR" denotes data with incorrect classifications with valid evidence. An increase in the number of WR samples results in a decrease in IR scores.

Experiments

Experimental Setup

Fine-tuning Methods. In this paper, we study fundamental methods including: (1) Zero-Shot (ZS), (2) Linear-Probing (LP), (3) Finetune Like CLIP Pretrain (FLCP), and (4) Finetuning (FT). For detailed information on these methods, please refer to our supplementary material in the extended version of our paper.

Models. We study four VLMs: the first two models are CLIP-ViT-B/32 & 16 (Radford et al. 2021) from OpenAI, which manifest powerful zero-shot performances on image classification. The next two models are ALBEF-ViT-B/16 (Li et al. 2021), pretrained on 14M image-text pairs, and BLIP-ViT-B/16 (Li et al. 2022c), pretrained on 129M image-text pairs, both developed by Salesforce. Their performances on the image classification task are also investigated in many works (Jonathan Roberts and Albanie 2023; Wang et al. 2022a).

Fine-tuning Setups. We maintain a consistent batch size and training epoch across all three fine-tuning methods (LP, FLCP, FT) for the same dataset and model. We employ the Adam (Kingma and Ba 2014) optimizer during the fine-tuning. For more details about fine-tuning, please consult our supplementary material.

Datasets. In this paper, we conduct experiments on several datasets, including ImageNet (Russakovsky et al. 2015), CalTech-101 (Li et al. 2022b), Stanford-Dogs (Khosla et al. 2011), CUB-200-2011 (Wah et al. 2011), and ImageNet-C (Hendrycks and Dietterich 2018). In CUB-200-2011 and CalTech-101 datasets, the 0-1 segmentation mask annotations directly serve as ground truth explanation masks. For

Methods	VLMs	Datasets				Avg.
		IN	CT	SD	CUB	
ZS	ALBEF-ViT-B/16	46.48	77.02	29.25	12.43	
	BLIP-ViT-B/16	46.30	85.89	32.38	16.88	53.74
	CLIP-ViT-B/16	63.30	84.22	60.61	54.94	
	CLIP-ViT-B/32	58.41	84.79	54.62	52.33	
LP	ALBEF-ViT-B/16	72.03	90.38	65.10	48.46	
	BLIP-ViT-B/16	72.46	90.26	64.23	47.77	72.50
	CLIP-ViT-B/16	76.69	94.64	74.14	70.14	
	CLIP-ViT-B/32	72.21	93.09	67.27	61.08	
FLCP	ALBEF-ViT-B/16	77.58	95.85	77.88	77.27	
	BLIP-ViT-B/16	78.67	94.99	77.89	68.85	80.99
	CLIP-ViT-B/16	72.41	96.20	80.70	80.76	
	CLIP-ViT-B/32	70.81	95.74	75.70	74.49	
FT	ALBEF-ViT-B/16	80.82	95.91	81.32	80.48	
	BLIP-ViT-B/16	80.75	92.74	78.68	68.98	**81.63**
	CLIP-ViT-B/16	81.19	93.03	81.56	79.25	
	CLIP-ViT-B/32	76.62	94.30	72.42	68.07	

Table 1: Comparisons of four methods regarding prediction accuracy (%). The best-averaged score among the four methods is **bolded**, while the second-place averaged score is underlined. Due to the space limit, we abbreviate the names of datasets. Here, "IN", "CT", "SD", "CUB" denote "ImageNet-1K", "CalTech-101", "Stanford-Dogs", "CUB-200-2011" respectively.

images with bounding box annotations surrounding predicted instances (ImageNet, ImageNet-C, Stanford-Dogs), we generate ground truth explanation masks as follows: given initial masks whose pixel values are all zero, we mark the mask areas within boxes as one. For more detailed information about these datasets, please refer to the supplementary material.

Weaknesses of Fine-tuning

Question: Will mainstream fine-tuning methods hurt the rationality of VLM predictions?

Answer: Surprisingly yes! The well-adopted fine-tuning methods decrease the trustworthiness of VLM predictions in most settings: causing more samples with correct predictions based on invalid evidence.

Although fine-tuning is able to improve the prediction accuracies of VLMs (see Table 1), we find mainstream fine-tuning methods lead to worse prediction trustworthiness, as shown in Table 2. For instance, in the ImageNet-1K dataset, with CLIP-ViT-B/16 model, compared with ZS, fine-tuning deteriorates "Prediction Trustworthiness (PT)" performances by 6.4%, 5.65% and 4.07% respectively. Our experimental results confirm the significant drawbacks of mainstream fine-tuning methods for VLMs: fine-tuning results in more instances where predictions are correct but the evidence which VLMs base on is invalid. This results in a reduced level of trustworthiness to VLM predictions. Lastly, there are rare exceptions with increased PT scores. This is likely due to the low zero-shot prediction accuracy of ALBEF (12.43%) and BLIP (16.88%) on CUB-200-2011. Fine-tuning introduces the missing knowledge to these models,

leading to increased PT.

To further support our observation, we provide visualizations of the explanation heatmaps in Figure 3. We observe that widely adopted fine-tuning methods often amplify the responses of VLMs to pixels containing information irrelevant to the predicted objects. For instance, from the leftmost first-row comparisons, fine-tuning makes VLMs enhance responses on the human body or background instead of the hat (predicted category). Here we only show results on the CLIP-ViT-B/32 model with ImageNet-1K datasets due to space constraints. Please refer to our supplementary material for more visualizations.

Why does finetuning decrease trustworthiness? (1) VLMs tend to exploit the easiest path to minimize loss during finetuning, often picking up on spurious correlations or shortcuts present in the data. For instance, if all images of a particular class contain a common watermark or background, VLMs may associate that feature with the class label instead of learning the actual characteristics of the object. (2) Standard fine-tuning objectives usually prioritize improving prediction accuracy, but they do not account for the validity of the evidence used. As a result, there is no built-in mechanism to guide the model to focus on valid evidence.

In recent years, there have been some discussions regarding the excellence of fine-tuning for VLMs. For example, existing work (Goyal et al. 2023) shows that FLCP leads to uniformly better prediction performances. They claim that FLCP should be adopted as the "standard" method for finetuning CLIP. However, based on our discoveries, we contend that this conclusion doesn't apply when considering the rationality of VLM predictions. Although FLCP significantly enhances VLMs' prediction accuracies, we find that FLCP leads VLMs to provide more invalid evidence when making correct predictions, weakening the prediction trustworthiness of VLMs than ZS. This disparity highlights the significance of considering different possibilities when evaluating VLMs' prediction rationality.

Strengths of Fine-tuning

Question: Will valid evidence help enhance predictions made by fine-tuned VLMs?

Answer: Yes, they exhibit good inference reliability; i.e., when focusing on the valid evidence of target objects, fine-tuned VLMs are more likely to make correct predictions.

This phenomenon indicates better inference reliability of fine-tuning compared with ZS, as shown in Table 2. For example, in the ImageNet-1K dataset, with the CLIP-ViT-B/16 model, LP, FLCP, and FT outperform ZS by 12.6%, 8.67%, and 16.92% respectively; with the CLIP-ViT-B/32 model, LP, FLCP, and FT outperform ZS by 13.82%, 10.75%, and 18.25% respectively. This indicates that fine-tuning approaches contribute to less WR than ZS. When VLMs identify valid evidence for target objects, fine-tuning is more likely to produce correct predictions.

Existing works (Kumar et al. 2022; Wortsman et al. 2022b;

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	
		BLIP-ViT-B/16	89.01	61.72	93.95	23.93	**71.26**
		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	
		BLIP-ViT-B/16	80.36	52.57	92.63	12.98	64.78
		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	
		BLIP-ViT-B/16	82.57	59.52	91.46	36.17	<u>67.95</u>
		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	
		BLIP-ViT-B/16	85.54	39.96	93.13	25.85	67.01
		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	
		BLIP-ViT-B/16	49.65	90.05	33.87	18.92	56.65
		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	
		BLIP-ViT-B/16	74.78	90.89	65.11	59.91	75.67
		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	
		BLIP-ViT-B/16	80.04	94.93	78.73	73.40	<u>81.71</u>
		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	
		BLIP-ViT-B/16	82.86	91.55	79.18	85.26	**83.52**
		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 <u>underlined</u>.

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 can enhance the prediction accuracy for out-of-distribution data.

2. However, the mainstream fine-tuning methods still compromise the prediction trustworthiness of VLM, which brings more samples with correct prediction based on invalid evidence, compared with zero-shot.

3. Fine-tuning tends to enhance the inference reliability of VLMs: when focusing on correct prediction objects, fine-tuned VLMs are more likely to give correct predictions.

Therefore, we extend our previous findings to scenarios involving out-of-distribution data, demonstrating the consistency of our discoveries.

Our conclusions also remain unaffected when the prediction accuracies decrease caused by corruption strength increases. Therefore, we think our findings may not change with variations in model prediction accuracy.

Ablations studies

To ensure the consistency of our findings across different experimental settings, we perform a comprehensive series of ablation studies. We investigate the effects under different setups including: (1) Experiments with another popular optimizer: AdamW (Loshchilov and Hutter 2017). (2) Experiments with another widely-used explanation method: gradient of attention ($\nabla \mathbf{A}$) based (Serrano and Smith 2019) method. The main idea of this method is to utilize the gradient of attention to the output as an explanation heatmap, where $\nabla \mathbf{A} := \frac{\partial y_t}{\partial \mathbf{A}}$ for y_t which is the model's output for the class t. (3) Results with different fine-tuning learning rates (abbreviated as "LR"): $5e-4$ for "LP", $3e-6$ for "FLCP", and $2e-5$ for "FT", compared with the original setup, where we set learning rates as $1e-3$ for "LP", $5e-6$ for "FLCP", 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-

Figure 4: Experimental results on out-of-distribution data. Our discoveries remain consistent across various types and magnitudes of distributional shifts. The x-axis in all figures represents the strength of corruption, where a strength of 0 indicates the results of different methods on the original ImageNet validation data. Due to space constraints, we only show results with CLIP-ViT-B/32 and four types of corruption in the main paper. For more results, please refer to our supplementary material.

tiview analysis (Tian, Krishnan, and Isola 2020) and egocentric video understanding (Wang et al. 2023). Recently, researchers have engaged in fine-tuning (Goyal et al. 2023) VLMs to better adapt them to specific downstream tasks. However, the impact of such training on the prediction rationality of these models remains an open research problem, one that warrants in-depth exploration and investigation.

Explainable Machine Learning

Explainable Machine Learning (XML) is crucial for promoting transparency, trust, accountability, and fairness in AI systems. Researchers frequently employ techniques to explain neural network operations and decision-making regarding input data. Activation heatmaps such as Grad-CAM (Selvaraju et al. 2020), visualize important regions for specific classes. In light of the proliferation of transformer-based models (Dosovitskiy et al. 2020), researchers start exploring the feasibility of utilizing attention maps, taking it as a way to provide explanations (Chefer, Gur, and Wolf 2021). In order to evaluate the quality of these explanation generation methods, existing works including (Petsiuk,

Setup	Evaluations	Methods			
		ZS	LP	FLCP	FT
AdamW Optimizer	Pred. Acc.(%) ↑	58.41	72.22	70.88	**76.53**
	PT(%) ↑	**89.39**	84.23	85.53	86.46
	IR(%) ↑	61.09	74.93	71.93	**79.26**
∇A Explanation Heatmap	Pred. Acc.(%) ↑	58.41	72.21	70.81	**76.62**
	PT(%) ↑	**74.79**	63.62	65.21	65.76
	IR(%) ↑	61.18	75.18	71.87	**79.68**
Different LRs Compared with Original	Pred. Acc.(%) ↑	58.41	72.28	70.05	**75.51**
	PT(%) ↑	**89.39**	84.72	86.19	86.59
	IR(%) ↑	61.09	74.91	71.21	**78.62**

Table 3: Ablation studies with prediction accuracy, and our proposed "Prediction Trustworthiness (PT)" and "Inference Reliability (IR)" metrics. Our findings are unaffected under different experimental setups. The best score is bolded.

Das, and Saenko 2018) study from the perspective of faithfulness; i.e., how accurately an explanation method reflects the true decision-making process of a model. In parallel, Mao et al. (Mao et al. 2023) propose the concept of a reliable model, emphasizing the importance of the "doubly-right" criterion: both accurate predictions and fine-grained language explanations of model decision-making. Recently, some works (Li, Ma, and Peng 2024a,b) have increasingly required VLM models to deliver not only accurate predictions but also correct rationales. In this paper, we explore the impact of widely accepted fine-tuning methods on the prediction rationality of VLMs for vision tasks such as image classification, providing novel insights about VLM fine-tuning within the XML research community. And we highlight that faithfulness is beyond the scope of our study due to two reasons. On the one hand, faithfulness evaluations primarily focus on assessing the correctness of heatmap explanation methods. On the other hand, existing work (Liu et al. 2022) verified the superiority of our employed explanation generation method.

Conclusion

Prediction rationality is an important aspect to consider when fine-tuning Vision-Language Models (VLMs), especially in high-stakes applications. This paper provides a comprehensive assessment of the commonly used fine-tuning approaches, presenting some insights on both advantages and disadvantages. On the one hand, they generally demonstrate strong inference reliability. More specifically, when focusing on the valid evidence of target objects, the fine-tuned VLMs are more likely to make correct predictions. On the other hand, fine-tuning often results in undermining the trustworthiness of VLM predictions by bringing more data samples with correct predictions based on invalid evidence. We further observe that our discoveries are consistent across various types and magnitudes of distributional shifts, and remain unaffected with multiple setups. To ensure that VLMs can be reliably used in high-stack applications, it will be crucial to study new fine-tuning methods that can improve VLM prediction rationality. We leave it as future works. We expect our research may provide useful experience and advance the study of VLM fine-tuning.

Acknowledgements

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Arras, L.; Osman, A.; and Samek, W. 2022. CLEVR-XAI: A Benchmark Dataset for the Ground Truth Evaluation of Neural Network Explanations. *Inf. Fusion*, 81(C): 14–40.

Brandt, R.; Raatjens, D.; and Gaydadjiev, G. 2023. Precise Benchmarking of Explainable AI Attribution Methods. *arXiv preprint arXiv:2308.03161*.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 397–406.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023. CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Chen, X.; and Wang, X. 2022. PaLI: Scaling Language-Image Learning in 100+ Languages. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Goyal, S.; Kumar, A.; Garg, S.; Kolter, Z.; and Raghunathan, A. 2023. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19338–19347.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 3–19. Springer.

Hendrycks, D.; and Dietterich, T. 2018. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jonathan Roberts, K. H.; and Albanie, S. 2023. SATIN: A Multi-Task Metadataset for Classifying Satellite Imagery using Vision-Language Models. *arXiv preprint arXiv:2304.11619*.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A.; Ma, T.; Liang, P.; and Raghunathan, A. 2022. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *Uncertainty in Artificial Intelligence*, 1041–1051. PMLR.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Yang, J.; and Liu, Z. 2023. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Li, D.; Li, J.; Li, H.; Niebles, J. C.; and Hoi, S. C. 2022a. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4953–4963.

Li, F.-F.; Andreeto, M.; Ranzato, M.; and Perona, P. 2022b. Caltech 101.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022c. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022d. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, T.; Ma, M.; and Peng, X. 2024a. Beyond Accuracy: Ensuring Correct Predictions With Correct Rationales. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Li, T.; Ma, M.; and Peng, X. 2024b. DEAL: Disentangle and Localize Concept-level Explanations for VLMs. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022e. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *International Conference on Learning Representations*.

Liu, Y.; Li, H.; Guo, Y.; Kong, C.; Li, J.; and Wang, S. 2022. Rethinking attention-model explainability through faithfulness violation test. In *International Conference on Machine Learning*, 13807–13824. PMLR.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mao, C.; Teotia, R.; Sundar, A.; Menon, S.; Yang, J.; Wang, X.; and Vondrick, C. 2023. Doubly Right Object Recognition: A Why Prompt for Visual Rationales. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2722–2732.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Qiao, F.; and Peng, X. 2023. Topology-aware Robust Optimization for Out-of-Distribution Generalization. In *The Eleventh International Conference on Learning Representations*.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to Learn Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.*, 128(2): 336–359.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy: Association for Computational Linguistics.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Suhr, A.; Lewis, M.; Yeh, J.; and Artzi, Y. 2017. A Corpus of Natural Language for Visual Reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Vinyals, O.; and Le, Q. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, J.; Chen, D.; Wu, Z.; Luo, C.; Zhou, L.; Zhao, Y.; Xie, Y.; Liu, C.; Jiang, Y.-G.; and Yuan, L. 2022a. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35: 5696–5710.

Wang, Q.; Zhao, L.; Yuan, L.; Liu, T.; and Peng, X. 2023. Learning from Semantic Alignment between Unpaired Multiviews for Egocentric Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3307–3317.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022b. Med-CLIP: Contrastive Learning from Unpaired Medical Images and Text. In *Conference on Empirical Methods in Natural Language Processing*.

Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022a. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7959–7971.

Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022b. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7959–7971.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, 493–510. Springer.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16793–16803.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Supplementary Material

This section contains supplementary material to support the main paper text. It includes:

- Additional experimental results with out-of-distribution data. These include more corruption types and more VLMs compared with the main paper, which served as the extension of Figure 4.

- Visualizations with ALBEF-ViT-B/16, BLIP-ViT-B/16, and CLIP-ViT-B/16 models (Extension of Figure 3).

- More detailed description of the fine-tuning methods utilized in our main paper (Extension of paragraph titled "Fine-tuning Methods" in Section "Experimental Setup").

- More detailed description of the datasets utilized in our study (Extension of paragraph titled "Datasets" in Section "Experimental Setup").

- More implementation details of fine-tuning VLMs, which served as the extension of paragraph titled "Fine-tuning Setups" in Section "Experimental Setup".

- Experiments with more fine-tuning techniques for VLMs, which served as the extension of Section "Ablation Studies".

Additional Results on Out-of-Distribution Data

Here, we utilize CLIP-ViT-B/16 and CLIP-ViT-B/32 models. For CLIP-ViT-B/16, we introduce six types of corruption from the ImageNet-C dataset. Compared to the main paper's employment of CLIP-ViT-B/32 (see Figure 4), we incorporate two additional types of corruption. As shown in Figure 5 and Figure 7, with more corruption types and VLMs, our conclusions are consistent with those presented in the main paper.

Figure 5: Additional results of testing on out-of-distribution data with CLIP-ViT-B/32 model.

Addditional Visualizations

From Figure 6, we find that mainstream fine-tuning methods still deteriorate the prediction ratinoality of VLMs besides using the CLIP-ViT-B/32 model (shown in Figure 3). More

specifically, fine-tuning tends to show enhanced responses to the background information, in contrast with ZS. For example, from the rightmost second row, we find that fine-tuning methods show more responses to pixels (such as floor, and human) compared with ZS. These pixels are irrelevant to the predicted category: "basketball". Another example is that from the leftmost third row, we find that fine-tuning methods show more responses to pixels that are not related to the prediction object: "hay". Our visualization evidence clarifies that compared with ZS VLMs, fine-tuned VLMs tend to base on unreasonable evidence even with correct predictions.

Detailed Information of Fine-tuning Methods

1. **Zero-Shot (ZS)**. The model is presented with images along with textual descriptions of target classes, where we follow (Radford et al. 2021) and use the template: "a photo of a c_i" given k classes $\{c_1, c_2, ..., c_k\}$. Here, we directly load the pretrained weights of the models and then evaluate by assigning the most similar class to each image among all classes in the dataset.

2. **Linear-Probing (LP)**. In this context, we develop a neural network designed to process natural images as input data. This model comprises two distinct components. The first component is the image encoder initialized with pre-trained weights obtained from the image encoder of VLMs. Importantly, the parameters of this image encoder are frozen throughout the fine-tuning process. The second component is the classification head, responsible for making predictions, which is trainable during the fine-tuning phase. During model training for each dataset, we utilize a loss function based on cross-entropy classification.

3. **Finetune Like CLIP Pretrain (FLCP)**. We load the pretrained weights of VLMs. We align the images along with textual descriptions of target classes in a contrastive manner during fine-tuning for each dataset, following the same training scenarios as the pretaining process of CLIP (Radford et al. 2021). And we follow the same evaluation protocol as ZS after fine-tuning.

4. **Fine-tuning (FT)**. In this case, we build the same neural network as LP method, with the same initialization and evaluation scenarios. The only difference is that the image encoder is trainable when fine-tuning the model.

Detailed Information of Datasets

We use datasets that are widely used in the community. The specific information about these datasets is as follows:

ImageNet (Russakovsky et al. 2015) is a large-scale image database that has played an important role in computer vision and deep learning research. Here we access **ImageNet-1K** which is one of the most commonly used subsets of ImageNet. **ImageNet-1K** is a large-scale image database that has played an important role in computer vision and deep learning research. It spans 1000 object classes and contains 1,281,167 training images, 50,000 validation images, and 100,000 test images. The bounding boxes for instances in this dataset are included. We utilize the validation set during evaluations.

Figure 6: Additional visualizations among different methods on ImageNet-1K validation sets. Here we also select the samples for which all four methods make correct predictions. We display bounding box annotations indicating the target positions.

CalTech-101 (Li et al. 2022b) contains pictures of objects belonging to 101 categories, which includes 40 to 800 images per category. The segmentation masks for instances in this dataset are included. The size of each image is roughly 300×200 pixels. We partition the dataset into a hold-out test set, consisting of 20% of the data, while the remaining 80% will be used for fine-tuning VLMs.

Stanford-Dogs (Khosla et al. 2011) contains images of 120 types of dogs from around the world. This dataset has been built using images and annotation from ImageNet for the task of fine-grained image categorization. It contains 20,580 images in total (12,000 for training and 8,580 for testing) and includes class labels, and bounding boxes for annotations.

CUB-200-2011 (Wah et al. 2011) is one of the most widely used datasets for fine-grained visual categorization tasks. It contains 11,788 images of 200 bird categories, 5,994 for training, and 5,794 for testing. Each image has one category label and one bird segmentation mask annotation.

ImageNet-C (Hendrycks and Dietterich 2018) is an open-source dataset that consists of algorithmically generated corruptions (such as blur, noise) applied to the ImageNet validation set. Similar to ImageNet-1K, each image is accompanied by a single bounding box annotation that delineates

the instances present within images. Unlike the four previously mentioned datasets, we only employ this dataset for evaluations of out-of-distribution data.

More VLM Finetuning Details

For all three studied fine-tuning methods (LP, FLYP, FT) in main paper, besides ImageNet-1K, where we fine-tune with four A6000 Nvidia GPUs, we fine-tune VLMs using one A6000 Nvidia GPU for the rest of the datasets. We set the batch size as 128 per GPU for the ImageNet-1K dataset, and 64 for the CalTech-101 and Stanford-Dogs datasets when finetuning. For the CUB-200-2011 dataset, We set the batch size as 64 when fine-tuning CLIP models and 128 when fine-tuning ALBEF and BLIP models. We set the training epochs as 10 for the ImageNet-1K and Stanford-Dogs datasets. For the CalTech-101 dataset, We set the training epochs as 10 when fine-tuning CLIP models and 5 when fine-tuning AL-BEF and 3 when fine-tuning BLIP models. For the CUB-200-2011 dataset, We set the training epochs as 20 when fine-tuning the ALBEF model and 10 when fine-tuning the rest of the models.

For all experiments, we use Python-3.9.16, PyTorch-1.9.1, TorchVision-0.10.1, and cudatoolkit-11.4.2. And the hyper-parameters of the Adam optimizer are all set as follows: betas= $(0.9, 0.999)$, eps= $1e - 8$, weight decay= 0.0. For

Evaluations	Methods	VLMs	Datasets				Avg.
			ImageNet-1K	CalTech-101	Stanford-Dogs	CUB-200-2011	
Prediction Accuracy (%) ↑	ZS	CLIP-ViT-B/16	63.30	84.22	60.61	54.94	64.15
		CLIP-ViT-B/32	58.41	84.79	54.62	52.33	
	CoOp	CLIP-ViT-B/16	76.06	92.17	73.54	68.17	75.80
		CLIP-ViT-B/32	71.21	92.28	68.95	64.03	
	TA	CLIP-ViT-B/16	76.34	93.14	72.31	75.27	**76.73**
		CLIP-ViT-B/32	71.94	93.09	67.18	64.57	
Prediction Trustworthiness (PT, %) ↑	ZS	CLIP-ViT-B/16	87.05	62.99	92.96	29.38	**70.05**
		CLIP-ViT-B/32	89.39	73.44	94.58	30.57	
	CoOp	CLIP-ViT-B/16	75.35	35.94	83.25	2.35	54.43
		CLIP-ViT-B/32	79.59	47.96	90.30	20.70	
	TA	CLIP-ViT-B/16	66.76	37.90	85.95	6.14	52.24
		CLIP-ViT-B/32	72.77	50.25	89.19	8.98	
Inference Reliability (IR, %) ↑	ZS	CLIP-ViT-B/16	66.33	85.58	61.96	68.05	67.65
		CLIP-ViT-B/32	61.09	85.23	56.12	56.80	
	CoOp	CLIP-ViT-B/16	78.50	96.13	75.41	69.93	78.84
		CLIP-ViT-B/32	73.90	92.82	70.99	73.00	
	TA	CLIP-ViT-B/16	78.42	96.06	74.39	86.73	**81.00**
		CLIP-ViT-B/32	74.60	94.50	69.09	74.17	

Table 4: Comparisons of three methods regarding prediction accuracy and our proposed PT and IR scores. Tip-Adapter is abbreviated as "TA" due to the space limit in this table. The best averaged scores are marked as **bold**. In this case, these fine-tuning methods (CoOp, Tip-Adapter) all tend to outperform ZS in prediction accuracy and IR. However, ZS achieves the best averaged performances with PT evaluations.

Methods	VLMs	Datasets			
		IN	CT	SD	CUB
FLCP	ALBEF-ViT-B/16	1e-5	3e-5	3e-5	3e-5
	BLIP-ViT-B/16	1e-5	3e-5	3e-5	1e-5
	CLIP-ViT-B/16	1e-6	3e-6	1e-5	1e-5
	CLIP-ViT-B/32	5e-6	2e-5	1e-5	1e-5
FT	ALBEF-ViT-B/16	3e-5	1e-4	3e-5	3e-5
	BLIP-ViT-B/16	3e-5	1e-4	3e-5	2e-5
	CLIP-ViT-B/16	1e-5	3e-5	1e-5	1e-5
	CLIP-ViT-B/32	1e-5	1e-5	1e-5	2e-5

Table 5: Learning rate settings when fine-tuning with FLCP and FT methods.

the FLYP method, the temperature value of the contrastive loss is set to 0.07. The learning rate of the LP method is set as $1e-3$. The learning rate settings of FLYP and FT methods as shown in Table 5. The reason why learning rates behave in variations under different training scenarios is that we find the appropriate learning rate varies in different situations. For example, when fine-tuning on ImageNet-1K, the learning rate with $1e-3$ is too big for FLYP and FT methods to let models converge during training. We provide the results from a single run. We did not observe any variation in results with the same setup.

Experiments with More Fine-tune Techniques

Recently, there have been other fine-tuning techniques proposed by the community. Typical ones include prompt tuning such as CoOp (Zhou et al. 2022), adapter tuning such as Tip-Adapter (Zhang et al. 2022). Here we experiment with these three techniques (CoOp, Tip-Adapter) with CLIP-ViT-B/16 and CLIP-ViT-B/32 models. The high-level concepts of these methods are introduced here:

1. **CoOp** uses learnable vectors to model the words in the prompt, while keeping the parameters of the pre-trained VLM fixed throughout the process. It considers two types of learnable prompts: the first is a unified context, where the learnable context is the same regardless of the sample's category; the second is a class-specific context, where each category has its own unique learnable context.

2. **Tip-Adapter** leverages VLM such as CLIP to construct a cache model, which stores classification knowledge from downstream training data. Based on this approach, Tip-Adapter-F turns the Keys part of the Cache Model into learnable parameters, allowing them to be updated through training. Here we employ the Tip-Adapter-F method in our study.

As shown in Table 4, we find our findings remain consistent. While these fine-tuning strategies (CoOp, Tip-Adapter) tend to enhance the inference reliability (IR) of VLMs, they often deteriorate the prediction trustworthiness (PT).

Figure 7: Experimental results on out-of-distribution data with CLIP-ViT-B/16 model. Due to the space limit in this figure, "JPEG Compression" and "Speckle Noise" are abbreviated as "JPEG Comp." and "Speckle Noi.", respectively.