Towards Calibrated Robust Fine-Tuning of Vision-Language Models

Changdae Oh[*] University of Seoul

Mijoo Kim[*] Chung-Ang University

Hyesu Lim[*] KAIST

Junhyeok Park[*] KAIST

Euiseog Jeong[*] Sungkyunkwan University

Zhi-Qi Cheng[†] CMU

Kyungwoo Song[†] Yonsei University

Abstract

While fine-tuning unlocks the potential of a pre-trained model for a specific task, it compromises the model's ability to generalize to out-of-distribution (OOD) datasets. To mitigate this, robust fine-tuning aims to ensure performance on OOD datasets as well as on an in-distribution (ID) dataset for which the model is being tuned. However, another criterion for reliable machine learning (ML) – *confidence calibration* has been overlooked despite its increasing demand for real-world high-stakes ML applications (e.g. autonomous driving and medical diagnosis). For the first time, we raise concerns about the calibration of fine-tuned vision-language models (VLMs) under distribution shift by showing that naive fine-tuning and even state-of-the-art robust fine-tuning methods hurt the calibration of pre-trained VLMs, especially on OOD datasets. To address this issue, we provide a simple approach, called calibrated robust fine-tuning (**CaRot**), that incentivizes calibration and robustness on both ID and OOD datasets. Empirical results on ImageNet-1K distribution shift evaluation verify the effectiveness of our method.

1 Introduction

Recently, large-scale vision-language pre-training has ushered in a new era of foundation models [1, 2, 3, 4] in computer vision. Due to their unprecedented generalization capability and well-aligned multimodal embeddings, vision-language models (VLMs) such as CLIP [2] are used in numerous downstream tasks such as zero-shot classification [2, 5], zero-shot segmentation [6, 7], and cross-modal generation [8, 3]. However, despite their promising results, the zero-shot performance still lags behind the fine-tuned performance on a specific downstream task [2, 9, 10].

To this end, there have been numerous lines of work on fine-tuning VLMs [9, 11, 12, 13]. Among them, robust fine-tuning, which aims to achieve good performance on both out-of-distribution (OOD) and in-distribution (ID) data, has attracted much attention. After Wortsman et al. [9] and Kumar et al. [10] started the discussion on the robustness of recent VLMs under distribution shifts, a wide range of research has followed [14, 15, 16]. However, there is a lack of ongoing studies investigating *confidence calibration*, which is a crucial factor in reliable machine learning (ML). With the increasing development of ML in real-world decision-making systems, the calibration (matching between the actual correctness and the confidence of their prediction) of neural networks has been actively studied to achieve reliable ML. It is crucial to avoid making incorrect predictions with high confidence, especially in high-stakes tasks such as autonomous driving and healthcare applications. After a seminal work [17] revealed the miscalibration problem of accurate neural

[*] Work done while they are at Carnegie Mellon University (CMU) as visiting students.

[†] Co-corresponding author.

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

networks, a plethora of attempts followed to improve the calibration of neural network models through post-hoc adjustments [18, 17, 19, 20, 21] or train-time regularizations [22, 23, 24, 25, 26].

Recently, Minderer et al. [27] systematically reviewed the calibration of modern vision models across different model architectures, model sizes, and amounts of pre-training, and LeVine et al. [28] raised concerns about the calibration of a zero-shot VLM (e.g., CLIP). However, to the best of our knowledge, no existing work addresses the calibration of *fine-tuned* VLMs, which is timely to be questioned. In this work, we initiate the investigation on the calibration of VLMs after fine-tuning under distribution shifts. We observe that the traditional cross-entropy-based fine-tuning severely hurts the calibration in terms of expected calibration error (ECE) [29](discussed in Section 2.1) on OOD datasets as well as on the ID dataset on which the model is fine-tuned. Furthermore, while a state-of-the-art (SOTA) robust fine-tuning method achieves strong prediction accuracy on OOD datasets and somewhat convincing ECE on ID, it still struggles with high ECE on OOD datasets. We argue that current robust fine-tuning methods urgently need more calibration on both ID and OOD settings, based on our salient observations. In this paper, we introduce simple yet effective solutions, label smoothing, and its data-dependent variant, to cope with the miscalibration during fine-tuning.

In the experiments on the ImageNet-1K distribution shift benchmark, we show that simply employing label smoothing [30] with contrastive loss brings remarkable improvement in ID and OOD calibration. Moreover, using multimodal knowledge distillation as a form of data-dependent label smoothing further boosts the performance in terms of ID/OOD calibration and ID/OOD generalization.

Contributions. 1) This is the first paper to discuss the calibration of a fine-tuned VLM under a distribution shift scenario: We show that the standard fine-tuning largely harms the calibration of pre-trained VLMs on both ID and OOD datasets, and that a SOTA fine-tuning method also induces unsatisfactory calibration results on OOD datasets; 2) We derive a simple yet intriguing argument that *adopting label smoothing not only increases the generalization capability of the VLM, but also makes it better calibrated on both ID and OOD datasets*; 3) By equipping the multimodal knowledge distillation term as data-dependent label smoothing, we minimize the drop in calibration compared to the pre-trained model, while achieving the highest accuracy on both ID and OOD datasets.

2 Calibration of Vision-Language Models in the Wild

2.1 Preliminaries

In a K class classification, let $X \in \mathbb{R}^d$ and $Y \in \{1, ..., K\}$ be random variables indicating inputs and labels, respectively. A data set with N independent samples from the joint distribution $\pi(X, Y) = \pi(Y|X)\pi(X)$ is denoted by $\mathcal{D} = \{(X_n, Y_n)\}_{n=1}^N$. Let f be a classifier and $f(X) = \hat{Z}$ be a logit. We can derive the confidence score $\hat{P}$ by taking the maximum among the probabilistic masses $\texttt{softmax}(\hat{Z})$. If the confidence score $\hat{P}$ matches the true probability p, we say the model is *calibrated* and it can be written as $\mathbb{P}(\hat{Y} = Y|\hat{P} = p) = p,\ \forall p \in [0, 1]$, where $\hat{Y}$ denotes the predicted class. With a simple modification, the calibration error can be derived as $\mathbb{E}_{\hat{P}}[|\mathbb{P}(\hat{Y} = Y|\hat{P} = p) - p|]$. However, since the data set size N is finite in practice, the calibration error should be approximated empirically. In this context, Naeini et al. [29] proposed the expected calibration error (ECE), one of the most representative metrics for uncertainty calibration. It divides the predictions on each example into M uniform confidence bins $\{\text{bin}_m\}_{m=1}^M$ and takes a weighted average of the gap between accuracy and confidence of each bin, i.e. $\text{ECE} = \sum_{m=1}^M \frac{|\text{bin}_m|}{N} |\text{acc}(\text{bin}_m) - \text{conf}(\text{bin}_m)|$. Meanwhile, as a popular post-hoc adjustment-based calibration method, temperature scaling [17] divides the logit by temperature T before applying the softmax function. By scaling the logit, the confidence distribution of each class becomes more uniform or sharper to mitigate over/under confidence problems.

2.2 Miscalibration of fine-tuned VLMs

As a quick validation, we visualize the calibration of CLIP ViT-B/16 through reliability diagrams [31] on ImageNet-1K (ID) and ImageNet-R (OOD) after fine-tuning it on ImageNet-1K. As shown in Figures 1 and 2, (1) the standard fine-tuning significantly hurts the calibration compared to the pre-trained VLM, especially on the OOD dataset, (2) while the SOTA robust fine-tuning method achieves relatively better ID calibration, it also suffers to achieve good calibration on the OOD dataset. These miscalibration issues of fine-tuned VLMs fuel us to devise a new fine-tuning method

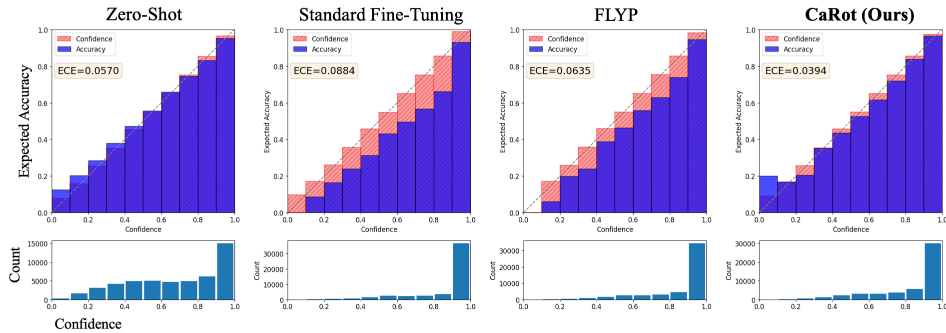

Figure 1: Reliability diagram [31] and confidence histogram of zero-shot CLIP and fine-tuning methods on ImageNet-1K. Each fine-tuning method is evaluated after fine-tuning on ImageNet-1K.

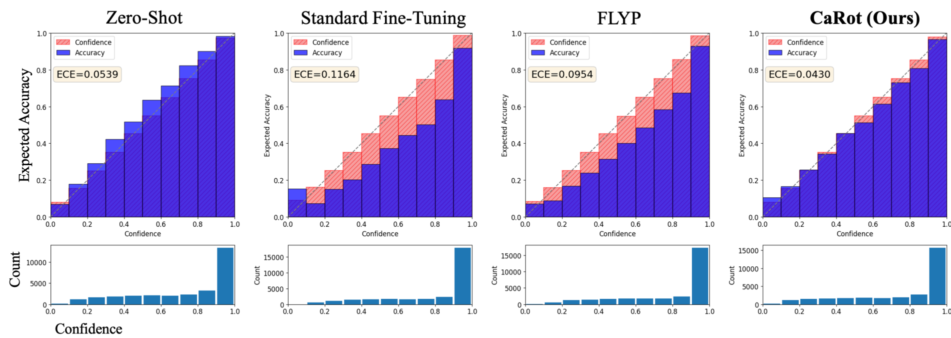

Figure 2: Reliability diagram [31] and confidence histogram of zero-shot CLIP and fine-tuning methods on ImageNet-R. Each fine-tuning method is evaluated after fine-tuning on ImageNet-1K.

accomplishing better calibration (rightmost panels of Figure 1 and 2). Please refer to Supplementary for additional results.

2.3 Towards calibrated robust fine-tuning of VLMs

Recently, Goyal et al. [15] proposed to fine-tune VLMs via contrastive loss, which is the same loss used for pretraining, and called the proposed approach as FLYP (finetune like you pretrain). By simply exploiting the pretraining objective for fine-tuning, FLYP showed robust performance both in ID and OOD settings without any additional techniques such as ensembling. Furthermore, we discovered that calibration on ID setting is much better when using FLYP than standard fine-tuning with cross-entropy loss (Section 3). Drawing inspiration from FLYP, we utilize contrastive loss as our fine-tuning objective. Let our vision-language model parameterized with $\theta = \{\theta_v, \theta_l\}$, which has image encoder $f_{\theta_v}(\cdot)$ and text encoder $g_{\theta_l}(\cdot)$. Given a downstream data minibatch $\mathcal{B} = \{(I_1, T_1), ..., (I_B, T_B)\}$ size of B, multimodal contrastive loss $\mathcal{L}_{\text{MCL}}$ can be written as:

$$\mathcal{L}_{\text{MCL}}(\mathcal{B}, \theta) := \sum_{i=1}^{B} - \log \frac{\exp(f_{\theta_v}(I_i) \cdot g_{\theta_l}(T_i))}{\sum_{j=1}^{B} \exp(f_{\theta_v}(I_i) \cdot g_{\theta_l}(T_j))} + \sum_{i=1}^{B} - \log \frac{\exp(f_{\theta_v}(I_i) \cdot g_{\theta_l}(T_i))}{\sum_{j=1}^{B} \exp(f_{\theta_v}(I_j) \cdot g_{\theta_l}(T_i))}. \quad (1)$$

Label smoothing. As one of the representative regularization strategies, label smoothing (LS) [30] pursues the generalization of a classification model by addressing the issue of overconfident predictions. To do so, it derives a new target probability distribution by adjusting the ground truth one-hot vector, i.e., reducing the value of 1 to the target class, and increasing the values of 0 on the other classes to the amount of smoothing. Mathematically, LS can be represented as $y^{LS} = (1-\epsilon)y + \epsilon\hat{y}$, where y is a one-hot vector of a target label, $\epsilon \in (0, 1)$ is a smoothing parameter, and $\hat{y} = \frac{1}{K}$ is the uniform distribution over the label space. Beyond the generalization, it is also shown that LS is beneficial for calibration [25], which motivates us to adopt it for VLM fine-tuning.

Knowledge distillation as a data-dependent label smoothing. While the standard form of LS always produces a fixed target regardless of the input data, one may want to alternatively design

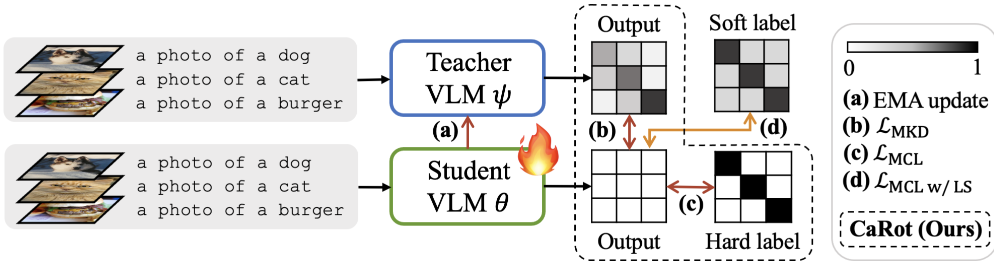

Figure 3: **Method overview.** We update the teacher VLM ψ via **(a)** EMA update. Our proposed CaRot updates the student VLM θ using **(b)** multimodal knowledge distillation $\mathcal{L}_{\text{MKD}}$ (Eq. 2) and **(c)** multimodal contrastive $\mathcal{L}_{\text{MCL}}$ (Eq. 1) losses as fine-tuning objective. Darker and lighter elements denote values closer to 1 and 0, respectively. The hard and soft labels denote one-hot encoded and label-smoothed target labels, respectively. While **(d)** $\mathcal{L}_{\text{MCL w/ LS}}$ gives the same target labels for every training instance, $\mathcal{L}_{\text{MKD}}$ produces data-dependent dynamic target labels.

smoothed labels in a data-dependent manner. Recently, [32, 33] revealed the connection between knowledge distillation and LS for computer vision models. This motivates us to investigate the feasibility of utilizing the knowledge distillation strategy as a data-dependent LS to achieve calibration of fine-tuned VLM. For this, we first set an exponentially moving average (EMA) of learning parameter θ as a teacher network that has $\psi = \{\psi_v, \psi_l\}$, which can be written as $\psi \leftarrow \alpha\psi + (1-\alpha)\theta$, where α controls the evolution speed of the teacher model. For stability, we update the EMA teacher model ψ per a certain number of iterations, which we set as 500 instead of updating every step (see Table 2 for detail). In this work, we employ a self-evolving EMA network as a teacher model rather than hosting another VLM, in accordance with the emerging evidence of the strong generalization capability of weight space ensembling approaches [34, 9, 35]. By interpolating the weight of a pre-trained VLM with that of a dynamically fine-tuned one, we can view our teacher model as an ensemble of *multi-domain calibrated predictor* (pre-trained VLM) and *in-domain calibrated predictor* (fine-tuning VLM), which is an analogy to the recently proposed ensemble approach *in-distribution calibrated ensembles* [36]. With the EMA teacher ψ and the learning student θ, we construct a multimodal knowledge distillation loss term for a minibatch $\mathcal{B}$ as:

$$\mathcal{L}_{\text{MKD}}(\mathcal{B}, \theta) := \sum_{i=1}^{B} [KL(\tilde{q}_i^I || q_i^I))) + KL(\tilde{q}_i^T || q_i^T)))], \tag{2}$$

where $\tilde{q}_i^I = \texttt{softmax}(\{f_{\psi_v}(I_i) \cdot g_{\psi_l}(T_j)\}_{j=1}^B)$ and $q_i^I = \texttt{softmax}(\{f_{\theta_v}(I_i) \cdot g_{\theta_l}(T_j)\}_{j=1}^B)$. Here, $\tilde{q}_j^T$ and q_j^T are similarly defined by exchanging the index i with j. As depicted in Figure 3, we finalize our learning objective as a sum of $\mathcal{L}_{\text{MCL}}$ and $\mathcal{L}_{\text{MKD}}$ with hyperparameter λ that controls the magnitude of distillation, i.e., $\mathcal{L} = \mathcal{L}_{\text{MCL}} + \lambda\mathcal{L}_{\text{MKD}}$. While this induces the same objective as Cheng et al. [37], we propose this from a fine-tuning context with consideration of robustness and calibration, which is different from data-efficient supervision during pre-training in terms of motivation and aim.

3 Experiments

Setup. We adopt CLIP ViT-B/16 as our VLM backbone and validate its calibration and robustness under distribution shift. We choose the ImageNet-1K classification under the distribution shift scenario as our testbed. Following previous robust fine-tuning literature [9, 10, 15], we fine-tune CLIP on ImageNet-1K [38], and then evaluate the ID accuracy and OOD accuracy. Moreover, to investigate the calibration of fine-tuned VLMs, we compute ECE [29] on ID and OOD settings. Here, the number of confidence bins as set to 10 following the previous work [28]. As baselines, we consider zero-shot inference with pre-trained CLIP (ZS) with the following adaptation methods: standard fine-tuning (FT), WiSE-FT [9], FLYP [15], and our proposal. For each method, we additionally conduct the post-hoc calibration strategy temperature scaling (TS). More details are in the Supplementary.

Main results. Table 1 shows that: 1) During adaptation on the ID dataset, FT somewhat sacrifices the OOD generalization capacity of ZS as well as ID/OOD calibration; 2) While WiSE-FT finds a better trade-off between ID Acc. and OOD Acc., it significantly degrades the calibration of a

Table 1: Fine-tuning of CLIP-ViT-B/16 on ImageNet-1K and evaluation on its ID testset and five OOD variants (ImageNet-V2 [39], ImageNet-R [40], ImageNet-A [41], ImageNet-Sketch [42], and ObjectNet [43]). Here, OOD performances are averaged over the five variants. We denote label smoothing as LS, temperature scaling as TS, zero-shot as ZS, and fine-tuned as FT.

Method	ID Acc. (↑)	OOD Acc. (↑)	w/o *TS* ID ECE (↓)	w/o *TS* OOD ECE (↓)	w/ *TS* ID ECE (↓)	w/ *TS* OOD ECE (↓)
ZS	0.6832	0.5840	0.0571	0.0836	0.0561	0.0748
FT	0.8153	0.5750	0.0884	0.2186	0.0629	0.1629
FT w/ LS	0.8223	0.5833	0.0460	0.1147	0.0481	0.1282
WiSE-FT	0.8043	0.6350	0.2129	0.1764	0.0872	0.1533
WiSE-FT w/ LS	0.8068	**0.6405**	0.5231	0.3601	0.3382	0.2425
FLYP	0.8258	0.5946	0.0643	0.1831	0.0392	0.1217
FLYP w/ LS	0.8271	0.5975	0.0459	0.1295	0.0427	0.1145
CaRot	**0.8319**	0.6197	**0.0395**	**0.1093**	**0.0380**	**0.0980**

pre-trained model; 3) FLYP achieves strong generalization on ID and OOD, and ID calibration, but lacks the OOD calibration; 4) TS helps calibration somewhat, but the gap between ZS OOD and fine-tuned ones still non-negligible; 5) LS remarkably improves the calibration on both ID and OOD datasets; 6) CaRot gets superior results overall metrics ID/OOD generalization and calibration which verify the effectiveness of data-dependent label smoothing coupled with contrastive loss.

Ablation study. This section provides an ablation study on the frequency of EMA updates for the teacher model as we discussed in Section 2.3. In Table 2, we see that updating the teacher model for every single iteration does not bring meaningful improvement on both ID/OOD accuracy and ECE compared to FLYP (refer Table 1). While a slowly updated teacher model (i.e., single update per 1000 iterations) induces a slightly lower ID accuracy and higher ECE and our default update frequency (i.e., 500), it largely improves OOD accuracy and ECE, implying that the blending frequency (and ratio) between a pre-trained checkpoint and fine-tuning ones make a huge influence on the trade-off between ID/OOD calibration as well as generalization.

Table 2: Ablation on EMA teacher update frequency. ✗ denotes pre-trained weights without updates.

EMA update freq.	ID Acc. (↑)	OOD Acc. (↑)	ID ECE (↓)	OOD ECE (↓)
1	0.8261	0.5956	0.0630	0.1824
500 (default)	**0.8319**	0.6197	**0.0395**	0.1093
1000	0.8303	**0.6266**	0.0511	**0.0781**
✗	0.7848	0.6199	0.1257	0.0936

4 Conclusion

In this paper, we pave the way for robust fine-tuning of visual foundation models with consideration of *confidence calibration*. Specifically, we reveal that naive fine-tuning and even SOTA robust fine-tuning methods struggle with achieving satisfactory confidence calibration on both ID and OOD datasets for the first time. By adopting a simple regularization technique (i.e., label smoothing), we show that the calibration issue on ID and OOD datasets is easily addressed. From these findings, we further propose the utilization of multimodal knowledge distillation as a form of data-dependent label smoothing (so-called CaRot), resulting in promising results in terms of generalization and calibration.

Acknowledgment

We appreciate Professor Alex Hauptmann at CMU and Professor Jaegul Choo at KAIST for giving us insightful feedback about the overall research design and method. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2021R1F1A1060117).

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[5] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15691–15701, October 2023.

[6] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.

[7] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, June 2023.

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[9] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[12] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

[13] Changdae Oh, Junhyuk So, Yongtaek Lim, Hoyoon Byun, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. *NeurIPS*, 2023.

[14] Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[16] Junjiao Tian, Zecheng He, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, and Zsolt Kira. Trainable projected gradient method for robust fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7836–7845, June 2023.

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[18] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

[19] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[20] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning (ICML)*, 2020.

[21] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. *arXiv preprint arXiv:2006.12800*, 2020.

[22] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[23] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[24] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[25] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[26] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[27] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

[28] Will LeVine, Benjamin Pikus, Pranav Raj, and Fernando Amat Gil. Enabling calibration in the zero-shot inference of large vision-language models. *arXiv preprint arXiv:2303.12748*, 2023.

[29] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery.

[32] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[33] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996, 2021.

[34] P Izmailov, AG Wilson, D Podoprikhin, D Vetrov, and T Garipov. Averaging weights leads to wider optima and better generalization. In *Uncertainty in Artificial Intelligence (UAI)*, 2018.

[35] Alexandre Ram'e, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning (ICML)*, 2023.

[36] Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *Uncertainty in Artificial Intelligence (UAI)*, 2022.

[37] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E. Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3119–3124, June 2021.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

[40] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[41] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[42] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[43] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[45] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

A Supplementary Material

A.1 Implementation details

To conduct ZS, FT, WiSE-FT, and FLYP, we follow the official protocol of previous works [9, 15]. We use the default configuration, hyperparameter, and prompt template proposed by the authors [15]. For all adaptation methods (FT and FLYP), we train the model using AdamW [44] with a learning rate 1e-5 (for FLYP and CaRot) and 3e-5 (for FT) and weight decay 0.1 during 10 epochs with batch size 512, resulting in about 25K training iterations. The ensemble coefficient of WiSE-FT is simply set as 0.5, which works well, as shown in the original paper. The linear classification head of FT was initialized with the text representation vectors for each class from pre-trained CLIP for stable fine-tuning.

We select the temperature scaling parameter based on ID validation set ECE over range $[0.5, 10.0]$, and label smoothing parameter over $\{0.01, 0.03, 0.05, 0.1, 0.2\}$. For CaRot, the distillation loss coefficient lambda and EMA momentum parameter are set to 0.75 and 0.9, respectively. When performing the EMA update, we linearly increased the momentum coefficient from a small value (0.05) to a final target value (0.9) during the first 20% of iterations and kept it constant for the remaining iterations by following [45].

A.2 Additional results

We provide reliability diagrams and confidence histograms of four methods (ZS, FT, FLYP, and CaRot) on ImageNet-1K (ID) and its four variants (-V2, -A, -R, -SKetch) without (Fig. 4) and with (Fig. 5) temperature scaling. While standard fine-tuning and state-of-the-art method (FLYP) significantly hurt the calibration of a pre-trained visual foundation model, especially on OOD datasets (second to fifth rows), our proposed CaRot robustly protects or even improves (second and fourth rows in Fig. 5, and the fourth row in Fig. 4) the calibration of a pre-trained model.

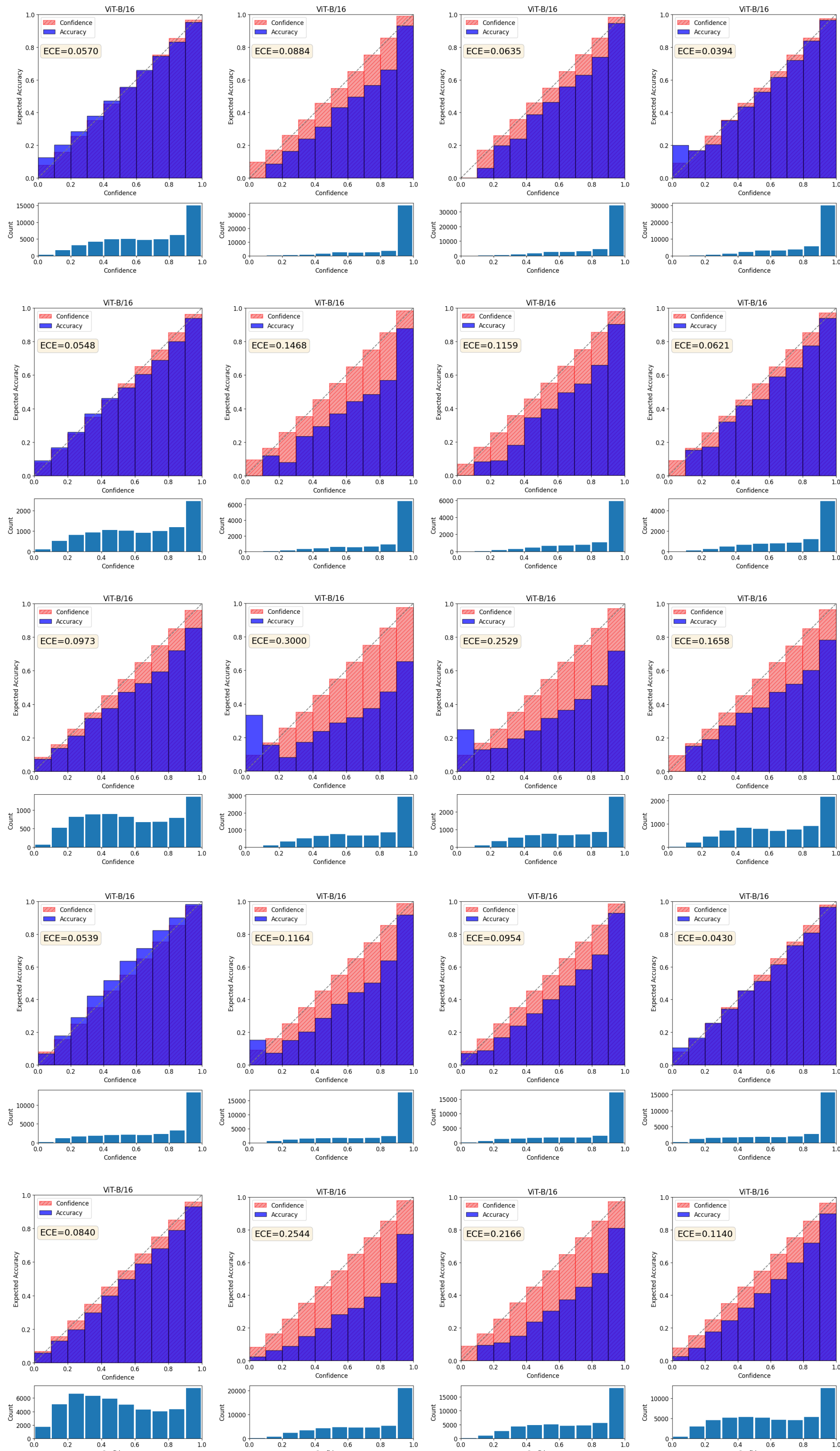

Figure 4: Reliability diagram [31] and confidence histogram of zero-shot CLIP and fine-tuning methods without temperature scaling. From top to bottom rows, ImageNet, -V2, -A, -R, and -Sketch were reported. From left to right columns, ZS, FT, FLYP, and CaRot were reported.

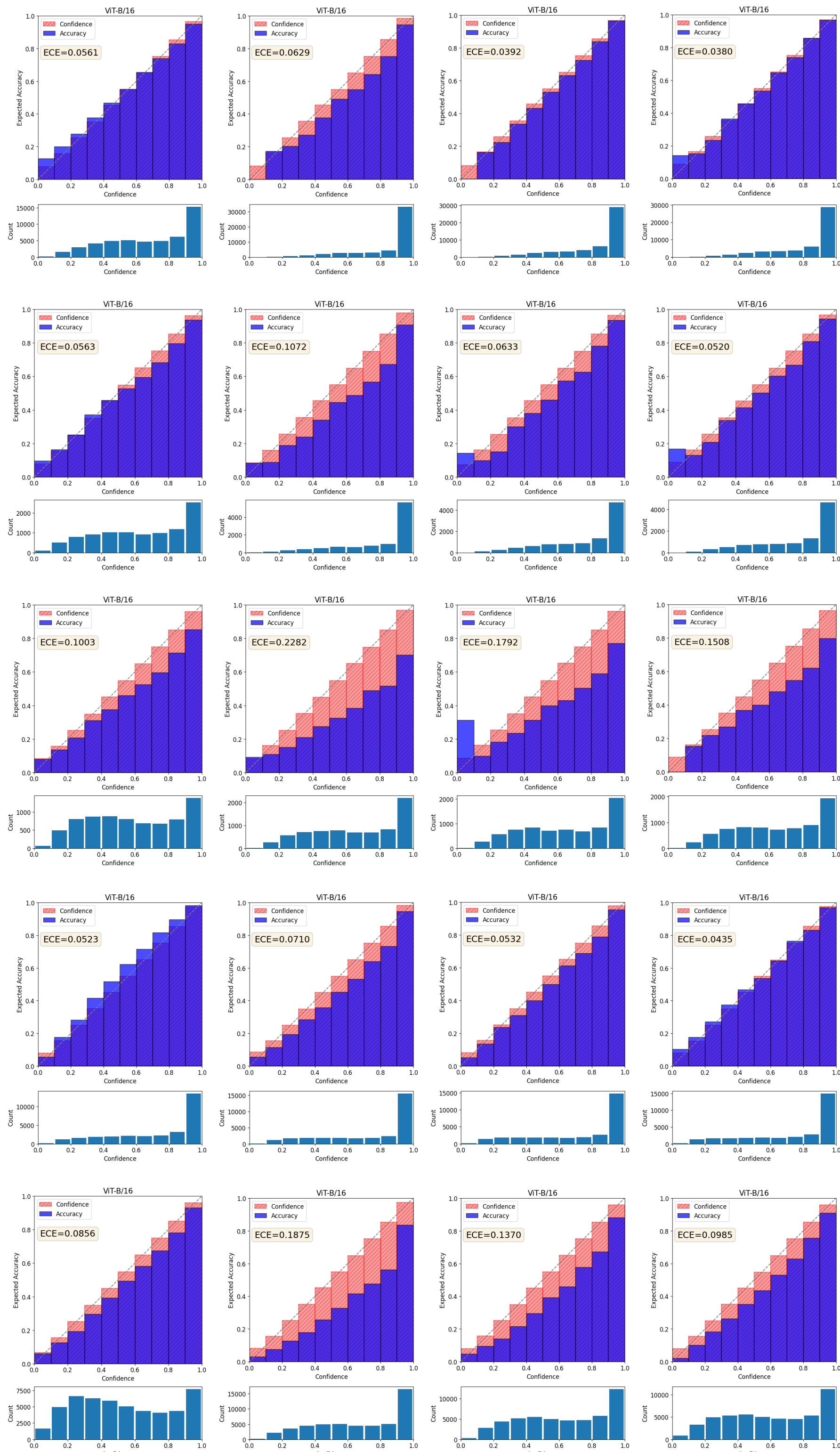

Figure 5: Reliability diagram [31] and confidence histogram of zero-shot CLIP and fine-tuning methods after applying temperature scaling. From top to bottom rows, ImageNet, -V2, -A, -R, and -Sketch were reported. From left to right columns, ZS, FT, FLYP, and CaRot were reported.