Robust Calibration of Large Vision-Language Adapters

Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz

ETS Montreal, Canada
balamurali.murugesan.1@ens.etsmtl.ca

Abstract. This paper addresses the critical issue of miscalibration in CLIP-based model adaptation, particularly in the challenging scenario of out-of-distribution (OOD) samples, which has been overlooked in the existing literature on CLIP adaptation. We empirically demonstrate that popular CLIP adaptation approaches, such as Adapters, Prompt Learning, and Test-Time Adaptation, substantially degrade the calibration capabilities of the zero-shot baseline in the presence of distributional drift. We identify the increase in logit ranges as the underlying cause of miscalibration of CLIP adaptation methods, contrasting with previous work on calibrating fully-supervised models. Motivated by these observations, we present a simple and model-agnostic solution to mitigate miscalibration, by scaling the logit range of each sample to its zero-shot prediction logits. We explore three different alternatives to achieve this, which can be either integrated during adaptation or directly used at inference time. Comprehensive experiments on popular OOD classification benchmarks demonstrate the effectiveness of the proposed approaches in mitigating miscalibration while maintaining discriminative performance, whose improvements are consistent across the three families of these increasingly popular approaches. The code is publicly available at: https://github.com/Bala93/CLIPCalib .

Keywords: Vision-language models · Few-shot adaptation · Domain generalization · Test-time adaptation · Network calibration

1 Introduction

Deep learning is undergoing a paradigm shift with the emergence of pre-training large-scale language-vision models, such as CLIP [45]. These models, and more particularly the variants integrating vision transformers, have demonstrated impressive generalization capabilities in visual recognition tasks, yielding exceptional zero-shot and few-shot performance. Nevertheless, in a dynamic and evolving open world, machine learning applications inevitably encounter the challenge of out-of-distribution (OOD) data, which typically hinders the scalability of these models to new domains. Existing literature based on CLIP faces this scenario

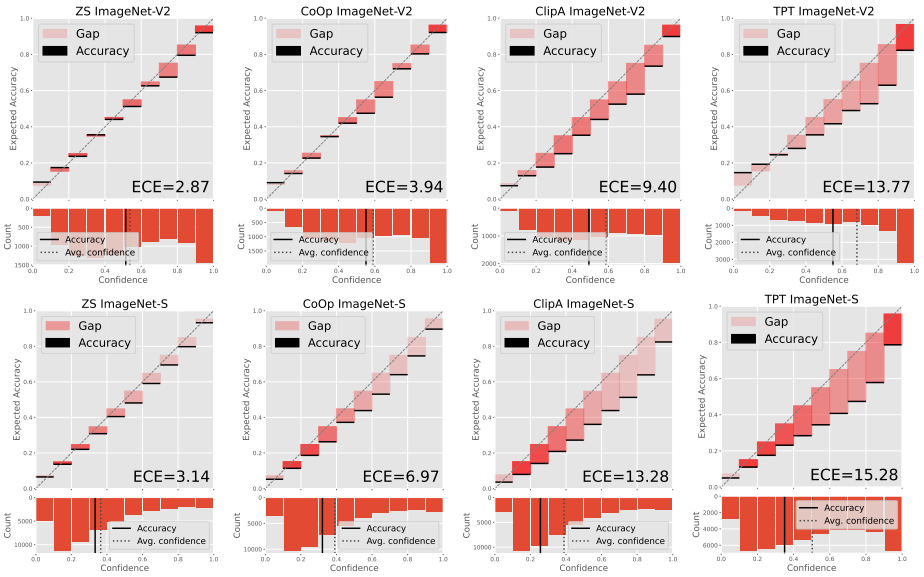

Fig. 1: CLIP-based adaptation methods are severely miscalibrated on Out-of-distribution (OOD) samples. Three families of popular approaches to adapt CLIP under different scenarios, i.e., Prompt Learning (CoOp [66]), Adapters (Clip-Ad [10]) and Test-time prompt tuning (TPT [49]), significantly degrade the miscalibration of the zero-shot baseline, despite improving its discriminative performance.

with different solutions to exacerbate robustness. In particular, freezing the entire vision backbone to re-use these generalizable features has been a popular choice, especially in the low-data regime [10, 66]. Thus, CLIP adaptation during training typically resorts to Adapters [10, 63] or Prompt Learning [65, 66] strategies, which leverage a few labeled samples to adapt the model with the hope that it will generalize properly to unseen related-domains. Furthermore, a more challenging scenario consists of adapting the model during inference without any access to labeled data, where a prevalent method is Test-Time Prompt Tuning (TPT) [49].

While these strategies have further improved the discriminative performance of a zero-shot baseline, we have observed that the accuracy of the uncertainty estimates of the predictions, i.e., calibration, is significantly degraded (see Fig. 1), regardless of the family of adaptation models or setting. Thus, after adaptation, model predictions are often over-confident, even if they are wrong. This represents a major concern, as inaccurate uncertainty estimates can carry serious implications in safety-critical applications, such as healthcare, where CLIP is emerging as a popular strategy [27, 31]. Nevertheless, despite its importance, the miscalibration issue has been overlooked in the CLIP adaptation literature.

Motivated by these observations, this paper addresses this critical issue, which has been disregarded in current literature. Indeed, few-shot adaptation

strategies, notably Prompt Learning and Adapters, are attracting wide attention recently, with an unprecedented surge in the number of methods proposed [12, 18, 22, 31, 50, 60, 63, 65, 66], albeit being a relatively recent research topic. Nevertheless, the main focus of this growing literature has been on improving the discriminative power of adapted models. Thus, given their increasing popularity, and quick adoption in real-world safety-critical problems, we believe that assessing the calibration performance of CLIP adaptation strategies in OOD scenarios is of paramount importance to deploy not only high-performing but also reliable models. We can summarize our contributions as follows:

1. We empirically demonstrate that popular CLIP adaptation strategies, such as Adapters, Prompt Learning, and Test-Time Prompt Tuning, substantially degrade the calibration capabilities of the zero-shot baseline in the presence of distributional drift.
2. For these adaptation strategies, we expose that the underlying cause of miscalibration is, in fact, the increase of the logit ranges. This contrasts with recent work in calibrating fully-supervised models [56], which suggests that the inherent cause of miscalibration is the increase of its norm instead, due to the standard cross-entropy loss used for training.
3. Based on these observations, we present a simple, and model-agnostic solution, which consists in scaling the logit range of each sample based on the zero-shot logits. We further present several alternatives to accommodate our solution, which can be implemented either at training or inference time.
4. Comprehensive experiments on popular OOD classification benchmarks empirically demonstrate the effectiveness of our approaches to reduce the miscalibration error, while keeping the discriminative performance.

2 Related Work

2.1 Vision language models

Text-driven pre-training of image representation, so-called vision-language models (VLMs) is revolutionizing the paradigm of transfer learning. These models can integrate massive web-scrabbled data sources thus learning robust feature representations. In particular, models such as CLIP [45] or ALIGN [19] train joint multi-modal embedding spaces via contrastive learning of paired images and text, using dual encoder architectures. Such strong vision-language alignment has demonstrated robust open-vocabulary zero-shot generalization capabilities [45, 61]. Given such potential, transferring pre-trained VLMs to a wide variety of tasks is gaining increasing popularity. Nevertheless, this process faces particular challenges. First, large-scale pre-training usually involves also scaling network sizes, which is a computational bottleneck for low-resource adaptation scenarios. Second, recent attempts to fine-tune VLMs have demonstrated a deterioration of their robustness against domain drifts [24, 57], especially when available data is limited. Thus, an emerging core of recent literature is focusing on novel alternatives to overcome these limitations. More concretely, freezing the pre-trained backbone, and reusing such features by training

a small set of parameters, via Prompt Learning [12, 22, 65–67], or black-box Adapters [10, 26, 39, 50, 60, 63, 64], is getting increasing attention.

2.2 Prompt based learning

CLIP models have shown encouraging results by hand-crafting personalized text descriptions of the target visual representation [34]. The automatizing of this cumbersome process raises the concept of Prompt Learning (PL) [66], a family of methods to adapt CLIP that inserts a set of continuous learnable tokens in the original text prompt at the input of the VLM language encoder. While the CLIP model remains frozen, PL optimizes the most discriminative text input, given a few-shot support set [22, 65–67]. CoOP [66] represents one of the initial attempts to study the effect of prompt tuning on different tasks, and proposed to learn the prompt's context words. CoCoOP [65], on the other hand, designed a simple network to predict the input text prompt through image features, as CoOP failed to match the zero-shot performance on generic tasks. TPT [49] extends PL to address time-test adaptation scenarios by updating the prompt for a batch with original and augmented samples through entropy minimization.

2.3 Black-box Adapters

Prompt Learning involves using the CLIP's encoder throughout the entire training process as the backpropagation of the gradient has to pass through it to update the prompts, which results in large computational constraints [10]. Adapter-based techniques provide an alternative to Prompt Learning for aligning to downstream tasks, leveraging uniquely pre-computed features with minimal additional parameters. A base version of such methods involves training a linear classifier via logistic regression, typically referred to as Linear Probing [45]. Nevertheless, leveraging only the vision features does not fully exploit the potential of VLMs. To this end, several methods have proposed enhanced Adapters, which further rely on zero-shot text-driven class-wise prototypes. In particular, Clip-Adapter [10] introduced additional fully connected layers and operated on the vision or language branch through residual style feature combination. Training-free methods such as Tip-Adapter [63] resorted to a key-value cache model based on the available few-shot supports. Likewise, TaskRes [60] introduced additional learning parameters and applied a residual modification of the text representation, which led to a better initialization when learning from few-shot supervision. More recently, [50] provided a wider look at the coupling of vision and text features in such Adapters, by pointing out that these methods largely build up their improved performance on initializing the logistic classifier weights with the zero-shot prototypes, proposing a simple solution, coined CLAP, for a better distillation of such prototypes.

2.4 Model calibration

Calibrating the confidence of deep learning models is paramount in developing reliable solutions, as the confidence is expected to correlate with correctness.

Given the importance and the potential impact of miscalibration, a growing literature to address this issue has emerged in the last years. Post-processing techniques have been widely used to achieve calibration, wherein a linear transformation [11, 20, 54] is applied to the predicted logits before converting it to softmax. Nevertheless, an important limitation is that these transformations are obtained using held-out validation data, which is assumed to follow the same distribution as the test data, limiting their applicability in the presence of domain drifts [40]. A popular alternative consists in calibrating the networks at training time. This can be achieved by incorporating explicit penalties that either penalize overconfident softmax predictions [4, 25, 41, 43] or encourage small logit differences [29, 30, 37]. Furthermore, [35, 36] demonstrated that popular classification losses, such as Focal Loss [28] or Label Smoothing [52], integrate an implicit term that maximizes the entropy of the network predictions, thus favoring low-confidence models. Other works to improve the accuracy of the uncertainty estimates during training include the use of MixUp [53,62], or enforcing a constant vector norm on the logits [56], among others. Nevertheless, all these methods have been proposed in the context of fully-supervised models, and the calibration of Prompt Learning and Adapter-based methods for CLIP remains unexplored in the literature.

3 Background

3.1 CLIP Zero-Shot Classification

CLIP [45] is a large vision-language model, trained via contrastive learning to produce visual representations from images x paired with their associated text descriptions T. To do so, CLIP consists of an image encoder and a text encoder . This generates the corresponding vision $z \in \mathbb{R}^d$ and class text $w_k \in \mathbb{R}^d$ embeddings, which are typically projected into an ℓ_2-normalized shared embedding space. Given a new task consisting in visually discriminating between K categories, the set containing all the text embeddings for all the K classes can be denoted as $W = \{w_k\}_{k=1}^{K}$, with $w_k =$ ("A PHOTO OF A [CLASS$_k$]"). At inference, this learning paradigm enables zero-shot prediction. More concretely, for a given set of K classes, and an ensemble of N different prompts per category, we can generate the set of available prompts as $T = \{\{T_{n,k}\}_{n=1}^{N}\}_{k=1}^{K}$. Then, a popular strategy [10, 45, 57] consists in obtaining a class zero-shot prototype, which is computed as $t_k = \frac{1}{N} \sum_{n=1}^{N} (T_{n,k})$. Then, for a given test image, the zero-shot prediction, $p = (p_k)_{1 \le k \le K}$, can be obtained as:

$$p_k = \frac{\exp(z^\top t_k /)}{\sum_{j=1}^{K} \exp(z^\top t_j /)} \tag{1}$$

where is a temperature hyperparameter, whose value is learned during training, and $z^\top t$ denotes the dot product operator[1].

[1] As vectors are ℓ_2-normalized, the dot product between these two vectors is equivalent to their cosine similarity.

3.2 Adaptation to novel tasks

Let us now consider a support set that contains a few labeled samples $S = \{(x_i, y_i)\}_{i=1}^{S}$, with $y \in \{0,1\}^K$ the ground truth vector associated with x. The vector of predicted logits of a given image i is defined as $l_i = (l_{ik})_{1 \leq k \leq K}$. In Prompt Learning methods, such as CoOp [66] or KgCoOp [65], the adaptation is done by modeling the input text T_k of a given class k as learnable continuous vectors. Thus, in contrast with zero-shot inference, where the resulting text embeddings are obtained as the mean over the different pre-defined prompts, in Prompt Learning these are optimized. To generate the logits, the learnable prompts are combined with the fixed visual embedding from the test image i, such that $l_{ik} = z_i^\top t_k / \tau$, which can then be integrated into Eq. (1) to minimize the cross entropy loss over the few labeled shots. The family of methods commonly referred to as Adapters [10, 50, 60, 63] proceeds differently, and learns transformations over the visual and text embeddings, yielding the following logits $l_{ik} = \theta_a(z_i, t_k, \theta)$, where θ_a is the set of learnable parameters of the Adapter. A more challenging scenario consists in adapting the text prompts at inference, which is commonly referred to as test-time prompt tuning [49]. As this setting does not include few-shot supports to adapt the prompts, the supervised cross-entropy objective is replaced by an unsupervised minimization of the Shannon entropy. Thus, regardless of the method selected, the objective is to optimize either t_k (Prompt Learning and test-time prompt tuning) or θ_a (Adapters) to minimize either the CE over the softmax predictions obtained from the few-shots, or the Shannon entropy on the test samples predictions at inference.

4 Constraining logits during adaptation

4.1 Impact of adaptation in logits

To understand the impact on calibration of using the cross-entropy (CE) loss to adapt CLIP, let us decompose the logit vector l into its Euclidean norm $\|l\| = \sqrt{l_1^2 + \ldots + l_K^2}$, (*magnitude*) and its unit vector $\hat{l}$ (*direction*), such that $l = \|l\|\hat{l}$. Considering now the *magnitude* and *direction* of the logit vector, the general form of the cross-entropy loss over a given support sample, using the softmax probabilities in Eq. (1), can be formulated as:

$$\log \frac{\exp\left(\|l_i\| \hat{l}_{ik}\right)}{\sum_{j=1}^{K} \exp\left(\|l_i\| \hat{l}_{ij}\right)} \tag{2}$$

This view of the cross-entropy implies that the direction of the logit vector $\hat{l}_i$ determines the predicted class of the image i. Thus, if the predicted category is incorrect, $\hat{l}_i$ will change to match the target class provided in the one-hot encoded label. Once the network prediction is correct, i.e., $y_i = \arg\max_j(l_{ij})$, the direction of the vector will remain unchanged. Nevertheless, the nature of the

cross-entropy loss will favor higher softmax probabilities for the predicted class. Recent literature [56] suggests that this is achieved by increasing $\|l_i\|$, indicating that the miscalibration issue originates from the augmentation of the logit norm. Nevertheless, in what follows we refute this argument and advocate for the increase of the logits range as the potential cause of miscalibration.

Proposition 1. *Let us consider the softmax cross entropy loss, where $(\cdot)$ denotes the softmax function. Assume that $l \neq 0$. Then, for any scalar $a > 0$, $\sigma_k(l) = \sigma_k(l + a)\ \forall k$, and $\|l + a\mathbf{1}\| > \|l\|$, where $\mathbf{1}$ denotes the vector of ones.*

Prop. 1 demonstrates that adding a strictly positive constant value $a \in \mathbb{R}_{++}$ to all the logits increases the norm of the vector l, but this does not necessarily lead to more confident predictions, whose probability scores remain unchanged.

Proposition 2. *Let $R(l) = \max(l) - \min(l)$ denotes the range of logit vector l, where $\max(l)$ (respectively $\min(l)$) denotes the largest (respectively smallest) value among the elements of vector l. Then, for any given scalar $a > 1$, and for $k = \arg\max_j(l_j)$, we have $\sigma_k(al) > \sigma_k(l)$ and $R(al) > R(l)$.*

The proofs of Propositions 1 and 2 are deferred to the Appendix. From the above proposition we find that increasing the range of a given logit vector results in higher softmax probability values. Thus, contrary to the widely spread belief that increasing the logit norm hinders model calibration, we argue that this effect of *logit distance magnification*, which yields higher softmax distributions, is a potential source of miscalibration[2]. This explains why, even though adaptation of CLIP yields performance gains in terms of accuracy, adapted models are worse calibrated than a zero-shot baseline. Furthermore, this analysis is supported empirically by the observations depicted in Fig. 2, where can observe that, while calibration has been degraded in the adapted models, the logit norm of their predictions has substantially decreased.

4.2 Our solution

From our previous analysis and empirical observations, we can derive that: *i)* despite improving their classification performance, state-of-the-art strategies to adapt CLIP suffer from miscalibration, particularly compared to the original zero-shot predictions, and *ii)* one of the main causes arises from the logit magnification issue introduced by the cross-entropy loss used during adaptation.

In light of these findings, we propose a simple but effective solution that can alleviate the miscalibration issue in CLIP adaptation. More concretely, we propose to constraint the range of the logits during the training of a main objective H, which results in the following constrained problem:

$$\begin{aligned} &\text{minimize} && H(Y;P) \\ &\text{subject to} && l_i^{\text{ZS-min}}\mathbf{1} \leq l_i \leq l_i^{\text{ZS-max}}\mathbf{1} \quad \forall i \in D; \end{aligned} \tag{3}$$

[2] Note that the same reasoning applies to TPT [49], whose learning objective to adapt the CLIP baseline is to minimize the Shannon entropy of the softmax distribution.

Fig. 2: Logit norm or logit range as the source of miscalibration? These gures clearly show that when the calibration of the zero-shot (ZS) model is degraded, the logit norm of its predictions is reduced (top), which discards an increase of the logit norm as the main cause for miscalibration. In contrast, there exists a correlation between the increase of the logit ranges and miscalibration (bottom).

where Y and P are matrices containing the sample-wise ground-truth and softmax-prediction vectors for all the samples involved in the training, $l_i^{\text{ZS-min}}$ and $l_i^{\text{ZS-max}}$ are the min and max logit magnitudes of the zero-shot prediction for sample x_i. $\mathcal{D}$ denotes a given set of available samples. Furthermore, in the test-time prompt tuning setting, we simply need to replace Y by P in Eq. (3). Directly solving the constrained problem in Eq. (3) in the context of deep models is not trivial [33], and Lagrangian-dual optimization has been typically avoided in modern deep networks involving millions of parameters. To address this issue, we propose several alternatives to approximate the constrained problem presented in Eq. (3), which are detailed below.

4.3 Zero-shot logit normalization during training (ZS-Norm)

The constraint in the presented problem, i.e., $l_i^{\text{ZS-min}} \mathbf{1} \leq l_i \leq l_i^{\text{ZS-max}} \mathbf{1}; \forall i \in \mathcal{D}$, can be integrated into the main objective by transforming the logits before computing the CE loss over the support set samples (here $\mathcal{D} = \mathcal{S}$). More concretely, the modi ed learning objective can be de ned as:

$$H(Y; P) = \sum_{i \in \mathcal{S}} \sum_{k=1}^{K} y_{ik} \log \left(\frac{\exp(l_{ik}')}{\sum_{j=1}^{K} \exp(l_{ij}')} \right); \tag{4}$$

where l_i' denotes the zero-shot normalized logit vector of x_i, obtained as:

$$l_i' = \frac{(l_i^{\text{ZS-max}} - l_i^{\text{ZS-min}})}{(l_i^{\max} - l_i^{\min})} (l_i - l_i^{\min} \mathbf{1}) + l_i^{\text{ZS-min}} \mathbf{1}; \tag{5}$$

with $l_i^{max} = max_j(l_{ij})$ and $l_i^{min} = min_j(l_{ij})$, respectively. While the calibration strategy formalized in Eq. 4 forces the direction of the logit vector to match the correct category encoded in the one-hot label, its magnitude is normalized according to the ZS logit range of image x_i. Note that this is different from the solution presented in [56], as the logit values are normalized by the logit norm, which does not guarantee that the logit values will be in a certain range.

4.4 Integrating explicit constraints in the learning objective (Penalty)

The problem in Eq. (3) can also be approximated by an unconstrained problem, for example by transforming the enforced inequality constraints into penalties, which are implemented with the ReLU function. The resulting learning objective can be formally defined as:

$$\min \quad H(Y;P) + \sum_{i \in S} \sum_{k=1}^{K} (\text{ReLU}(l_{ik} - l_i^{\text{ZS-max}}) + \text{ReLU}(l_i^{\text{ZS-min}} - l_{ik})); \quad (6)$$

where λ controls the trade-off between the main loss and the penalties. The intuition behind the penalties is that when the constraint in Eq. (3) is not satisfied, i.e., there exist logit magnitudes outside the zero-shot logit range, the value of the penalty term increases, backpropagating gradients to modify the logit values according to the enforced constraint. We would like to stress that a natural solution to tackle the constrained problem in Eq. (6) would be the use of Lagrangian multipliers. Nevertheless, as stated earlier, in the context of deep learning, these methods suffer from several well-known limitations, which include training instability and non-convergence due to the difficulty of convexifying loss functions [1, 2, 48]. Thus, despite its simplicity, the use of penalties has proven to be effective in constraining deep models on a myriad of problems, such as image segmentation [21], adversarial attacks [47], or modeling thermal dynamics [8].

4.5 Sample-adaptive logit scaling (SaLS)

Last, we explore a simple but efficient solution that is closely related to temperature scaling (TS) [11]. In particular, TS is a single-parameter variant of Platt scaling [44], which consists in learning the scaling hyperparameter α in Eq. (1). While this strategy has led to very competitive results, it requires an external validation set to fine-tune the value of α, which limits its use to learning scenarios with abundant labeled data and absence of distributional drifts [40]. Furthermore, α is fixed for a whole dataset, which is suboptimal from a sample-wise standpoint. To alleviate these issues, we propose to use the logit normalization defined in Eq. (5) at inference time to obtain the final softmax probability in Eq. (1). More concretely, for each sample i to be classified, we compute its zero-shot prediction, whose min and max logit values are used in Eq. (5) to scale the logit distribution of that sample i provided by the adaptation method selected.

This can be viewed as an unsupervised sample-wise temperature scaling during testing, which does not require additional validation samples to fix its value, and adapts to the specificity of each sample, regardless of distributional drifts.

5 Experiments

5.1 Setup

Datasets. We use popular datasets for benchmark few-shot [60, 66] and test-time [49] CLIP adaptation. Domain Generalization: the adaptation robustness to domain shifts is evaluated using ImageNet [6] distributions. Concretely, we sample a 16-shot training subset from ImageNet's training partition which is directly evaluated on out-of-distribution test data from ImageNetV2 [46], ImageNet-Sketch [55], ImageNet-A [17], and ImageNet-R [15]. Fine-grained tasks: calibration during test-time adaptation is assessed on an assembly of 11 datasets that include heterogeneous discriminative tasks. These include Imagenet [6], Caltech101 [9], OxfordPets [42], StanfordCars [23], Flowers102 [38], Food101 [3], FGVCAircraft [32], SUN397 [58], DTD [5], EuroSAT [14], and UCF101 [51] datasets. Note that for test time adaptation we uniquely employed their corresponding test partitions.

Selected methods. Our proposed calibration framework is agnostic to any adaptation strategy of zero-shot models. We evaluate its performance across different popular settings and state-of-the-art methods for CLIP adaptation. Prompt Learning (PL): CoOp [66], CoCoOp [65], ProGrad [67] and MaPLe [22] are considered as the baselines. Adapters: CLIP-Adapter [10], TIP-Adapter [63], and TaskRes [60] are used. Test-Time Adaptation: TPT [49] is selected as the primary method for test time prompt tuning, together with C-TPT [59], a concurrent method recently proposed for calibrating TPT.

CLIP adaptation. We now describe the experimental details for training the selected adaptation methods. Backbones: All experiments build upon CLIP [45], using its ResNet-50 [13] and ViT-B/16 [7] pre-trained weights. Text prompts: The textual descriptions for zero-shot representation of the target concepts used are the hand-crafted text prompts used in CoOp [66]. Image augmentations: For few-shot adaptation, we applied random zoom, crops, and flips, following [60, 65]. Regarding Prompt Learning methods, these transformations are applied continuously during training, while for Adapters, since feature representations are pre-computed, the number of augmentations per support sample is set to 20, following [50]. Finally, regarding test-time prompt tuning (TPT), we employed AugMix [16] as in [49] to form a 64-image batch from each original image. Training details: Adapters are trained following the recent benchmark in [50]. We optimized the Adapters for 300 epochs, using SGD optimizer with a Momentum of 0:9 and an initial learning rate of 0:1. In the case of PL, we set the context length of the prompt to 4 and trained CoOp and CoCoOp for 50 and 10 epochs, respectively. We set the same training schedule, optimizer, and learning

as in [66]. For ProGrad and MaPLe, we follow the training settings considered for domain generalization reported in theirs respective works [22, 67]. Likewise, for TPT, we optimized the learned prompt by doing a single step with AdamW optimizer, with the learning rate set to 0.005, as in [49].

Evaluation metrics. To measure the discriminative performance of the di erent methods, we use classi cation accuracy (ACC). In terms of calibration, we follow the standard literature and resort to the Expected Calibration Error (ECE). In particular, with N samples grouped into M bins $\{b_1, b_2, \dots, b_M\}$, the ECE is calculated as: $\sum_{m=1}^{M} \frac{|b_m|}{N} |acc(b_m) - conf(b_m)|$, where $acc(\cdot)$ and $conf(\cdot)$ denote the average accuracy and con dence in bin b_m.

Calibration details. We introduced three di erent alternatives to alleviate the miscalibration of adapted models (Sec. 4.2). For ZS-Norm and Penalty , we incorporated such modi cations during training (i.e. adaptation), and kept all implementation details previously presented. Furthermore, the penalty-based calibration weight in Eq. 6 is set to 10 and remains xed across all settings.

5.2 Results

I) Task 1: Few-shot domain generalization. Table 1 introduces the average few-shot generalization (OOD) results using black-box Adapters, whereas Table 2 presents the same for PL approaches. We refer the reader to Appendix for the detailed results per dataset. First, results consistently show a miscalibration phenomenon when CLIP models are adapted, regardless of the CLIP backbone

Table 1: Results for robust Adapters calibration. The average over the four ImageNet OOD datasets is reported. In brackets, we highlight the di erence with respect to each baseline, to stress the impact of the proposed methods (ZS-Norm , Penalty , and SaLS). Detailed results are reported in Appendix.

Method	Avg. OOD ACC	Avg. OOD ECE
Zero-Shot [45]	40.62	7.18
CLIP-Ad [10]	34.07	15.45
w/ ZS-Norm	30.06 (-4.01) #	21.27 (+5.82) "
w/ Penalty	35.20 (+1.13) "	11.22 (-4.23) #
w/ SaLS	34.07	8.95 (-6.50) #
TIP-Ad(f) [63]	41.45	19.04
w/ ZS-Norm	41.73 (+0.28) "	19.80 (+0.76) "
w/ Penalty	43.73 (+2.28) "	12.18 (-6.86) #
w/ SaLS	41.45	8.13 (-10.91) #
TaskRes [60]	41.18	11.25
w/ ZS-Norm	41.30 (+0.12) "	9.07 (-2.18) #
w/ Penalty	41.29 (+0.11) "	10.62 (-0.63) #
w/ SaLS	41.18	9.03 (-2.22) #

(a) ResNet-50

Method	Avg. OOD ACC	Avg. OOD ECE
Zero-Shot [45]	57.15	4.78
CLIP-Ad [10]	50.61	7.82
w/ ZS-Norm	49.73 (+0.88) #	12.53 (+4.71) "
w/ Penalty	51.59 (+0.98) "	6.38 (-1.44) #
w/ SaLS	50.61	4.38 (-3.44) #
TIP-Ad(f) [63]	25.86	63.63
w/ ZS-Norm	41.64 (+15.78) "	58.27 (-5.36) #
w/ Penalty	49.23 (+23.37) "	40.98 (-22.65) #
w/ SaLS	25.86	44.37 (-19.26) #
TaskRes [60]	58.01	7.52
w/ ZS-Norm	58.41 (+0.40) "	5.72 (-1.80) #
w/ Penalty	58.31 (+0.30) "	6.65 (-0.87) #
w/ SaLS	58.01	6.21 (-1.31) #

(b) ViT-B/16

Table 2: Results for robust Prompt Learning calibration. The average over the four ImageNet OOD datasets is reported. In brackets, we highlight the difference with respect to each baseline, to stress the impact of the proposed methods (ZS-Norm, Penalty and SaLS). Detailed results are reported in Appendix.

Method	Avg. OOD ACC	Avg. OOD ECE
Zero-Shot [45]	40.62	7.18
CoOp [66]	40.86	10.97
w/ ZS-Norm	41.59 (+0.73) ↑	10.19 (-0.78) ↓
w/ Penalty	41.87 (+1.01) ↑	8.06 (-2.91) ↓
w/ SaLS	40.86	7.82 (-3.15) ↓
CoCoOp [65]	43.36	7.69
w/ ZS-Norm	43.70 (+0.34) ↑	7.12 (-0.57) ↓
w/ Penalty	43.86 (+0.50) ↑	6.15 (-1.54) ↓
w/ SaLS	43.36	6.82 (-1.87) ↓
ProGrad [67]	42.32	7.66
w/ ZS-Norm	42.21 (+0.11) ↑	7.98 (+0.32) ↑
w/ Penalty	42.57 (+0.25) ↑	6.84 (-0.82) ↓
w/ SaLS	42.32	6.90 (-0.76) ↓

(a) ResNet-50

Method	Avg. OOD ACC	Avg. OOD ECE
Zero-Shot [45]	57.15	4.78
CoOp [66]	58.41	6.61
w/ ZS-Norm	58.75 (+0.34) ↑	4.35 (-2.26) ↓
w/ Penalty	59.18 (+0.77) ↑	4.91 (-1.70) ↓
w/ SaLS	58.41	4.90 (-1.71) ↓
CoCoOp [65]	59.74	4.83
w/ ZS-Norm	59.90 (+0.16) ↑	3.94 (-0.89) ↓
w/ Penalty	60.20 (+0.46) ↑	3.89 (-0.94) ↓
w/ SaLS	59.74	4.81 (-0.00)
MaPLe [22]	60.07	4.13
w/ ZS-Norm	60.09 (+0.02) ↑	3.59 (-0.14) ↓
w/ Penalty	60.62 (+0.55) ↑	3.78 (-0.35) ↓
w/ SaLS	60.07	4.38 (+0.25) ↑

(b) ViT-B/16

used, or the transferability approach. Few-shot Adapters calibration: We find that miscalibration is especially occurring in few-shot black-box Adapters. For example, Clip-Ad or TaskRes in Tab. 1 (a) show ECE increments of +8.3 and +4.0 respectively. This is further magnified when using the popular TIP-Adapter method. Few-shot PL calibration: PL approaches are relatively more robust in this setting (e.g. +3.8 CoOp in Tab. 2 (a)). On the impact of logit range regularization: Results show the potential of logit range scaling among its different proposed variants, improving calibration for all Prompt Learning approaches, and most of the used Adapters. Impact of different strategies to adjust logit range: . The only strategy that does not allow for consistent performance gains is ZS-Norm, which deteriorates performance in some Adapters (see Clip-Ad in Table 1). We believe that the re-parameterization in Eq 4 might not properly prevent logit range de-adjustment before normalization, and thus overfit to the few support samples. In contrast, Penalty constraint directly regularizes such values, showing consistent ECE decreases for both Adapters (e.g. -22.0 for TIP-Ad(f) using ViTs, or -4.3 for CLIP-Ad using RN50) and PL (e.g. -2.9 for CoOp using RN50, or -0.94 for CoCoOp using ViTs). Interestingly, as a side effect, we also observed accuracy improvements for domain generalization for several methods. Nevertheless, the best calibration performance is provided by a simple, yet effective post-processing standardization, SaLS. This is especially relevant, since this method does not require any modification of the adaptation strategy, and can be potentially applied to the output of any few-shot model.

II) Task 2: Test Time Adaptation (TTA). We report in Table 3 the performance for test-time prompt tuning across 11 fine-grained adaptation datasets for

Table 3: Test-time Prompt Learning calibration. Results for the popular TPT, as well as the concurrent work in [59], with ResNet-50 backbone, where our three solutions are implemented. Results on ViT-16 backbone are provided in the Appendix.

		Avg.	INet	CAL	PET	CAR	FLW	FOO	AIR	SUN	DTD	SAT	UCF
ACC	Zero-shot [45]	56.03	58.17	85.68	83.62	55.75	61.67	73.96	15.69	58.82	40.43	23.69	58.90
	TPT [49]	58.03	60.74	87.22	84.49	58.36	62.81	74.97	17.58	61.17	42.08	28.40	60.61
	w/ ZS-Norm	57.94	60.69	87.38	84.41	58.45	62.12	75.01	17.13	61.09	41.96	28.53	60.59
	w/ Penalty	57.69	60.74	87.06	84.30	58.13	61.84	75.17	17.22	61.11	42.02	26.60	60.35
	w/ SaLS	58.03	60.74	87.22	84.49	58.36	62.81	74.97	17.58	61.17	42.08	28.40	60.61
	C-TPT [59]	57.54	60.02	87.18	83.65	56.41	64.80	74.89	16.62	60.72	41.55	27.06	60.01
	w/ ZS-Norm	57.63	60.00	87.06	83.65	56.57	65.04	74.82	16.86	60.58	41.61	27.51	60.27
	w/ Penalty	57.52	60.06	86.94	83.51	56.78	64.76	74.88	16.29	60.67	41.90	26.63	60.32
	w/ SaLS	57.54	60.02	87.18	83.65	56.41	64.80	74.89	16.62	60.72	41.55	27.06	60.01
ECE	Zero-shot [45]	5.04	1.90	3.56	5.64	4.17	2.10	2.35	6.31	3.79	8.60	14.40	2.66
	TPT [49]	11.27	11.34	4.10	3.78	3.70	13.66	5.18	15.57	9.20	25.29	21.00	11.20
	w/ ZS-Norm	10.57	10.81	4.29	3.71	3.62	13.29	4.73	15.28	8.50	23.95	17.61	10.49
	w/ Penalty	9.58	11.31	3.99	1.57	2.26	13.94	4.27	14.51	8.88	23.10	11.82	9.78
	w/ SaLS	9.26	9.81	4.45	2.90	2.50	12.01	3.91	15.23	8.64	21.09	12.31	9.05
	C-TPT [59]	6.33	3.05	2.60	2.46	0.87	3.91	1.62	11.30	2.73	21.38	13.58	2.88
	w/ ZS-Norm	5.74	2.85	2.29	2.69	0.78	3.53	1.61	10.94	2.72	20.94	12.17	2.65
	w/ Penalty	3.14	5.93	2.26	2.66	0.81	3.79	1.64	11.58	2.74	20.49	10.83	2.51
	w/ SaLS	5.22	2.21	3.41	3.94	2.55	1.75	1.78	10.15	2.58	12.92	10.41	2.71

ResNet-50 backbone. Our results show that compared to zero-shot prediction, TPT largely deteriorates the calibration. Despite this degradation is somehow alleviated by C-TPT, further integrating our approaches show promising potential for better calibration of such methods, with consistent improvements for both strategies (e.g., 2:0 and 0:9 in ECE for TPT and C-TPT with SaLS).

III) Further constraining the logit range to smaller values. ZS predictions are well calibrated. Nevertheless, during adaptation, the model improves its discriminative performance at the cost of degrading its calibration capabilities. While in this work we advocate for increases of the logit range as a cause of miscalibration, decreasing this range should be done with care. In particular, further decreasing the logit range approaches a scenario of maximum entropy, where the predicted probabilities are semantically meaningless, leading to worse discrimination performance. This reasoning is empirically supported in Table 4, where we can see that, regardless of the learning paradigm, signi cantly decreasing the logit range yields higher ECE scores, i.e., miscalibration is magni ed.

IV) E ect on logits. Following one of our main observations (Fig. 2), we argued that the source of miscalibration in CLIP adaptation models is the increase of the logit range of their predictions, and not the logit norm. To empirically validate this hypothesis, we depict in Fig. 3 both the logit norm and logit ranges for a relevant method of each category, as well as the version improved with our SaLS solution, across the four OOD datasets of ImageNet. We can observe that, indeed, applying our approach (which improves calibration) leads to reduced logit ranges (bottom), whereas the logit norm (top) typically increases.

Table 4: What if the logit range is further decreased? ECE scores on ImageNet shifts (V2, S, A and R) for representative methods when reducing the original ZS logit range (denoted as 1) to half (1=2) and one quarter (1=4) in SaLS .

	CLIP-Ad			CoOp			TPT		
ZS-Range	1	1/2	1/4	1	1/2	1/4	1	1/2	1/4
RN50	8.95	21.31	31.51	7.82	24.44	37.72	16.74	25.15	40.7

Fig. 3: E ect of calibrating adapted CLIP models. Mean of the distribution of logit norms (top) and logit ranges (bottom) across the four ImageNet OOD datasets for a relevant Adapter-based (CLIP-Ad), Prompt Learning (CoOp) and TPT approach.

6 Conclusions

We have investigated the miscalibration issue of popular CLIP adaptation approaches on the challenging task of few-shot and zero-shot adaptation under distributional drifts. We have analyzed the source of this issue and demonstrated that, in contrast to existing evidence pointing to the logit norm, increases in the range of predicted logits might be a potential cause of miscalibration on the adapted models. To overcome this issue, we have presented three simple solutions, which consist in constraining the logit ranges to the values of the zero-shot predictions, either at training or test time. Extensive experiments on multiple models from the three categories, and popular OOD benchmarks, demonstrate that incorporating our simple solution to existing CLIP adaptation approaches considerably enhances their calibration performance, without sacri cing model accuracy. The proposed approach is model-agnostic, and demonstrate superior performance regardless of the family of approaches or setting, making of our model an appealing yet simple solution for zero-shot and few-shot CLIP adaptation, particularly in the challenging scenario of out-of-distribution data.

Acknowledgments. This work is supported by the National Science and Engineering Research Council of Canada (NSERC), via its Discovery Grant program. We also thank Calcul Quebec and Compute Canada.

References

1. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series). Athena Scienti c, 1 edn. (1996)
2. Birgin, E.G., Castillo, R.A., Martínez, J.M.: Numerical comparison of augmented lagrangian algorithms for nonconvex problems. Computational Optimization and Applications 31(1), 31 55 (2005)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 mining discriminative components with random forests. In: European Conference on Computer Vision (ECCV) (2014)
4. Cheng, J., Vasconcelos, N.: Calibrating deep neural networks by pairwise constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13709 13718 (2022)
5. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3606 3613 (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248 255 (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR) (2021)
8. Drgo¬a, J., Tuor, A.R., Chandan, V., Vrabie, D.L.: Physics-constrained deep learning of multi-zone building thermal dynamics. Energy and Buildings 243, 110992 (2021)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Worskshops (CVPRW). pp. 178 178 (2004)
10. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision (IJCV) (2024)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning (ICML). pp. 1321 1330. PMLR (2017)
12. Hantao Yao, Rui Zhang, C.X.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
14. Helber, P., Bischke, B., Dengel, A., Borth, D.: Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classi cation. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 3606 3613 (2018)
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). p. 8340 8349 (2021)

16. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple method to improve robustness and uncertainty under data shift. In: International Conference on Learning Representations (ICLR) (2020)
17. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15262 15271 (2019)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (ICML). pp. 4904 4916 (2021)
20. Joy, T., Pinto, F., Lim, S.N., Torr, P.H., Dokania, P.K.: Sample-dependent adaptive temperature scaling for improved calibration. In: Proceedings of the AAAI Conference on Arti cial Intelligence. vol. 37, pp. 14919 14926 (2023)
21. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B.: Constrained-cnn losses for weakly supervised segmentation. Medical image analysis54, 88 99 (2019)
22. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19113 19122 (2023)
23. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for ne-grained categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 3498 3505 (2012)
24. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR). pp. 1 42 (2022)
25. Larrazabal, A., Martinez, C., Dolz, J., Ferrante, E.: Maximum entropy on erroneous predictions (MEEP): Improving model calibration for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2023)
26. Li, X., Lian, D., Lu, Z., Bai, J., Chen, Z., Wang, X.: Graphadapter: Tuning vision-language models with dual knowledge graph. Advances in Neural Information Processing Systems (NeurIPS)36 (2024)
27. Liang, X., Wu, Y., Han, J., Xu, H., Xu, C., Liang, X.: E ective adaptation in multi-task co-training for uni ed autonomous driving. Advances in Neural Information Processing Systems (NeurIPS)35, 19645 19658 (2022)
28. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 2980 2988 (2017)
29. Liu, B., Ben Ayed, I., Galdran, A., Dolz, J.: The devil is in the margin: Margin-based label smoothing for network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 80 88 (2022)
30. Liu, B., Rony, J., Galdran, A., Dolz, J., Ben Ayed, I.: Class adaptive network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16070 16079 (2023)
31. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and

tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21152–21164 (2023)
32. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. In: ArXiv Preprint (2013)
33. Márquez-Neila, P., Salzmann, M., Fua, P.: Imposing hard constraints on deep networks: Promises and limitations. arXiv preprint arXiv:1706.02025 (2017)
34. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: International Conference on Learning Representations (ICLR). pp. 1–17 (2023)
35. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems (NeurIPS) **33**, 15288–15299 (2020)
36. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Advances in neural information processing systems (NeurIPS) **32** (2019)
37. Murugesan, B., Adiga Vasudeva, S., Liu, B., Lombaert, H., Ben Ayed, I., Dolz, J.: Trust your neighbours: Penalty-based constraints for model calibration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 572–581 (2023)
38. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (2008)
39. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Black box few-shot adaptation for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
40. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems (NeurIPS) **32** (2019)
41. Park, H., Noh, J., Oh, Y., Baek, D., Ham, B.: Acls: Adaptive and conditional label smoothing for network calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3936–3945 (2023)
42. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 3498–3505 (2012)
43. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: International Conference on Learning Representations (ICLR) (2017)
44. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
46. Recht, B., Roelofs, R., Schmidt, L., , VaishaalShankar: Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning (ICML). pp. 5389–5400 (2019)
47. Rony, J., Granger, E., Pedersoli, M., Ben Ayed, I.: Augmented lagrangian adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7738–7747 (2021)

48. Sangalli, S., Erdil, E., Hötker, A., Donati, O.F., Konukoglu, E.: Constrained optimization to train neural networks on critical and under-represented classes. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
49. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems (NeurIPS) **35**, 14274–14289 (2022)
50. Silva-Rodriguez, J., Hajimiri, S., Ayed, I.B., Dolz, J.: A closer look at the few-shot adaptation of large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2024)
51. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. In: ArXiv Preprint (2012)
52. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 2818–2826 (2016)
53. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)
54. Tomani, C., Cremers, D., Buettner, F.: Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In: European Conference on Computer Vision (ECCV). pp. 555–569. Springer (2022)
55. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
56. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning (ICML). pp. 23631–23644. PMLR (2022)
57. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7959–7971 (2022)
58. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3485–3492 (2010)
59. Yoon, H.S., Yoon, E., Tee, J.T.J., Hasegawa-Johnson, M.A., Li, Y., Yoo, C.D.: C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In: International Conference on Learning Representations (ICLR) (2024)
60. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10899–10909 (2023)
61. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18123–18133 (2022)
62. Zhang, L., Deng, Z., Kawaguchi, K., Zou, J.: When and how mixup improves calibration. In: International Conference on Machine Learning. pp. 26135–26160. PMLR (2022)

63. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. In: European Conference on Computer Vision (ECCV). pp. 1–19 (11 2022)
64. Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., Li, H.: Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15211–15222 (2023)
65. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
66. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) (2022)
67. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15659–15669 (2023)