Open-Vocabulary Calibration for Fine-tuned CLIP

Shuoyuan Wang [1 2 †] **Jindong Wang** [3] **Guoqing Wang** [4] **Bob Zhang** [2] **Kaiyang Zhou** [5] **Hongxin Wei** [1]

Abstract

Vision-language models (VLMs) have emerged as formidable tools, showing their strong capability in handling various open-vocabulary tasks in image recognition, text-driven visual content generation, and visual chatbots, to name a few. In recent years, considerable efforts and resources have been devoted to adaptation methods for improving the downstream performance of VLMs, particularly on parameter-efficient fine-tuning methods like prompt learning. However, a crucial aspect that has been largely overlooked is the confidence calibration problem in fine-tuned VLMs, which could greatly reduce reliability when deploying such models in the real world. This paper bridges the gap by systematically investigating the confidence calibration problem in the context of prompt learning and reveals that existing calibration methods are insufficient to address the problem, especially in the open-vocabulary setting. To solve the problem, we present a simple and effective approach called Distance-Aware Calibration (DAC), which is based on scaling the temperature using as guidance the distance between predicted text labels and base classes. The experiments with 7 distinct prompt learning methods applied across 11 diverse downstream datasets demonstrate the effectiveness of DAC, which achieves high efficacy without sacrificing the inference speed. Our code is available at https://github.com/ml-stat-Sustech/CLIP_Calibration.

[†]Work done while working at SUSTech as a visiting scholar. [1]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China. [2]Department of Computer and Information Science, University of Macau, Taipa, Macau [3]William & Mary, Williamsburg, Virginia, USA [4]School of Computer Science and Engineering, University of Electronic Science and Technology of China, China. [5]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China. Correspondence to: Hongxin Wei <weihx@sustech.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Vision-language models (VLMs) like CLIP (Radford et al., 2021) have achieved strong performance in open-vocabulary image recognition by leveraging natural language supervision (Radford et al., 2021). To strengthen performance for VLMs in downstream applications, numerous fine-tuning methods have been studied in the literature, among which prompt learning (Zhou et al., 2022b) has gained the most attention due to its parameter-efficient design and robustness. Ideally, in an open-vocabulary setting a fine-tuned VLM should make accurate and reliable predictions on both seen (base) and unseen (novel) classes. While prompt learning algorithms can achieve considerable improvement in accuracy, the reliability issue is still underexplored, preventing the deployment of such models in high-stakes applications like medical diagnosis and autonomous driving.

Existing research has shown that the pre-trained CLIP model is well-calibrated in the zero-shot setting (Minderer et al., 2021). However, after fine-tuning for downstream tasks, the model struggles with miscalibration, i.e., the predicted class probabilities do not align with the true correctness likelihood (Oh et al., 2023). Existing research has only studied the calibration of fine-tuned CLIP on base classes. In this work, we present the first study on calibrating fine-tuned VLMs under an open-vocabulary setting. Our experiments reveal that fine-tuned VLMs exhibit overconfidence in novel classes but underconfidence in base classes. Moreover, we find that the current state-of-the-art post-hoc calibration methods can mitigate the miscalibration problem on base classes. However, the calibration result fails to transfer to novel classes, which motivates us to design post-hoc methods for open-vocabulary calibration.

In this work, we show that the miscalibration issue on novel classes can be mitigated through a simple fix to temperature scaling (Guo et al., 2017)—the most popular method for post-hoc calibration—by scaling the temperature with class information. Our approach, *Distance-Aware Calibration* (**DAC**), is motivated by our analysis on the deviation degree of novel classes. Specifically, we show that prompt tuning leads to a large gap between the text features of base classes and those of novel classes. Furthermore, as the deviation degree of a novel class increases, the model becomes more "confident" in its predictions.

Concretely, the key idea behind DAC is to scale the temperature value based on the distance between the text embeddings of a novel class and those of base classes. This can be achieved by estimating a textual deviation score for each novel class, which quantifies the deviation degree of the corresponding normalized text features. With the class-dependent scaling factors, our method generates high temperatures for those novel classes far away from the base classes, improving the calibration performance under open-vocabulary settings.

We evaluate our approach DAC using 7 distinct prompt learning methods applied across 11 diverse downstream datasets. The results show that DAC consistently improves open-vocabulary calibration on all prompt learning methods. For example, DAC achieves an average of 6.84% reduction in Expected Calibration Error (ECE) for CoOp (Zhou et al., 2022b) over the 11 datasets, with maximum reduction reaching 16%. The same improvement can be achieved for other prompt learning methods as well, such as MaPLe (Khattak et al., 2023a) and PromptSRC (Khattak et al., 2023b). Beyond prompt tuning methods, we also show that DAC can boost the open-vocabulary calibration performance for existing post-hoc methods such as Density-Ratio Calibration (Xiong et al., 2023).

As a post-hoc calibration method, DAC can be easily integrated with various prompt tuning methods in a plug-and-play manner. Our method is computationally efficient and does not require hyperparameter tuning (See Figure 5). In short, the main contributions of this paper are summarized as follows:

- **Problem**: We find that fine-tuned VLMs often suffer from miscalibration, and existing post-hoc calibration methods often fail in the open-vocabulary setting.
- **Analysis**: We empirically study the correlation between the calibration and the textual distribution gap. We show that after prompt learning, VLMs tend to be overconfident on classes far away from base classes.
- **Method**: We introduce DAC, a simple and effective post-hoc calibration method that rectifies the predicted confidence while maintaining classification accuracy.
- **Evaluation**: Extensive experiments with 7 tuning VLM tuning methods applied across 11 datasets demonstrate the effectiveness of our approach.

2. Background

Contrastive Language-Image Pretraining (CLIP). CLIP is a vision-language model that measures the alignment between image and text, which enables zero-shot inference for open-vocabulary classes (Radford et al., 2021). Let ϕ and θ denote CLIP's image and text encoders, respectively. Given an image instance $\boldsymbol{x} \in \mathbb{R}^d$ and a text label c, the logit function of CLIP can be formulated as:

$$L_c^{clip}(\boldsymbol{x}_i) = \tau \cdot \operatorname{sim}\left(\phi(\boldsymbol{x}), \psi(\boldsymbol{t}_c)\right), \tag{1}$$

where ϕ and ψ denote the image and text encoders in CLIP. $\boldsymbol{t}_c$ is derived from a hand-crafted prompt like "a photo of a {class}", where the "{class}" is filled with the text label c. τ is generally set as a constant of 100.

For multi-class classification, we predict by selecting the label with the highest probabilities among the label candidate set $\mathcal{C} = \{c_i\}_{i=1}^K$, as shown below:

$$c^* = \arg\max_{c \in \mathcal{C}} p(c|\boldsymbol{x}) = \arg\max_{c \in \mathcal{C}} \frac{e^{L_c^{clip}(\boldsymbol{x})}}{\sum_{i=1}^K e^{L_i^{clip}(\boldsymbol{x})}} \tag{2}$$

where $p(c|\boldsymbol{x})$ is the predicted probability of class c for the instance $\boldsymbol{x}$.

Prompt tuning. To strengthen the CLIP model in downstream applications, prompt tuning algorithms are normally designed to optimize the context prompt (Zhou et al., 2022b;a) for high efficiency.

In particular, CoOp (Zhou et al., 2022b) replace the hand-crafted textual tokens with a set of learnable textual token $\mathcal{T} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M\}$, where M is the length of tokens. Then, the logit function is:

$$L_c^{tuned}(\boldsymbol{x}) = \tau \cdot \operatorname{sim}\left(\phi(\boldsymbol{x}), \psi(\boldsymbol{t}'_c)\right), \tag{3}$$

where $\boldsymbol{t}'_c = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, ...\boldsymbol{v}_M, \boldsymbol{c}]$ and $\boldsymbol{c}$ denotes the textual embedding of class c. During tuning, the learnable textual tokens $\mathcal{T}$ are optimized to minimize the cross-entropy loss ℓ_{ce} with labeled few-shot samples $\mathcal{D} = \{(\boldsymbol{x}_i, c_i)\}_{i=1}^N$. Formally, the optimization of CoOp is:

$$\mathcal{T}^\star = \arg\min_{\mathcal{T}} \frac{1}{N} \sum_{i=1}^N \ell_{ce}(p(c_i|\boldsymbol{x}_i))$$

To maintain the generalization on the unseen classes within the same task, i.e., open-vocabulary classes, CoCoOp (Zhou et al., 2022a) employs instance-conditional prompts $\boldsymbol{t}_m(\boldsymbol{x}) = \boldsymbol{t}_m + h_\theta(\boldsymbol{x})$, where $h_\theta(\cdot)$ is a neural network parameterized by θ and $m \in \{1, 2, \ldots, M\}$. While the tuning methods achieve promising performance, the reliability of the prediction probability $p(c|\boldsymbol{x})$ is still a mystery for tuned VLMs. We proceed by introducing the calibration in the prompt tuning setting.

Confidence calibration. For multi-class classification tasks, vision-language models are designed to produce the class probabilities $p(c|\boldsymbol{x})$ as defined in Equation 2. In addition to high accuracy, it is generally expected for vision-language models to be well calibrated, i.e., the predicted

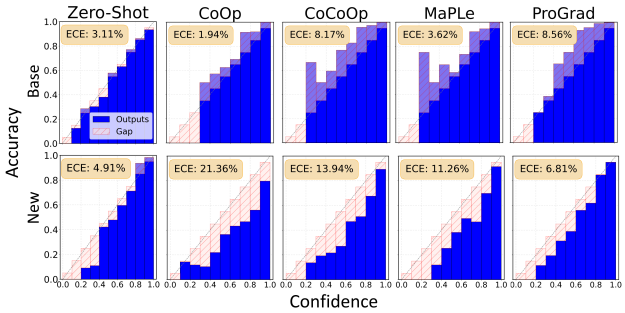


Figure 1. Reliability of fine-tuned CLIP (ViT-B/16) on the Flower102 dataset. ECE: Expected Calibration Error (lower is better). Miscalibration is depicted in pink for overconfidence and purple for underconfidence.

class probabilities can faithfully estimate the true probabilities of correctness. Formally, a *perfectly* calibrated model satisfies $Pr(c^\star = c \mid p(c|\boldsymbol{x}) = q) = q, \forall q \in [0, 1]$.

To quantify the degree of miscalibration, the *Expected Calibration Error* (ECE) is defined as the difference between accuracy and confidence. With N samples grouped into K bins $\{b_1, b_2, \ldots, b_K\}$, the ECE is calculated as:

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|b_k|}{N} \left|\mathrm{acc}\left(b_k\right) - \mathrm{conf}\left(b_k\right)\right|, \tag{4}$$

where $\mathrm{acc}\left(\cdot\right)$ and $\mathrm{conf}\left(\cdot\right)$ denotes the average accuracy and confidence in bin b_k.

3. An Empirical Study

In the literature, it has been shown that pre-trained CLIP is well calibrated in zero-shot inference (Minderer et al., 2021). In this section, we empirically study the calibration performance of tuned VLMs. In particular, we compare the zero-shot CLIP (Radford et al., 2021) with a variety of popular prompt tuning algorithms, including: CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), MaPLe (Khattak et al., 2023a), and ProGrad (Zhu et al., 2023).

3.1. Are fine-tuned VLMs well-calibrated?

To showcase the miscalibration in tuned VLMs, we fine-tune the pre-trained CLIP with 7 tuning methods on 11 downstream datasets and use ECE as the calibration metric. The descriptions of compared methods and datasets can be found in Subsection 5.1. The miscalibration of tuned CLIP on the Flower102 dataset is visualized in Figure 1 through reliability diagrams. Intuitively, one might expect that deep models will demonstrate increased confidence in the learned base classes after tuning and tend to make more conservative predictions under open-vocabulary conditions since they are not explicitly learned during the tuning process. On the contrary, we make a counter-intuitive observation that *tuned VLMs tend to be underconfident in base classes while prone to overconfident in new classes.* We present the detailed calibration results in Appendix C.1.

Table 1. ECE (%) of fine-tuned CLIP with different calibration methods. We use ProDA to fine-tune CLIP-ViT-B/16 on ImageNet-1K. "ZS" means zero-shot CLIP and "Conf" means confidence score without calibration after tuning. "–" means the results are not applicable. "Conf" shows underconfidence in base classes. "TS" and "DEN" show overconfidence in new classes.

	ZS	Conf	TS	DEN	HB	IR	MIR
Base classes	3.58	4.82	1.94	0.73	4.23	2.09	0.82
New classes	2.09	1.59	3.90	3.86	-	-	-

3.2. Can fine-tuned VLMs be calibrated?

To address the above miscalibration in tuned CLIP, we utilize commonly-used post-hoc confidence calibration for tuned VLMs and expect to calibrate the model well on both base and new classes. Specifically, post-hoc calibration can be split into two categories: *scaling-based methods* and *binning-based methods*. Specifically, regarding the former, we use Temperature Scaling (TS) (Guo et al., 2017) and Density-Ratio Calibration (DEN) (Xiong et al., 2023). For the latter, we use Histogram Binning ((Zadrozny & Elkan, 2001)), Isotonic Regression (IR) (Zadrozny & Elkan, 2002) and Multi-Isotonic Regression (MIR) (Zhang et al., 2020). The complete calibration results are available in Appendix C.2. The calibration results on ImageNet-1K are shown in Table 1, and we can make the following findings:

(1) **Post-hoc calibration can remedy miscalibration in base classes.** As a complement of the observation in the recent works (Tu et al., 2023; Oh et al., 2023), we observe that both scaling-based methods (e.g., DEN) and bin-based approaches (e.g., MIR) can effectively reduce miscalibration in base classes. (2) **Post-hoc calibration on base classes can not transfer to new classes.** For instance, the calibration performance of "TS" and "DEN" is worse than the model without calibration ("Conf") or tuning ("ZS"). This phenomenon indicates that simply using the calibrator learned in base classes is not optimal for wider unseen distributions (Yu et al., 2022). For bin-based methods, they need the probabilities from base classes as the input, which are incompatible with zero-shot prediction. This limitation leads to an open question: *How to calibrate tuned VLMs in open-vocabulary classes?*

4. Open-Vocabulary Calibration

In this section, we first investigate the open-vocabulary calibration with a visualization of the joint vision-language embedding space and show that the textual gap between the

base and new classes is crucial for calibration. Next, we quantify this gap by a distance-based metric for the deviation degree of a novel class. We then present our method and its advantages for open-vocabulary calibration.

4.1. Feature Space Analysis

In the prompt tuning of VLMs, the primary influence of tuning is reflected in the textual features. This fact motivates us to explore how these changes in text features correlate to the miscalibration in the new classes. To understand the open-vocabulary miscalibration in VLMs after prompt tuning, we first analyze it from the joint vision-language representation space. Specifically, following *modality gap*[1] (Liang et al., 2022), paired image-text of the downstream datasets are fed into zero-shot / tuned CLIP and their embeddings are visualized in 2D using singular value decomposition (SVD).

As mentioned in (Liang et al., 2022), we observe that images and text are embedded at arm's length in the shared representation space. In zero-shot CLIP, the embeddings of the same modality are relatively concentrated. However, as is shown in Figure 2b, there is a distinct textual distribution gap between the base and new classes after tuning. Furthermore, ECE has been significantly increased, indicating that the model had been miscalibrated. Informally, we hypothesize that *the deviation degree in the textual gap is crucial for open-vocabulary calibration.*

4.2. Motivation

To verify the hypothesis, we first introduce a distance-based metric *proximity* to quantify the deviation degree of features.

Definition 4.1 (Proximity (Xiong et al., 2023))**.** Consider a feature $\mathbf{z} \in \mathbb{R}^d$ as the embedding of a test sample and the held-out feature embeddings $\mathcal{E} \in \mathbb{R}^{n \times d}$, proximity is a function inversely correlates with the mean distance between the test sample and its K nearest neighbors in held-out sets:

$$P(\mathbf{z}, \mathcal{E}) = \sigma\left(\frac{1}{K} \sum_{\mathbf{x}_i \in \mathcal{N}_K(\mathbf{z}, \mathcal{E})} \operatorname{dist}(\mathbf{z}, \mathbf{x}_i)\right), \quad (5)$$

where $\mathcal{N}_K(\mathbf{z}, \mathcal{E})$ denotes the set of K nearest neighbors of $\mathbf{z}$ and $\operatorname{dist}(\cdot, \cdot)$ is a predefined metric for distance calculation σ is a non-negative, monotonically decreasing function. This definition implies a greater likelihood that the test sample originates from the same distribution as the held-out set as its proximity increases.

In VLMs, we have a set of l_2-normalized (Radford et al., 2021) textual embeddings of seen class $\mathcal{W} = \{\boldsymbol{w}_j\}_{j=1}^{S}$. Given an arbitrary class $c_i \in \mathcal{U}$ in the test time and its corresponding class token is $\boldsymbol{c}_i$. We first obtain its normalized textual embedding $\boldsymbol{w}_i = \theta(\boldsymbol{t}_i)$, where $\boldsymbol{t}_i = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, ...\boldsymbol{v}_M, \boldsymbol{c}_i\}$ represents the textual token. we can define the textual distribution gap between this class c_i and the seen class as follows:

$$P(\boldsymbol{w}_i, \mathcal{W}) = \exp\left(-\frac{1}{K} \sum_{\boldsymbol{w}_j \in \mathcal{N}_K(\boldsymbol{w}_i, \mathcal{W})} \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2\right). \quad (6)$$

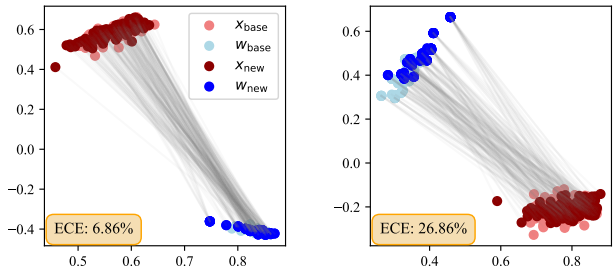
Here we use e^{-x} as $\sigma(\cdot)$ and l_2-distance for $\operatorname{dist}(\cdot, \cdot)$.

We present the correlation between the deviation degree of textual feature and calibration in Figure 3. For each textual embedding of new classes, we can calculate its corresponding proximity to the hold-out set of base textual features. We can observe that:

Lower proximity correlates with higher confidence and ECE. Specifically, if the given text is far from the base textual distribution, tuned VLM tends to present a higher confidence, which results in worse calibration performance. Similar to the pre-training stage (Liang et al., 2022), contrastive loss in V-L tuning bridges a connection between visual and textual modality. Each modality will be embedded into a relatively concentrated distribution. However, in the test time, if the given class is an outlier from the textual distribution learned in tuning, tuned CLIP may not be closely aligned across modalities and fail to give a proper predicted confidence level. Furthermore, as shown in Figure 3b, commonly used calibration techniques such as Temperature Scaling (TS) do not alleviate but deteriorate this issue. We present more details of the experiment are included in Appendix B.1. This motivated us to modify the confidence level when the prediction exhibits lower textual proximity.

4.3. Distance-Aware Calibration

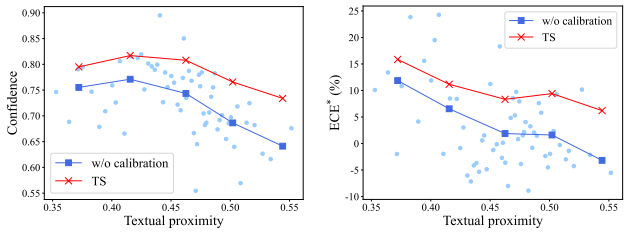
Per the analysis above, the predicted class with lower textual proximity can result in higher confidence and worse



(a) Zero-Shot (b) CoOp

Figure 2. Paired inputs from image (x) / text (w) are sampled from the DTD dataset fed into zero-shot / tuned CLIP and are visualized in 2D using SVD. Compared with zero-shot CLIP, CoOp has a larger textual distribution gap between the base and new classes

[1] https://github.com/Weixin-Liang/Modality-Gap/



(a) Confidence (b) ECE*

Figure 3. Class-wise performance on StanfordCars dataset after tuning. ECE* with a positive (negative) value denotes overconfidence (underconfidence). The scatters represent the origin results and the broken line denotes the bin-based results. Confidence and ECE* increase as proximity decreases. Temperature scaling (TS) can not mitigate the overconfidence.

calibration performance. In this work, we propose a post-hoc method to rectify the predicted confidence level. More specifically, we introduce a textual deviation-based score to scale the temperature value, thereby updating the confidence level to align with the true correctness likelihood.

Textual deviation estimation. Ideally, the model is expected to give highly uncertain predictions for examples from novel classes, with relatively low accuracy. With this in mind, we define the score by calculating the deviation distances of novel classes to base classes using tuned CLIP models, compared to those of pre-trained CLIP.

Let $\boldsymbol{w}_i$ and $\boldsymbol{w}'_i$ be the normalized text features of class c_i from the pre-trained and tuned VLMs respectively. The **T**extual **D**eviation (**TD**) score for class c_i is formulated as:

$$\gamma(c_i) = \frac{P(\boldsymbol{w}'_i, \mathcal{W}')}{P(\boldsymbol{w}_i, \mathcal{W})}, \tag{7}$$

where $\mathcal{W}$ and $\mathcal{W}'$ denote the sets of normalized textual features for base classes. For example, a low TD score indicates that the corresponding class is more likely to be deviant for the textual distribution of base classes. As demonstrated in Subsection 4.2, tuned VLMs tend to be overconfident in examples from such classes. Thus, we propose to update the confidence level according to the TD score for each class.

Calibrated inference. Before evaluation, we first collect the TD scores $\Gamma = \{q_i\}_{i=1}^{K}$ for each class. Given a test instance $\boldsymbol{x}$, we first calculate its predicted class $\hat{c} = \arg\max_c p(c \mid \boldsymbol{x})$ using the tuned CLIP model. Finally, the rectified logit is calculated as :

$$L_c^{dac}(\boldsymbol{x}) = \gamma(\hat{c}) \cdot \tau \cdot \operatorname{sim}\left(\phi(\boldsymbol{x}), \psi(\boldsymbol{t}'_c)\right). \tag{8}$$

where $\boldsymbol{t}'_c$ is the textual token of class c, defined in Equation 3. Here, we rectify the confidence level of the prediction by scaling the temperature value τ with TD score $\gamma(\hat{c})$. For base classes, we set $\gamma_{\hat{c}} = 1$, and our DAC is degraded to vanilla temperature scaling. Thus, our method would not affect the calibration performance on base classes. Noticeably, DAC offers several compelling advantages:

- **Realistic**: The calculation procedure of DAC only utilizes the information of text labels. Thus, our method does not assume access to the visual information from open-vocabulary concepts, which is challenging to obtain in real applications.
- **Algorithm-agnostic**: DAC can easily combine with existing prompt tuning algorithms, consistently improving the calibration performance in novel classes.
- **Easy to use**: our method does not require hyperparameter tuning (see Table 5) and does not introduce extra computational cost (See Table 6).

5. Experiments

In this section, we verify the effectiveness of our proposed DAC with 11 downstream classification benchmark datasets.

5.1. Experimental Setup

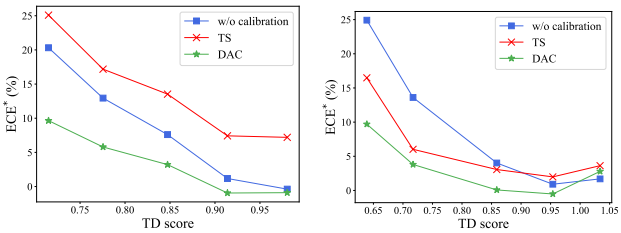
Benchmark setting. We evaluate the confidence calibration of existing popular prompt tuning methods. Following the few-shot evaluation protocol (Zhou et al., 2022a), the datasets are split into base and new classes. The model is trained only on the base classes in a few-shot setting and we generally report the calibration performance on new classes.

Datasets. In this work, we follow the standard base-to-new setting (Zhou et al., 2022b;a) and use 11 image recognition datasets in our experiment. The datasets cover diverse classification tasks. It includes general object datasets like ImageNet (Deng et al., 2009) and Caltech101 (Fei-Fei et al., 2004), alongside fine-grained classification datasets such as OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and FGVCAircraft (Maji et al., 2013) . For specialized categories, it features SUN397 (Xiao et al., 2010) for scene recognition, UCF101 (Soomro et al., 2012) for action recognition, DTD (Cimpoi et al., 2014) for texture classification, and EuroSAT (Helber et al., 2019) with its satellite images.

Compared methods. We compare and incorporate our method with existing VLM tuning algorithms, which can be split into two categories: *prompt tuning* and *multi-modal tuning*: For prompt tuning, we use Context Optimization (CoOp) (Zhou et al., 2022b), Conditional Context Optimization (CoCoOp) (Zhou et al., 2022a), Prompt Distri-

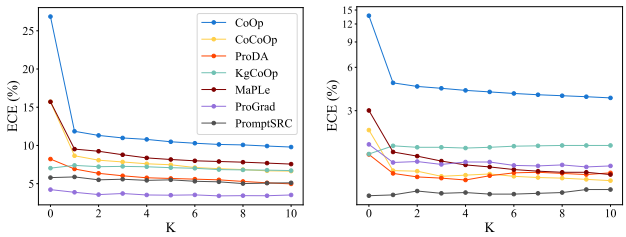
Table 2. Average calibration performance across 11 datasets. "Conf" represents the origin performance on open-vocabulary classes with existing tuning methods. "DAC" to our method applied to existing tuning methods. ↓ indicates smaller values are better. Calibration error is given by $\times 10^{-2}$. **Bold** numbers are significantly superior results.

Method	ECE(↓)		ACE(↓)		MCE(↓)		PIECE(↓)	
	Conf	DAC	Conf	DAC	Conf	DAC	Conf	DAC
CoOp	13.84	**7.00**	13.76	**6.91**	3.80	**1.71**	14.71	**9.02**
CoCoOp	6.29	**4.82**	6.21	**4.77**	1.79	**1.40**	8.07	**7.15**
ProDA	4.27	**3.99**	4.35	**4.08**	1.27	1.32	6.57	**6.35**
KgCoOp	4.36	4.32	4.43	4.38	1.18	1.13	6.67	6.63
MaPLe	5.77	**4.61**	5.71	**4.64**	1.82	**1.42**	7.59	**6.98**
ProGrad	4.22	**3.74**	4.27	**3.74**	1.22	**1.09**	6.75	**6.55**
PromptSRC	3.84	**3.63**	3.92	**3.69**	1.09	1.08	6.26	6.17



(a) StanfordCars (b) SUN397

Figure 4. ECE* (%) performance with difference calibration methods. ECE* with a positive (negative) value denotes overconfidence (underconfidence). Our proposed DAC largely mitigates the overconfidence in the predicted classes with lower TD score.



(a) DTD (b) SUN397

Figure 5. Hyperparameter sensitivity of the number of neighbors used in computing textual proximity. The miscalibration will be noticeably mitigated if $K > 1$.

bution Learning (ProDA) (Lu et al., 2022), Knowledge-guided Context Optimization (KgCoOp) (Yao et al., 2023), Prompt-aligned Gradient (ProGrad) (Zhu et al., 2023). For multi-modal tuning, we use Multi-modal Prompt Learning (MaPLe) (Khattak et al., 2023a), Prompting with Self-regulating Constraints (PromptSRC) (Khattak et al., 2023b).

Evaluation metrics. 4 standard metrics are used in our evaluation of open-vocabulary confidence calibration: Expected Calibration Error (ECE) (Guo et al., 2017), Maximum Calibration Error (MCE) (Guo et al., 2017), Adaptive Calibration Error (ACE) (Nixon et al., 2019) and Proximity-Informed Expected Calibration Error (Xiong et al., 2023).

Implementation details. For the main results, We use CLIP (ViT-B/16) (Radford et al., 2021) as the pre-trained VLM throughout our experiments and report results averaged over 3 runs. We fine-tune the model with 16 samples per class in a few-shot setting (Zhou et al., 2022a). For the compared tuning methods, we adopt them from the corresponding official implementation. For our proposed DAC, for simplicity, we set the number of nearest textual features K=5 throughout the paper to calculate TD scores. More details on datasets, pre-trained models, hyperparameters, and implementation details can be found in Appendix A.

5.2. Results

DAC improves open-vocabulary calibration in existing prompt tuning. Table 2 shows the average calibration error of all datasets. For the open-vocabulary calibration performance, our proposed DAC effectively mitigates the miscalibration in various tuning methods across all metrics. For instance, DAC significantly reduces the ECE from 13.84% to 7.00%, which makes CoOp more reliable in terms of predicted confidence. Furthermore, note that DAC is also applicable to *multi-modal tuning* (e.g., MaPLe and PromptSRC). This approach adjusts both vision and language branches and DAC consistently improves the open-vocabulary calibration performance of these methods. The results indicate that the textual distribution gap widely exists in various tuning methods and our proposed DAC can consistently boost open-vocabulary calibration upon existing tuning methods. Due to space constraints, we provide detailed calibration results in Appendix C.4.

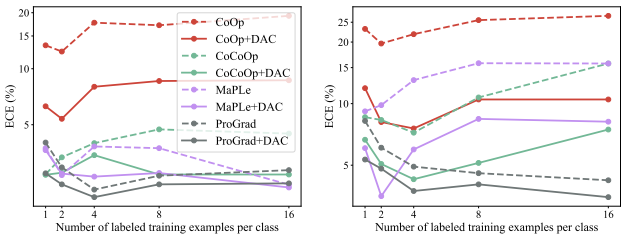
DAC significantly reduces calibration error, especially

Table 3. Calibration results of ECE (%) across different confidence levels. Δ shows the improvement achieved by DAC. **Bold** numbers denote the top-3 most significant improvements.

Method		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CoOp	Conf	0.00	19.24	18.86	15.87	20.42	33.28	30.30	37.84	41.60	18.57
	DAC	0.00	4.95	8.17	11.33	6.51	11.42	24.12	17.41	11.37	-0.94
	Δ	0.00	-14.29	-10.69	-4.54	-13.91	**-21.86**	-6.18	**-20.43**	**-30.23**	-19.51
MaPLe	Conf	0.00	0.00	-12.80	3.90	16.73	10.50	38.07	23.93	19.11	12.13
	+DAC	0.00	-3.62	-15.32	5.72	6.74	3.12	15.45	6.16	9.29	6.55
	Δ	0.00	-3.62	-2.52	1.82	**-10.99**	-7.38	**-22.62**	**-17.77**	-9.82	-5.58
ProGrad	Conf	0.00	-3.82	0.14	-0.10	4.29	6.31	3.48	8.11	1.23	4.86
	+DAC	0.00	-0.71	0.03	1.30	-1.32	0.40	-0.65	-0.04	0.44	-0.34
	Δ	0.00	3.11	-0.11	1.40	**-5.61**	**-5.91**	-4.13	**-8.15**	-0.79	-5.20

Table 4. Comparison results of ECE (%) using different visual backbones on Flower102 dataset. The smaller values are better.

	CoOp		CoCoOp		ProGrad	
Backbone	Conf	DAC	Conf	DAC	Conf	DAC
RN50	15.72	**8.03**	6.00	**4.88**	4.1	**3.39**
ViT-B-32	21.07	**11.72**	9.71	**6.57**	5.11	**4.36**
ViT-B-16	18.34	**10.19**	11.49	**7.74**	5.45	**5.04**



(a) UCF101 (b) DTD

Figure 6. Comparison results of ECE (%) using different shots. Miscalibration is a common issue and DAC can reduce it across different shots. The Y-axis is presented in an exponential form for a better view.

for high-confidence predictions. To verify where DAC helps the most over the baseline methods, we analyzed the calibration performance across various confidence levels. Specifically, we measure the gap between confidence and accuracy for each bin on the DTD dataset in Table 3. A negative gap indicates underconfident prediction, and vice versa. The difference between the baseline and our method is presented for easy comparison of improvements in each bin. The results show that DAC improves the calibration performance mainly in those confident predictions (the bins 0.6-0.9). In particular, our method can significantly alleviate the overconfidence issue. For example, in the case of CoOp, our method reduces the gap from 41.60 to 11.37, a direct improvement of 30.23, in the bin with confidence 0.9.

DAC regularizes the prediction with low TD score for better calibration. Since DAC adjusts the temperature according to the predicted class, we present a detailed visualization of class-wise results in 4. For a comprehensive evaluation, we also compare our results with Temperature Scaling (Guo et al., 2017). As the TD score decreases, the tuned model shows higher confidence in open-vocabulary classes, and TS cannot help and even exacerbates the issue. In particular, DAC addresses this overconfidence by reducing the logit scale for the prediction with low textual proximity, thereby ensuring more reliable predictions. We present the detailed results in the Appendix B.2.

DAC is robust to the number of nearest neighbors. We evaluate how K in our method affects the open-vocabulary calibration performance. Specifically, we present this ablation with ECE based on DTD and SUN397 datasets. We vary the number of textual neighbors $k = \{1, 2, 3, ..., 10\}$. As is shown in Figure 5, an increase in the number of neighbors beyond 0 leads to an evident reduction in ECE, and the performance starts to reach a point of saturation with the further addition of neighbors. It's noteworthy that DAC shows robustness to the choice of K, even if we set $K = 1$ yielding significant calibration improvements. This phenomenon indicates that it is beneficial to adjust the confidence level according to the textual distribution gap between base and new classes. For simplicity, we employ a moderate range of neighbors and use $K = 5$ throughout our experiments.

DAC is effective across various few-shot settings. We consider two options in few-shot learning: shots and visual backbone, where "shot" is the number of labeled samples per class in tuning. Unlike accuracy, which tends to increase as the number of shots increases (Zhou et al., 2022b), calibration errors are prevalent across diverse few-shot settings. We visualize the ECE performance across different shots in Figure 6 and demonstrate that our proposed DAC successfully mitigates this issue. Moreover, we show that DAC is effective across a variety of backbones in Table 4.

Table 5. Calibration of CoOp and FLYP using ECE (%) across 11 datasets. **Bold** numbers are significantly superior results.

	CoOp		FLYP	
Dataset	Conf	+DAC	Conf	+DAC
Caltech101	4.08	**3.17**	2.80	**2.01**
OxfordPets	1.83	1.82	0.73	0.51
StanfordCars	12.50	**5.16**	19.90	**15.62**
Flowers102	18.34	**10.19**	20.58	**19.09**
Food101	3.83	**1.78**	5.74	**4.13**
FGVCAircraft	28.44	**17.38**	36.52	**26.84**
SUN397	13.70	**4.05**	21.72	**19.09**
DTD	26.86	**10.48**	12.30	**6.49**
EuroSAT	12.71	**8.62**	5.32	5.30
UCF101	19.24	**8.67**	11.98	**8.18**
ImageNet	10.69	**5.67**	8.42	**7.19**
AVG	13.84	**7.00**	13.27	**10.40**

6. Discussion

How does DAC perform in full fine-tuning? To evaluate the flexibility of DAC, we provide additional results with a most recent fine-tuning approach named FLYP(Goyal et al., 2023), where both image and text encoders will be updated. Following the official implementation, we fine-tune the model for 10 epochs with a learning rate of 1e-5.

As is shown in Table 5, DAC successfully reduces the calibration error of models finetuned by FLYP on open-vocabulary classes, which is consistent with the results of prompt-tuning methods. Furthermore, we observe that fine-tuning methods like FLYP also tend to overfit the base classes on several datasets, making it more challenging to calibrate on open-vocabulary classes.

Visual proximity vs. Textual proximity. Proximity plays an important role in trustworthy machine learning (Jiang et al., 2018; Liu et al., 2023; Xiong et al., 2023; Yuksekgonul et al., 2023). Recent works mainly focus on the proximity estimation of the model input and largely mitigate the miscalibration in supervised learning. However, we show that textual proximity is more important than visual proximity in prompt tuning. To this end, we compare the calibration results and inference efficiency of our proposed DAC on ImageNet-1K dataset with Density-Ratio Calibration (Xiong et al., 2023).

We present the results in Table 6. For calibration performance, our proposed DAC outperforms DEN in the calibration of new classes, and we can achieve better calibration results by combining DAC with DEN. Additionally, we assess the efficiency of inference. In terms of computational cost, our approach maintains high efficiency with negligible extra computational burden compared to DEN and the baseline (Conf). Overall, our proposed DAC shows performance superiority and inference efficiency on calibration in new classes compared with methods based on image proximity.

Table 6. Calibration performance and inference time on ImageNet-1K. CoOp is used for prompt tuning. "base" denotes the performance in base classes. DAC shows performance superiority and inference efficiency on calibration in new classes.

	Conf	DAC	DEN	DAC+DEN
ECE_{base}	1.65	1.65	1.19	1.19
ECE_{new}	10.69	5.67	8.08	3.29
time (s)	0.12	0.13	15.88	15.91

Textual feature normalization is critical. In DAC, we normalize the features from the textual encoder for calculating TD scores. However, the significance of such normalization is not immediately evident. In this ablation, we contrast the performance of textual proximity-based calibration with and without feature normalization. As an illustration, we report the calibration results of tuned CLIP with CoOp.

Our results in Figure 7 show that employing normalized features can have a better estimation of class-wise proximity, which achieves better calibration performance than raw features. Previous works (Jiang et al., 2024; Sun et al., 2022; Xiong et al., 2023) demonstrated that feature norms may misestimate the K-nearest-neighbor-based score with Euclidean distance. Empirically, we verify that normalization plays a key role in our distance-based calibration.

Calibration with data from novel classes. Generally, it is impractical for a user to access the image-text pairs from the novel classes during tuning. As an analysis, we verify the feasibility of open-vocabulary calibration utilizing the knowledge from new classes for post-hoc calibration. To this end, We present the calibration results that the model will be tuned in base classes and calibrated using open-vocabulary image-text pairs, which utilize Temperature Scaling (TS) as the post-hoc calibration method.

As illustrated in Table 7, TS effectively calibrates the model using image-text pairs of the new classes. However, such calibration compromises the model's ability to deliver a reliable confidence level for the base classes. Given that TS modifies the logits scale with a single scalar, it is infeasible to calibrate the model on both base and new class. Overall, as a training-free calibration, our proposed DAC boosts the calibration performance according to textual proximity.

7. Related work

Prompt tuning. Thanks to the rich natural language supervision, pre-trained large vision-language models (Jia et al., 2021; Radford et al., 2021) can effectively comprehend vi-

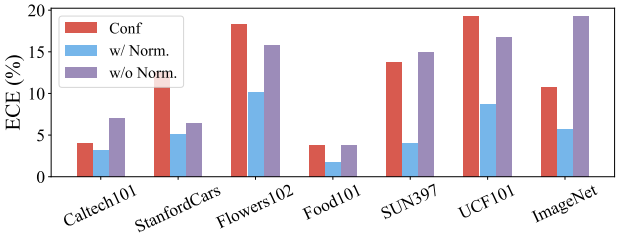


Figure 7. Ablation results of textual feature normalization with CoOp. We compare the effect of using normalization in the textual feature vs. without normalization.

sual concepts and apply them in downstream tasks (e.g. image classification (Radford et al., 2021; Zhou et al., 2022b; Lu et al., 2022; Naeem et al., 2023), image captioning (Hu et al., 2022; Wang et al., 2023b;a), visual question answering (Parelli et al., 2023)) in a zero-shot manner. Despite VLM's effectiveness in generalizing new visual concepts, the performance of zero-shot VLM still lags behind the fine-tuned performance on specific downstream tasks (Zhang et al., 2023). To further boost the downstream adaptation of pre-trained VLMs, many parameter-efficient tuning methods like prompt tuning (Zhou et al., 2022b;a; Yao et al., 2023; Zhu et al., 2023; Zang et al., 2022) have been proposed for high efficiency. Recently, several works have explored multimodel tuning (Khattak et al., 2023a;b; Shen et al., 2024), which adapts both vision and language branches of CLIP for better modality alignment. Despite the great success of V-L tuning, their effectiveness on safety-related evaluation is less explored. In this work, we first investigate the calibration performance for fine-tuned VLMs. We present that fine-tuned CLIP generally suffers from miscalibration, and our proposed DAC can effectively mitigate the issue by rectifying the logit scale based on textual information.

Confidence calibration. Confidence calibration has been widely studied to ensure that the confidence levels output by models accurately reflect their true performance. To achieve this, A common practice is to employ post-hoc calibration techniques after model training. Generally, post-hoc calibration can be split into two paradigms: *scaling-based* methods (Guo et al., 2017; Tomani et al., 2022; Yu et al., 2022; Zhang et al., 2020; Joy et al., 2023; Xiong et al., 2023) and *bin-based* methods (Zadrozny & Elkan, 2001; 2002; Zhang et al., 2020; Roelofs et al., 2022). For scaling-based calibration, a representative method is temperature scaling (Guo et al., 2017), which learns a single scalar for rescaling the logit. Mix-n-Match (Zhang et al., 2020) leverages ensemble and composition techniques to achieve data efficiency and maintain accuracy in confidence estimates. ATS (Joy et al., 2023) modifies the predicted confidence by per-data-point adaptive temperature. For binning-based calibration, the samples are segmented into various bins according to confidence levels, with each bin undergoing calibration individually. Recently, several works have explored calibration in VLMs (Oh et al., 2023; LeVine et al., 2023). In this work, we focus on the open-vocabulary calibration of fine-tuned VLMs. We first show that established post-hoc calibration on base classes can not transfer to new classes. To this end, our DAC fixes the logit scale of prediction via textual deviation-informed score in new classes.

Table 7. Average performance of ECE (%) on 11 datasets. CoOp is used to tune the pre-trained CLIP. TS_{new} / TS_{base} denotes the calibration on data of base/new classes.

Class	Conf	TS_{base}	TS_{new}	DAC
Base	2.95	2.57	13.05	2.95
New	13.84	12.95	7.56	7.00

8. Conclusion

In this paper, we introduce Distance-Aware Calibration (DAC), a simple fix to temperature scaling in fine-tuned VLMs, that enhances existing prompt tuning methods in open-vocabulary calibration. By scaling the temperature value according to the distance between the predicted text label and base classes, fine-tuned CLIP tends to give a more reliable confidence level for new classes. Extensive experiments show that DAC can improve the calibration in new classes after prompt tuning while maintaining the performance in base classes. Our proposed DAC can be easily incorporated with existing scaling-based calibration without compromising the inference speed. We hope our findings inspire further research on the calibration of VLMs.

Acknowledgements

This research is supported by the Shenzhen Fundamental Research Program (Grant No. JCYJ20230807091809020). Bob Zhang is supported by the Science and Technology Development Fund (FDCT), Macao, SAR, under Grant 0028/2023/RIA1. Guoqing Wang is supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Key R&D Program of Zhejiang under grant 2024SSYS0091. We gratefully acknowledge the support of the Center for Computational Science and Engineering at the Southern University of Science and Technology for our research.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition workshop*, pp. 178–178. IEEE, 2004.

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

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17980–17989, 2022.

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

Jiang, H., Kim, B., Guan, M., and Gupta, M. To trust or not to trust a classifier. *Advances in Neural Information Processing Systems*, 31, 2018.

Jiang, W., Cheng, H., Chen, M., Wang, C., and Wei, H. DOS: Diverse outlier sampling for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Joy, T., Pinto, F., Lim, S.-N., Torr, P. H., and Dokania, P. K. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14919–14926, 2023.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023a.

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M.-H., and Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

LeVine, W., Pikus, B., Raj, P., and Gil, F. A. Enabling calibration in the zero-shot inference of large vision-language models. *ICLR workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Liu, J. Z., Padhy, S., Ren, J., Lin, Z., Wen, Y., Jerfel, G., Nado, Z., Snoek, J., Tran, D., and Lakshminarayanan, B. A simple approach to improve single-model deep uncertainty via distance-awareness. *Journal of Machine Learning Research*, 24(42):1–63, 2023.

Lu, Y., Liu, J., Zhang, Y., Liu, Y., and Tian, X. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5215, 2022.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Naeem, M. F., Khan, M. G. Z. A., Xian, Y., Afzal, M. Z., Stricker, D., Van Gool, L., and Tombari, F. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15169–15179, 2023.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *Computer Vision and Pattern Recognition workshops*, volume 2, 2019.

Oh, C., Kim, M., Lim, H., Park, J., Jeong, E., Cheng, Z.-Q., and Song, K. Towards calibrated robust fine-tuning of vision-language models. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., and Hofmann, T. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5606–5611, 2023.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505. IEEE, 2012.

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*, pp. 8748–8763. PMLR, 2021.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, 2022.

Shen, S., Yang, S., Zhang, T., Zhai, B., Gonzalez, J. E., Keutzer, K., and Darrell, T. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5656–5667, 2024.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Tomani, C., Cremers, D., and Buettner, F. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *European Conference on Computer Vision*, pp. 555–569. Springer, 2022.

Tu, W., Deng, W., and Gedeon, T. A closer look at the robustness of contrastive language-image pre-training (clip). In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wang, N., Xie, J., Wu, J., Jia, M., and Li, L. Controllable image captioning via prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2617–2625, 2023a.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023b.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492. IEEE, 2010.

Xiong, M., Deng, A., Koh, P. W., Wu, J., Li, S., Xu, J., and Hooi, B. Proximity-informed calibration for deep neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yao, H., Zhang, R., and Xu, C. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6757–6767, 2023.

Yu, Y., Bates, S., Ma, Y., and Jordan, M. Robust calibration with multi-domain temperature scaling. *Advances in Neural Information Processing Systems*, 35:27510–27523, 2022.

Yuksekgonul, M., Zhang, L., Zou, J., and Guestrin, C. Beyond confidence: Reliable models should also consider atypicality. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, volume 1, pp. 609–616, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 694–699, 2002.

Zang, Y., Li, W., Zhou, K., Huang, C., and Loy, C. C. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.

Zhang, J., Kailkhura, B., and Han, T. Y.-J. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pp. 11117–11128. PMLR, 2020.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

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

Zhu, B., Niu, Y., Han, Y., Wu, Y., and Zhang, H. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023.

A. Experimental Setting

A.1. Dataset

Following the base-to-new generalization setting in CoCoOp (Zhou et al., 2022a), we split the dataset into two subsets *base* and *new*. The labels for these subsets do not overlap. As is shown in Table 8, we split the dataset into three folds through our experiments. The pre-trained CLIP will be first tuned on base classes. For former calibration methods like (Guo et al., 2017), they need a calibration set to align the predicted probabilities more closely with the actual likelihoods of the outcomes. Since we can not access new (open-vocabulary) classes, we sample a calibration set from the datasets of base classes which does not overlap with the training set. Finally, we evaluate the model on both base classes and new classes. In our experiments, we mainly focus on the calibration in new classes.

Table 8. The split strategy of downstream datasets. We use data from base classes to tune and calibrate the pre-trained CLIP.

Dataset	Train†	Calibration†	Test$^{\psi}$
Caltech101	800	200	916
OxfordPets	304	76	1,788
StanfordCars	1,568	392	4039
Flowers102	816	204	1,410
Food101	816	204	15,000
FGVCAircraft	800	200	1,667
SUN397	3,184	796	9,900
DTD	384	96	828
EuroSAT	80	20	3,900
UCF101	816	204	1,849
ImageNet	8,000	25,000	25,000

$\dagger$: Base class, ψ: New class.

A.2. Compared methods

We use CLIP (ViT-B/16) (Radford et al., 2021) as the pre-trained VLM throughout our experiments and report results averaged over 3 runs. We fine-tune the model with 16 samples per class in a few-shot setting (Zhou et al., 2022a). Following the official implementation, We list the general hyperparameters in Table 9. Here, we briefly introduce the corresponding exclusive hyperparameters of each VLM tuning method. All the methods are adopted from their official implementation. For CoOp and CoCoOp, they do not contain other hyperparameters. For ProDA, we set $\lambda = 0.1$. For KgCoOp, we set $\lambda = 8.0$. For MaPLe, we set prompt depth J to 0 and the language and vision prompt lengths to 2. For ProGrad, we set $\lambda = 0.8$. For PromptSRC, we set deep prompting with $V = T = 4$. $\lambda_1 = 10$ and $\lambda_2 = 25$ are used in weight loss. For textual diversity, we use a total of $N = 60$ standard prompt templates. We reproduce their methods in our framework based on the standard codebase [2].

Table 9. Hyperparameters for VLM tuning methods. "BS" denotes the batch size. "LR" denotes the learning rate. "CTX" is the context length of the learnable prompt.

Methods	Epochs	BS	LR	CTX
CoOp	200	32	0.002	16
CoCoOp	10	1	0.002	4
ProDA	100	4	0.001	16
KgCoOp	200	32	0.002	16
MaPLe	5	4	0.0026	2
ProGrad	100	32	0.002	16
PromptSRC	50	4	0.0025	4

[2]https://github.com/KaiyangZhou/CoOp

B. Detailed class-wise results of open-vocabulary calibration

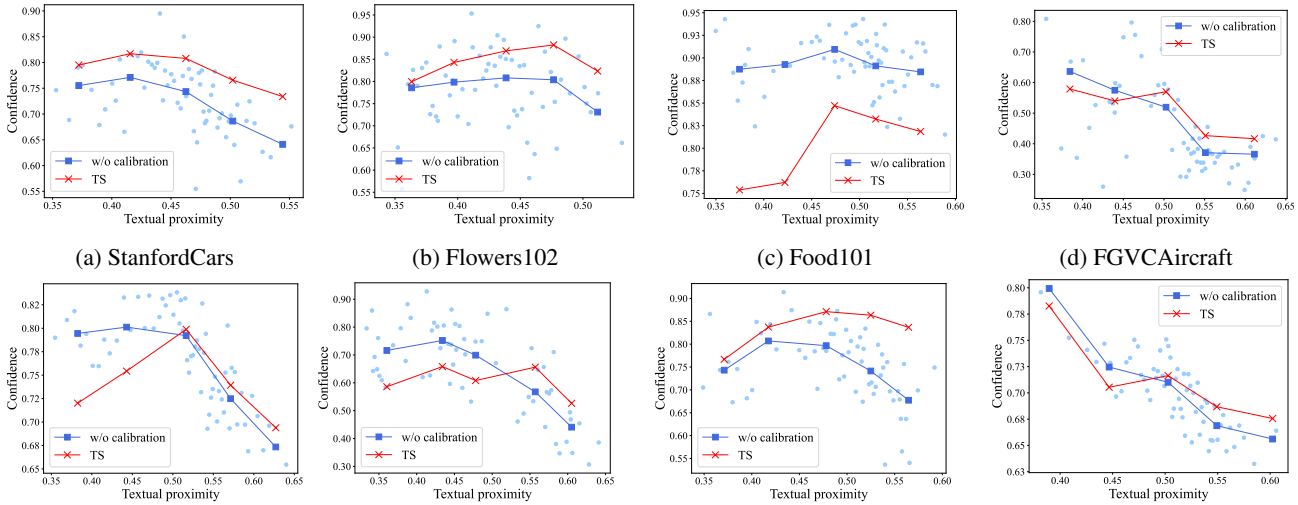
We present the miscalibration in tuned CLIP and the effectiveness of our proposed DAC via class-wise visualization.

Experiment setup. For tuning methods, we use Context Optimization (CoOp) (Zhou et al., 2022b), Conditional Context Optimization (CoCoOp) (Zhou et al., 2022a), Knowledge-guided Context Optimization (KgCoOp) (Yao et al., 2023), Prompt-aligned Gradient (ProGrad) (Zhu et al., 2023), Multi-modal Prompt Learning (MaPLe) (Khattak et al., 2023a). We calculate their class-wise results according to the *prediction* (pseudo label). For the datasets, we present the results of various image recognition datasets including ImageNet (Deng et al., 2009), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and FGVCAircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), UCF101 (Soomro et al., 2012) and DTD (Cimpoi et al., 2014). For a better view, we visualize bin-base results by broken lines, and the origin results are given by scatters.

B.1. Textual Proximity

To showcase that the deviation degree in the textual gap is crucial for open-vocabulary calibration, we first introduce a distance-based metric named *proximity* (see Section. 4.2) to quantify the deviation degree of textual features. We present the correlation between the deviation degree of textual feature and confidence / ECE in Table 8 and 9. Here we use ECE^* with a positive (negative) value denoting overconfidence (underconfidence) for a better view. The results show that Lower proximity correlates with higher confidence and ECE.

B.1.1. Confidence

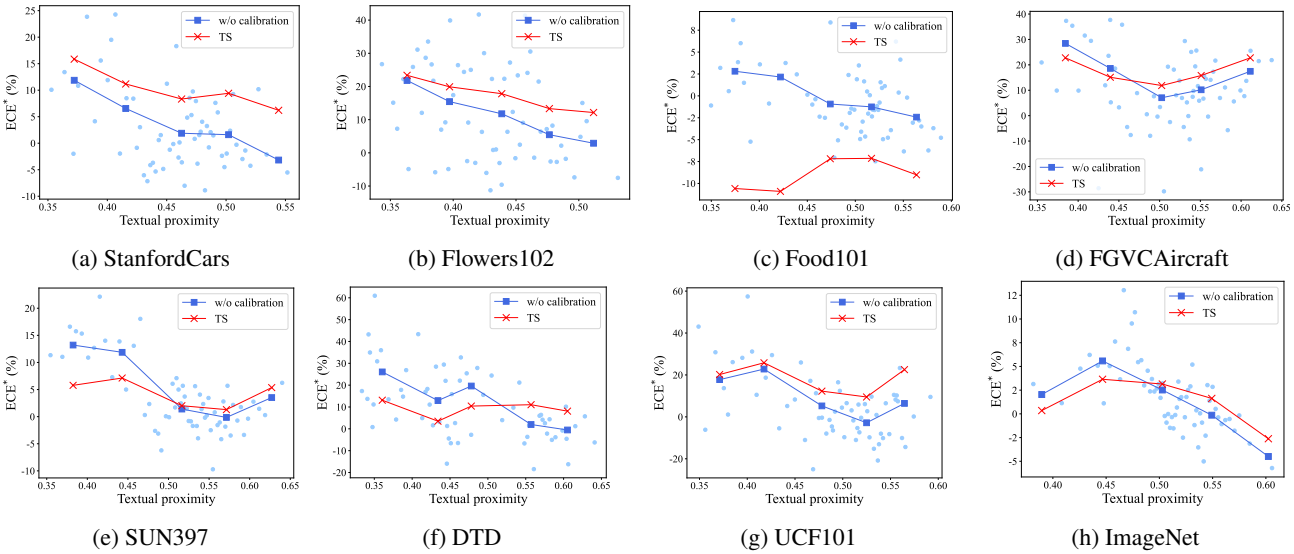


(a) StanfordCars (b) Flowers102 (c) Food101 (d) FGVCAircraft

(e) SUN397 (f) DTD (g) UCF101 (h) ImageNet

Figure 8. Class-wise confidence on downstream dataset after tuning.

B.1.2. ECE*



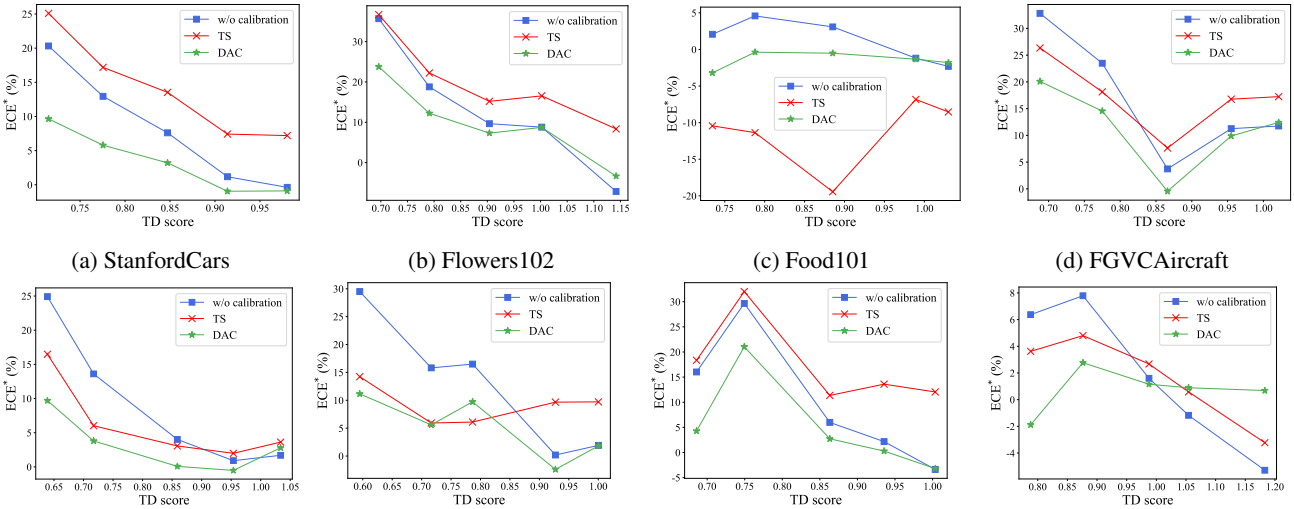
(a) StanfordCars (b) Flowers102 (c) Food101 (d) FGVCAircraft

(e) SUN397 (f) DTD (g) UCF101 (h) ImageNet

Figure 9. Class-wise performance on downstream dataset after tuning.

B.2. TD score

To verify the effectiveness of our proposed method, we present the class-wise results with DAC. As the TD score decreases in Figure 10, the tuned model shows higher confidence in open-vocabulary classes and TS cannot help the issue. In particular, DAC addresses this overconfidence by reducing the logit scale for the prediction with low textual proximity, thereby ensuring more reliable predictions.



(a) StanfordCars (b) Flowers102 (c) Food101 (d) FGVCAircraft

(e) SUN397 (f) DTD (g) UCF101 (h) ImageNet

Figure 10. Class-wise performance on downstream dataset after tuning.

C. Detailed Experimental Results

In this section, we illustrate the experimental results in detail.

C.1. The calibration performance of tuned VLMs

To showcase the miscalibration in tuned VLM, we fine-tune the pre-trained CLIP with 7 tuning methods on 11 downstream datasets and use ECE as the calibration metric. We report the detailed calibration results in this section. The datasets and compared methods are listed in Section 5.1.

As is shown in Table C.1, for the calibration of the base class, CoOp shows the best calibration results in base classes. We can also observe that distillation-based tuning (KgCoOp, ProGrad and PromptSRC) restrains the model's confidence in base classes. For the calibration of the new class, distillation-based tuning utilizes zero-shot CLIP as the teacher model to improve the generalization ability, which generates relatively reliable predictions in unseen classes as expected.

More importantly, we observe that tuned VLMs tend to be underconfident in base classes while prone to overconfident in new classes. This phenomenon motivates us to further investigate the calibration of fine-tuned VLM.

Table 10. Expected Calibration Error (ECE) on diverse downstream datasets using various tuning methods for CLIP-ViT-B/16. The calibration performance is averaged across three variants. "base" / "new" denotes, the performance on seen/unseen classes.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	6.49	2.25	3.74	3.11	1.57	3.03	1.59	4.53	8.35	3.24	1.51	3.58
CoOp	0.75	0.97	2.76	2.26	2.64	5.56	4.42	6.47	1.33	3.64	1.65	2.95
CoCoOp	1.72	2.29	7.01	7.41	1.02	3.66	1.84	2.85	8.21	2.17	2.58	3.71
ProDA	1.85	3.48	11.16	5.03	1.95	8.24	3.15	2.38	9.60	4.54	4.82	5.11
KgCoOp	2.32	2.96	11.36	9.97	1.35	5.88	4.68	7.97	10.84	4.17	2.63	5.83
MaPLe	1.12	1.92	6.65	3.66	0.74	3.06	1.22	3.12	4.35	2.29	2.19	2.76
ProGrad	2.76	2.71	9.62	8.92	1.39	5.68	3.68	10.99	10.07	5.04	2.49	5.76
PromptSRC	2.24	2.39	8.05	4.59	0.96	4.68	2.23	2.77	8.95	2.52	2.01	3.76

(a) Base

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.60	3.42	3.31	4.91	1.83	6.55	3.48	6.86	9.12	5.52	2.09	4.43
CoOp	4.08	1.83	12.50	18.34	3.83	28.44	13.70	26.86	12.71	19.24	10.69	13.84
CoCoOp	3.89	2.19	2.31	11.49	1.70	11.26	2.20	15.71	12.32	4.48	1.65	6.29
ProDA	4.61	6.60	3.02	6.67	1.51	5.36	1.49	8.22	4.84	3.11	1.59	4.27
KgCoOp	1.96	3.16	2.81	5.99	1.98	12.53	1.28	7.03	6.82	2.64	1.8	4.36
MaPLe	1.33	2.51	2.02	13.43	1.26	10.38	3.01	15.71	9.06	2.38	2.39	5.77
ProGrad	1.91	2.98	2.88	5.45	2.04	13.18	1.75	4.22	7.40	2.85	1.81	4.22
PromptSRC	1.62	3.08	2.16	5.77	1.60	9.79	0.77	5.79	7.35	2.64	1.71	3.84

(b) New

C.2. Calibrated performance for tuned VLM

We utilize commonly used post-hoc confidence calibration for tuned VLMs and expect to calibrate them well on both base and new classes. We can observe that established calibration can remedy miscalibration in base classes. However, such efficacy can not transfer to new classes.

C.2.1. BASE CLASS

Post-hoc calibration can remedy miscalibration in base classes. In the latest works (Tu et al., 2023; Oh et al., 2023), scaling-based methods like TS have been successfully adopted in VLM calibration, and their findings reveal that this calibration can handle the miscalibration well in the close-world setting. As a complement, we observe that both scaling-based (e.g., DEN) and bin-based (e.g., MIR) calibration can notably reduce miscalibration in base classes.

Post-hoc calibration on base classes can not transfer to new classes. For instance, scaling-based methods like TS scale the logits by a single scalar temperature value. If tuned VLMs show underconfidence in base classes, TS sharpens the logit scale before the softmax function and makes the prediction more confident. As a cost, the confidence distribution of tuned VLM becomes sharper and causes the VLMs to display exaggerated overconfidence in new classes. for bin-based methods, they need the probabilities from base classes as the input, which are incompatible with zero-shot prediction.

C.2.2. NEW CLASS

Table 12. Expected Calibration Error (ECE) on new classes in diverse downstream datasets using various tuning methods for CLIP-ViT-B/16, with calibration performance averaged across three variants.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.60	3.42	3.31	4.91	1.83	6.55	3.48	6.86	9.12	5.52	2.09	4.43
CoOp	4.08	1.83	12.50	18.34	3.83	28.44	13.70	26.86	12.71	19.24	10.69	13.84
CoCoOp	3.89	2.19	2.31	11.49	1.70	11.26	2.20	15.71	12.32	4.48	1.65	6.29
ProDA	4.61	6.60	3.02	6.67	1.51	5.36	1.49	8.22	4.84	3.11	1.59	4.27
KgCoOp	1.96	3.16	2.81	5.99	1.98	12.53	1.28	7.03	6.82	2.64	1.8	4.36
MaPLe	1.33	2.51	2.02	13.43	1.26	10.38	3.01	15.71	9.06	2.38	2.39	5.77
ProGrad	1.91	2.98	2.88	5.45	2.04	13.18	1.75	4.22	7.40	2.85	1.81	4.22
PromptSRC	1.62	3.08	2.16	5.77	1.60	9.79	0.77	5.79	7.35	2.64	1.71	3.84

(a) w/o calibration

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	3.00	1.54	1.68	5.14	8.29	7.60	3.34	10.86	11.07	1.62	3.1	5.20
PromptSRC	2.6	1.87	6.72	13.45	4.55	12.91	1.3	5.8	10.05	12.72	2.48	6.77
ProGrad	2.66	1.48	10.14	18.50	3.16	21.31	4.10	14.75	16.70	16.54	3.43	10.25
ProDA	5.66	3.99	7.26	14.38	5.12	15.17	1.11	6.49	6.42	15.20	3.9	7.70
MaPLe	1.61	1.86	4.89	18.02	3.89	13.15	2.12	6.68	11.04	7.45	2.76	6.68
KgCoOp	3.04	1.36	5.72	16.21	3.54	21.33	2.55	14.24	10.42	12.37	2.57	8.49
CoOp	4.43	2.28	15.78	25.16	6.01	24.21	7.44	10.67	15.71	23.47	7.33	12.95
CoCoOp	4.55	1.11	6.38	19.95	3.88	17.50	2.13	12.36	19.01	11.78	2.32	9.18

(b) Temperature Scaling

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	3.61	1.13	4.01	7.23	2.62	6.07	3.54	10.57	9.02	5.16	2.19	5.01
PromptSRC	3.86	1.42	6.17	12.96	1.68	16.34	2.70	10.35	24.47	9.42	3.04	8.40
ProGrad	3.90	0.96	8.36	18.32	1.40	22.28	4.62	14.08	69.06	13.13	3.77	14.53
ProDA	6.40	2.82	6.12	14.70	2.73	12.93	2.72	10.51	13.39	11.67	3.86	7.99
MaPLe	2.01	1.82	4.86	18.13	1.41	13.96	4.16	16.19	26.05	4.93	4.22	8.89
KgCoOp	4.43	1.20	5.88	14.48	1.54	21.33	3.94	16.05	16.16	10.86	2.76	8.97
CoOp	5.89	4.95	16.35	24.31	4.81	26.77	9.57	23.31	16.90	26.56	8.08	15.23
CoCoOp	5.32	1.62	5.33	19.03	2.02	17.18	4.35	16.11	9.47	9.75	2.18	8.40

(c) Density-Ratio Calibration

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.60	3.42	3.31	4.91	1.83	6.55	3.48	6.86	9.12	5.52	2.09	4.43
CoOp	3.17	1.82	5.16	10.19	1.78	17.38	4.05	10.48	8.62	8.67	5.67	7.00
CoCoOp	3.65	2.43	2.21	7.74	1.64	9.03	1.09	7.47	13.49	2.70	1.61	4.82
ProDA	4.87	4.72	3.28	6.32	0.70	7.40	1.06	5.68	3.33	4.14	2.42	3.99
KgCoOp	1.84	3.11	3.12	5.90	1.94	11.78	1.67	7.09	6.59	2.69	1.83	4.32
MaPLe	1.26	2.51	2.75	11.28	1.50	9.06	1.22	8.16	8.55	2.30	2.11	4.61
ProGrad	1.97	3.31	2.29	5.04	1.85	10.46	1.32	3.49	6.90	2.42	2.11	3.74
PromptSRC	1.58	2.98	2.39	5.03	1.55	8.55	0.79	5.50	7.24	2.46	1.84	3.63

(d) Distance-Aware Calibration

Table 11. Expected Calibration Error (ECE) on base classes in diverse downstream datasets using various tuning methods for CLIP-ViT-B/16, with calibration performance averaged across three variants.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	6.49	2.25	3.74	3.11	1.57	3.03	1.59	4.53	8.35	3.24	1.51	3.58
CoOp	0.75	0.97	2.76	2.26	2.64	5.56	4.42	6.47	1.33	3.64	1.65	2.95
CoCoOp	1.72	2.29	7.01	7.41	1.02	3.66	1.84	2.85	8.21	2.17	2.58	3.71
ProDA	1.85	3.48	11.16	5.03	1.95	8.24	3.15	2.38	9.60	4.54	4.82	5.11
KgCoOp	2.32	2.96	11.36	9.97	1.35	5.88	4.68	7.97	10.84	4.17	2.63	5.83
MaPLe	1.12	1.92	6.65	3.66	0.74	3.06	1.22	3.12	4.35	2.29	2.19	2.76
ProGrad	2.76	2.71	9.62	8.92	1.39	5.68	3.68	10.99	10.07	5.04	2.49	5.76
PromptSRC	2.24	2.39	8.05	4.59	0.96	4.68	2.23	2.77	8.95	2.52	2.01	3.76

(a) w/ calibation

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.33	0.82	1.82	2.73	8.29	2.62	1.69	4.54	7.84	7.86	1.79	3.76
CoOp	0.66	1.12	1.71	0.99	4.81	2.83	1.32	5.72	1.80	5.83	1.5	2.57
CoCoOp	0.96	1.14	1.45	1.15	4.00	3.07	1.37	4.74	10.90	5.43	1.56	3.25
ProDA	0.79	1.61	2.78	1.43	4.76	3.51	1.97	2.96	6.20	5.56	1.94	3.05
KgCoOp	0.62	0.67	2.26	1.87	3.76	2.62	1.65	3.88	6.43	5.60	1.48	2.80
MaPLe	0.51	1.15	1.89	1.77	4.40	2.71	1.21	5.98	1.93	5.22	1.91	2.61
ProGrad	0.93	1.18	2.07	1.30	3.26	2.64	1.21	3.41	3.34	6.47	1.23	2.46
PromptSRC	0.75	1.1	2.04	1.15	4.63	2.86	1.5	2.9	3.82	5.72	1.30	2.52

(b) Temperature Scaling

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	2.72	3.60	2.43	4.44	2.28	3.37	1.04	2.72	7.40	5.88	0.44	3.30
CoOp	1.09	5.77	2.20	1.01	3.18	3.85	1.59	1.09	1.50	6.67	1.19	2.65
CoCoOp	0.96	1.25	2.86	1.85	2.18	4.81	1.19	0.96	5.64	4.34	0.67	2.43
ProDA	0.70	2.13	2.63	1.40	1.74	5.58	0.98	0.70	3.58	5.56	0.73	2.34
KgCoOp	0.93	1.61	2.94	2.40	1.80	5.42	1.40	0.93	3.29	4.66	0.58	2.36
MaPLe	0.85	3.17	1.61	1.22	1.69	4.56	1.59	0.85	5.73	3.68	0.66	2.33
ProGrad	0.74	1.98	2.97	0.98	1.34	5.40	1.42	0.74	26.48	5.29	0.72	4.37
PromptSRC	0.84	0.65	2.05	0.79	1.77	4.27	1.17	0.84	8.71	4.27	0.80	2.38

(c) Density-Ratio Calibration

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	2.20	1.54	8.81	8.87	2.87	1.70	5.18	10.93	9.27	7.37	4.78	5.77
CoOp	0.89	1.39	7.92	1.21	3.42	11.44	5.26	7.79	1.35	7.21	3.95	4.71
CoCoOp	0.99	0.79	9.94	2.30	2.08	7.80	6.07	7.07	3.39	7.89	4.7	4.82
ProDA	0.87	1.46	8.74	1.98	2.28	7.27	4.81	7.54	3.23	8.03	4.23	4.59
KgCoOp	0.72	0.51	9.03	1.74	2.45	7.27	5.46	8.33	2.79	8.66	4.95	4.72
MaPLe	0.82	1.03	10.76	1.39	2.64	8.68	5.21	4.47	2.16	8.14	5.02	4.57
ProGrad	1.00	1.36	7.14	1.84	3.15	10.09	5.32	7.87	2.96	8.35	4.98	4.91
PromptSRC	0.52	1.10	7.48	1.59	2.22	10.28	4.38	5.26	3.77	5.83	4.73	4.29

(d) Histogram Binning

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	3.07	6.97	9.28	4.99	6.46	8.96	7.76	7.34	8.56	11.71	2.08	7.02
CoOp	5.23	4.70	11.85	2.00	8.46	12.23	9.36	8.33	8.17	15.68	1.16	7.92
CoCoOp	5.05	4.96	9.45	3.70	10.17	12.20	8.14	9.70	9.93	13.77	1.81	8.08
ProDA	3.98	4.20	8.69	3.03	8.26	9.98	7.26	6.48	11.88	16.61	2.09	7.50
KgCoOp	1.32	5.37	9.15	2.72	8.91	10.88	7.57	6.52	9.53	15.45	1.79	7.20
MaPLe	5.43	5.63	11.03	1.49	9.29	11.54	8.22	5.07	6.29	16.70	1.49	7.47
ProGrad	2.47	5.50	8.72	3.10	9.72	12.41	7.25	8.17	11.34	16.88	1.77	7.94
PromptSRC	3.56	4.23	10.09	1.86	10.11	12.40	8.18	7.28	12.76	10.25	1.53	7.48

(e) Isotonic Regression

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	3.54	1.07	3.19	2.43	2.81	1.97	2.03	3.91	2.86	4.26	0.31	2.58
CoOp	0.54	1.23	2.69	1.10	1.36	3.60	2.17	2.54	1.05	5.09	0.19	1.96
CoCoOp	0.71	0.78	2.71	2.36	1.94	3.26	2.49	4.74	6.61	3.32	0.48	2.67
ProDA	0.69	0.79	4.13	1.54	2.22	5.42	1.67	3.89	5.39	3.88	0.82	2.77
KgCoOp	0.84	1.13	4.60	2.20	1.40	3.35	2.13	4.49	3.34	3.59	0.48	2.50
MaPLe	0.55	1.72	2.97	1.36	1.64	4.19	1.62	4.44	1.40	2.95	0.33	2.11
ProGrad	0.81	1.98	3.47	1.56	1.77	3.33	2.10	3.45	1.41	4.03	0.39	2.21
PromptSRC	0.91	0.85	3.20	0.82	1.29	4.54	1.76	3.11	5.06	2.74	0.27	2.23

(f) Multi-Isotonic Regression

C.3. Accuracy & Confidence of tuned VLMs

In this section, we analyze the miscalibration from the classification performance. Since open-vocabulary classes are not explicitly seen during the tuning, some might believe that this miscalibration arises from rapid accuracy decreasing and leads to a larger Expected Calibration Error (ECE). This hypothesis leads to an open question: "*Should we blame accuracy degeneration for miscalibration?*"

We report the average performance on 11 datasets in Table 13 and 14. Although the classification performance on new classes may be comparable or lower, the average predictive confidence of tuned CLIP is still relatively higher than zero-shot CLIP, which results in a larger ECE. Moreover, we take a close look at the confidence level of MaPLe in new classes. Although the accuracy is comparable with zero-shot CLIP, we can find that the confidence level is greatly improved, which indicates the overconfidence issue.

C.3.1. Accuracy

Table 13. Accuracy comparison of existing prompt tuning in the base-to-new generalization setting.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	97.20	91.30	63.60	71.80	90.10	27.70	69.40	53.00	57.00	71.00	72.40	69.50
CoOp	98.23	93.57	80.07	97.90	87.83	42.57	80.60	80.80	92.63	84.77	75.60	83.14
CoCoOp	97.73	95.07	70.77	94.73	90.57	35.77	79.47	77.00	87.27	82.40	76.03	80.62
ProDA	98.53	96.13	76.43	97.93	90.77	39.77	82.37	82.80	90.47	85.43	77.57	83.47
KgCoOp	97.73	94.93	75.33	96.40	90.57	38.53	80.87	81.23	89.57	84.33	76.00	82.32
MaPLe	97.93	95.27	72.80	96.43	90.60	35.50	80.93	80.87	93.57	82.90	77.00	82.16
ProGrad	98.30	94.47	78.77	96.23	90.37	40.43	81.33	76.43	90.40	84.70	76.80	82.57
PromptSRC	98.27	95.57	80.30	98.27	90.63	44.37	83.10	83.80	92.70	87.73	77.83	84.78

(a) Base

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	94.10	97.10	75.00	77.50	91.10	35.90	75.50	60.60	63.80	78.60	68.10	74.30
CoOp	91.00	91.33	58.53	57.10	84.30	22.93	63.70	44.37	65.30	57.30	59.07	63.18
CoCoOp	93.13	97.70	72.70	69.80	91.33	33.47	76.60	53.63	62.37	74.23	70.63	72.33
ProDA	92.47	97.37	74.23	75.90	89.10	34.87	78.00	58.60	66.57	72.60	71.40	73.74
KgCoOp	94.20	97.40	74.17	72.83	91.60	30.40	75.33	50.10	66.67	74.70	69.73	72.47
MaPLe	95.30	97.53	75.07	71.50	91.63	36.20	77.10	57.83	67.40	78.67	70.60	74.44
ProGrad	93.73	97.47	69.40	71.80	90.10	29.37	72.40	51.57	60.57	71.83	67.33	70.51
PromptSRC	94.03	97.37	75.33	76.57	91.47	37.53	78.63	61.83	70.73	78.13	70.23	75.62

(b) New

C.3.2. Confidence

Table 14. Confidence comparison of existing prompt tuning in the base-to-new generalization setting.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	0.91	0.89	0.60	0.73	0.88	0.27	0.70	0.52	0.64	0.72	0.72	0.69
CoOp	0.98	0.93	0.77	0.96	0.90	0.48	0.85	0.87	0.91	0.89	0.77	0.85
CoCoOp	0.97	0.93	0.64	0.88	0.90	0.32	0.78	0.76	0.79	0.81	0.73	0.77
ProDA	0.97	0.93	0.65	0.93	0.89	0.32	0.79	0.82	0.81	0.82	0.73	0.79
KgCoOp	0.96	0.92	0.64	0.86	0.89	0.33	0.76	0.73	0.79	0.80	0.73	0.76
MaPLe	0.97	0.94	0.66	0.93	0.90	0.35	0.81	0.83	0.89	0.83	0.75	0.81
ProGrad	0.96	0.92	0.69	0.87	0.89	0.35	0.78	0.65	0.80	0.80	0.74	0.77
PromptSRC	0.96	0.93	0.72	0.94	0.90	0.40	0.81	0.82	0.84	0.86	0.76	0.81

(a) Base

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	0.95	0.94	0.72	0.79	0.89	0.41	0.72	0.52	0.62	0.73	0.68	0.72
CoOp	0.95	0.92	0.71	0.75	0.88	0.51	0.78	0.71	0.75	0.77	0.70	0.77
CoCoOp	0.96	0.95	0.74	0.81	0.90	0.45	0.78	0.70	0.71	0.78	0.70	0.77
ProDA	0.97	0.91	0.74	0.83	0.88	0.40	0.77	0.66	0.63	0.74	0.71	0.75
KgCoOp	0.95	0.94	0.72	0.78	0.90	0.43	0.74	0.56	0.69	0.74	0.69	0.74
MaPLe	0.95	0.95	0.75	0.85	0.90	0.46	0.80	0.73	0.69	0.80	0.73	0.78
ProGrad	0.94	0.95	0.72	0.77	0.88	0.42	0.74	0.53	0.68	0.73	0.69	0.73
PromptSRC	0.95	0.94	0.76	0.82	0.90	0.47	0.78	0.66	0.69	0.80	0.71	0.77

(b) New

C.4. Detailed results of main experiment

In this section, We present the detailed results on the calibration of new classes to verify that DAC can open-vocabulary calibration in existing prompt tuning. For comprehensive evaluation, 4 standard metrics are used in our evaluation of open-vocabulary confidence calibration: Expected Calibration Error (ECE) (Guo et al., 2017), Maximum Calibration Error (MCE) (Guo et al., 2017), Adaptive Calibration Error (ACE) (Nixon et al., 2019) and Proximity-Informed Expected Calibration Error (Xiong et al., 2023).

Table 15. Expected Calibration Error (ECE) on new classes in downstream datasets using various tuning methods for CLIP-ViT-B/16.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.60	3.42	3.31	4.91	1.83	6.55	3.48	6.86	9.12	5.52	2.09	4.43
CoOp	4.08	1.83	12.50	18.34	3.83	28.44	13.70	26.86	12.71	19.24	10.69	13.84
CoCoOp	3.89	2.19	2.31	11.49	1.70	11.26	2.20	15.71	12.32	4.48	1.65	6.29
ProDA	4.61	6.60	3.02	6.67	1.51	5.36	1.49	8.22	4.84	3.11	1.59	4.27
KgCoOp	1.96	3.16	2.81	5.99	1.98	12.53	1.28	7.03	6.82	2.64	1.8	4.36
MaPLe	1.33	2.51	2.02	13.43	1.26	10.38	3.01	15.71	9.06	2.38	2.39	5.77
ProGrad	1.91	2.98	2.88	5.45	2.04	13.18	1.75	4.22	7.40	2.85	1.81	4.22
PromptSRC	1.62	3.08	2.16	5.77	1.60	9.79	0.77	5.79	7.35	2.64	1.71	3.84

(a) w/o calibration

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.60	3.42	3.31	4.91	1.83	6.55	3.48	6.86	9.12	5.52	2.09	4.43
CoOp	3.17	1.82	5.16	10.19	1.78	17.38	4.05	10.48	8.62	8.67	5.67	7.00
CoCoOp	3.65	2.43	2.21	7.74	1.64	9.03	1.09	7.47	13.49	2.70	1.61	4.82
ProDA	4.87	4.72	3.28	6.32	0.70	7.40	1.06	5.68	3.33	4.14	2.42	3.99
KgCoOp	1.84	3.11	3.12	5.90	1.94	11.78	1.67	7.09	6.59	2.69	1.83	4.32
MaPLe	1.26	2.51	2.75	11.28	1.50	9.06	1.22	8.16	8.55	2.30	2.11	4.61
ProGrad	1.97	3.31	2.29	5.04	1.85	10.46	1.32	3.49	6.90	2.42	2.11	3.74
PromptSRC	1.58	2.98	2.39	5.03	1.55	8.55	0.79	5.50	7.24	2.46	1.84	3.63

(b) Distance-Aware Calibration

Table 16. Adaptive Calibration Error (ACE) on new classes in downstream datasets using various tuning methods for CLIP-ViT-B/16.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.45	3.41	3.34	4.77	1.78	6.06	3.64	8.93	9.39	5.50	2.15	4.58
CoOp	3.46	1.66	12.47	18.34	3.83	28.44	13.93	26.86	12.71	19.18	10.69	13.78
CoCoOp	3.30	2.02	2.30	11.49	1.65	11.19	2.00	15.73	12.68	4.38	1.48	6.20
ProDA	4.36	6.59	3.13	6.87	1.39	5.58	1.68	8.10	5.16	3.37	1.57	4.35
KgCoOp	1.56	3.05	2.78	6.79	1.99	12.48	1.52	7.34	6.83	2.61	1.73	4.43
MaPLe	1.11	2.31	2.04	13.43	1.24	10.31	2.95	15.38	9.10	2.53	2.39	5.71
ProGrad	1.22	2.86	3.13	5.90	2.04	12.96	1.65	5.18	7.66	2.55	1.84	4.27
PromptSRC	1.59	2.95	2.34	6.05	1.54	9.67	0.99	5.75	7.54	2.98	1.75	3.92

(a) w/o calibration

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	1.45	3.41	3.34	4.77	1.78	6.06	3.64	8.93	9.39	5.50	2.15	4.58
CoOp	2.60	1.70	5.17	10.18	1.75	17.27	4.00	10.51	8.58	8.63	5.66	6.91
CoCoOp	2.97	2.27	2.35	7.69	1.62	8.92	1.09	7.73	13.49	2.66	1.69	4.77
ProDA	4.64	4.72	3.37	6.54	0.64	7.73	1.04	5.64	3.99	4.19	2.37	4.08
KgCoOp	1.56	3.02	3.11	6.61	1.93	11.74	1.82	7.26	6.63	2.70	1.78	4.38
MaPLe	1.19	2.44	2.57	11.28	1.48	8.90	1.37	8.24	9.12	2.32	2.12	4.64
ProGrad	1.21	3.19	2.28	5.66	1.88	10.29	1.42	3.82	7.10	2.20	2.12	3.74
PromptSRC	1.65	2.86	2.34	5.50	1.48	8.46	1.03	5.42	7.43	2.54	1.83	3.69

(b) Distance-Aware Calibration

Table 17. Maximum Calibration Error (MCE) on new classes in downstream datasets using various tuning methods for CLIP-ViT-B/16.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	0.60	1.19	0.71	1.07	0.62	1.84	0.63	2.57	2.44	1.16	0.51	1.21
CoOp	2.60	0.62	2.40	5.46	1.43	5.00	4.36	7.08	4.42	6.09	2.67	3.83
CoCoOp	2.18	0.97	0.59	3.08	0.55	2.61	0.66	3.79	3.81	1.13	0.42	1.80
ProDA	3.56	1.66	0.65	1.34	0.61	1.44	0.40	1.72	1.42	0.89	0.33	1.27
KgCoOp	0.63	1.25	0.78	1.46	0.65	3.41	0.34	1.51	1.96	0.55	0.42	1.18
MaPLe	0.47	1.03	0.48	5.66	0.45	2.47	0.88	4.24	2.70	0.91	0.74	1.82
ProGrad	0.60	1.10	1.05	1.28	0.61	3.16	0.60	1.32	2.40	0.72	0.55	1.22
PromptSRC	0.66	1.02	0.51	1.2	0.56	2.41	0.23	1.59	2.65	0.65	0.54	1.09

(a) w/o calibration

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	0.60	1.19	0.71	1.07	0.62	1.84	0.63	2.57	2.44	1.16	0.51	1.21
CoOp	1.50	0.66	1.21	2.08	0.49	3.57	1.03	2.24	3.03	1.70	1.28	1.71
CoCoOp	1.92	1.06	0.55	1.60	0.53	2.34	0.28	1.75	4.41	0.58	0.43	1.40
ProDA	3.98	1.27	0.81	1.20	0.28	1.65	0.42	1.61	1.13	1.32	0.81	1.32
KgCoOp	0.62	1.21	0.85	1.37	0.64	3.02	0.42	1.33	1.75	0.77	0.41	1.13
MaPLe	0.39	1.06	0.77	4.42	0.50	2.26	0.40	1.95	2.55	0.66	0.64	1.42
ProGrad	0.71	1.17	0.88	1.18	0.58	2.60	0.40	1.18	1.96	0.69	0.6	1.09
PromptSRC	0.64	0.96	0.58	1.23	0.55	2.31	0.22	1.33	2.72	0.79	0.57	1.08

(b) Distance-Aware Calibration

Table 18. Proximity-Informed Expected Calibration Error (PIECE) on new classes in downstream datasets using various tuning methods for CLIP-ViT-B/16.

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	3.95	3.88	4.79	7.62	2.17	10.49	4.31	13.75	10.37	7.72	3.05	6.55
CoOp	6.21	4.26	12.73	18.80	4.07	29.03	14.01	27.29	15.07	19.79	10.71	14.72
CoCoOp	5.49	2.96	4.54	12.35	2.19	13.52	3.46	18.64	15.08	7.68	2.54	8.04
ProDA	6.08	6.89	5.56	8.90	2.39	8.87	2.88	12.84	7.96	7.46	2.48	6.57
KgCoOp	4.14	3.86	4.95	8.81	2.33	14.87	3.08	12.17	9.27	7.16	2.73	6.67
MaPLe	3.39	3.27	4.53	14.18	1.90	12.33	3.93	18.35	11.95	6.51	3.17	7.59
ProGrad	4.45	3.43	5.29	8.31	2.49	15.21	3.33	10.64	10.95	7.15	2.96	6.75
PromptSRC	3.80	3.82	4.41	8.02	2.07	12.34	2.81	12.10	10.31	6.49	2.64	6.26

(a) w/o calibration

Methods	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	ImageNet	AVG
ZeroshotCLIP	3.95	3.88	4.79	7.62	2.17	10.49	4.31	13.75	10.37	7.72	3.05	6.55
CoOp	5.89	4.15	6.84	12.07	2.58	19.43	4.92	13.13	11.88	12.23	6.08	9.02
CoCoOp	5.37	3.19	4.47	10.14	2.14	12.19	2.88	12.78	15.69	7.11	2.65	7.15
ProDA	6.25	5.24	5.61	8.80	1.59	10.06	2.74	11.44	7.26	7.61	3.20	6.35
KgCoOp	4.13	3.81	5.05	8.73	2.28	14.47	3.34	12.21	9.13	7.03	2.71	6.63
MaPLe	3.59	3.33	4.90	12.40	2.12	11.77	2.88	14.09	11.96	6.70	2.99	6.98
ProGrad	4.77	3.69	5.00	8.07	2.42	13.64	3.12	10.86	10.40	6.90	3.15	6.55
PromptSRC	3.75	3.74	4.48	8.00	1.98	11.71	2.84	12.41	10.10	6.08	2.74	6.17

(b) Distance-Aware Calibration