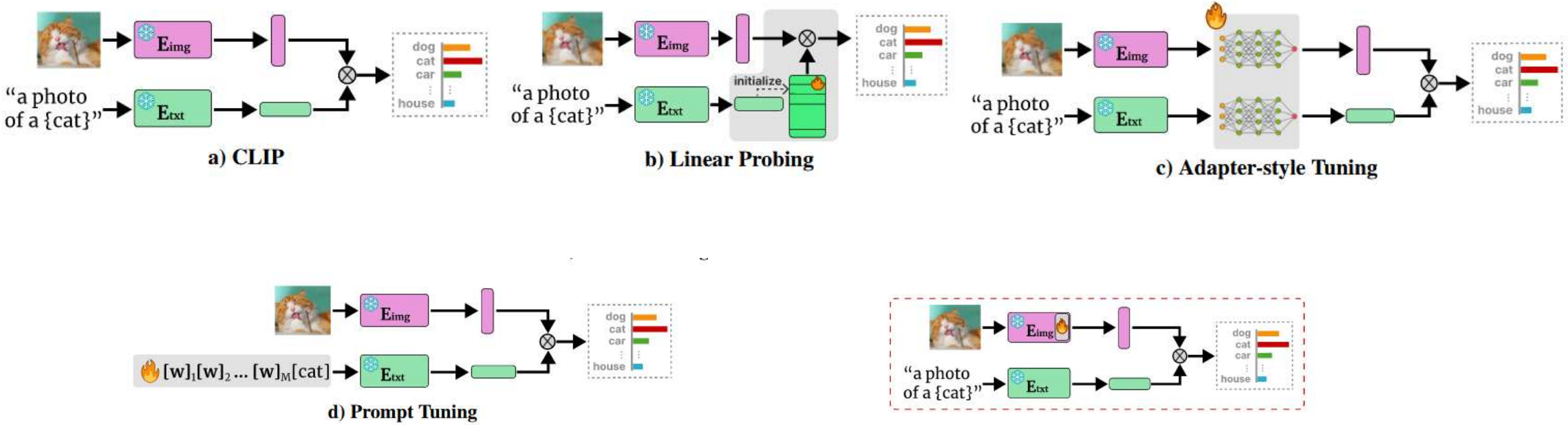
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

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]

ICML 2024

Background



Few-shot classification with CLIP

Background

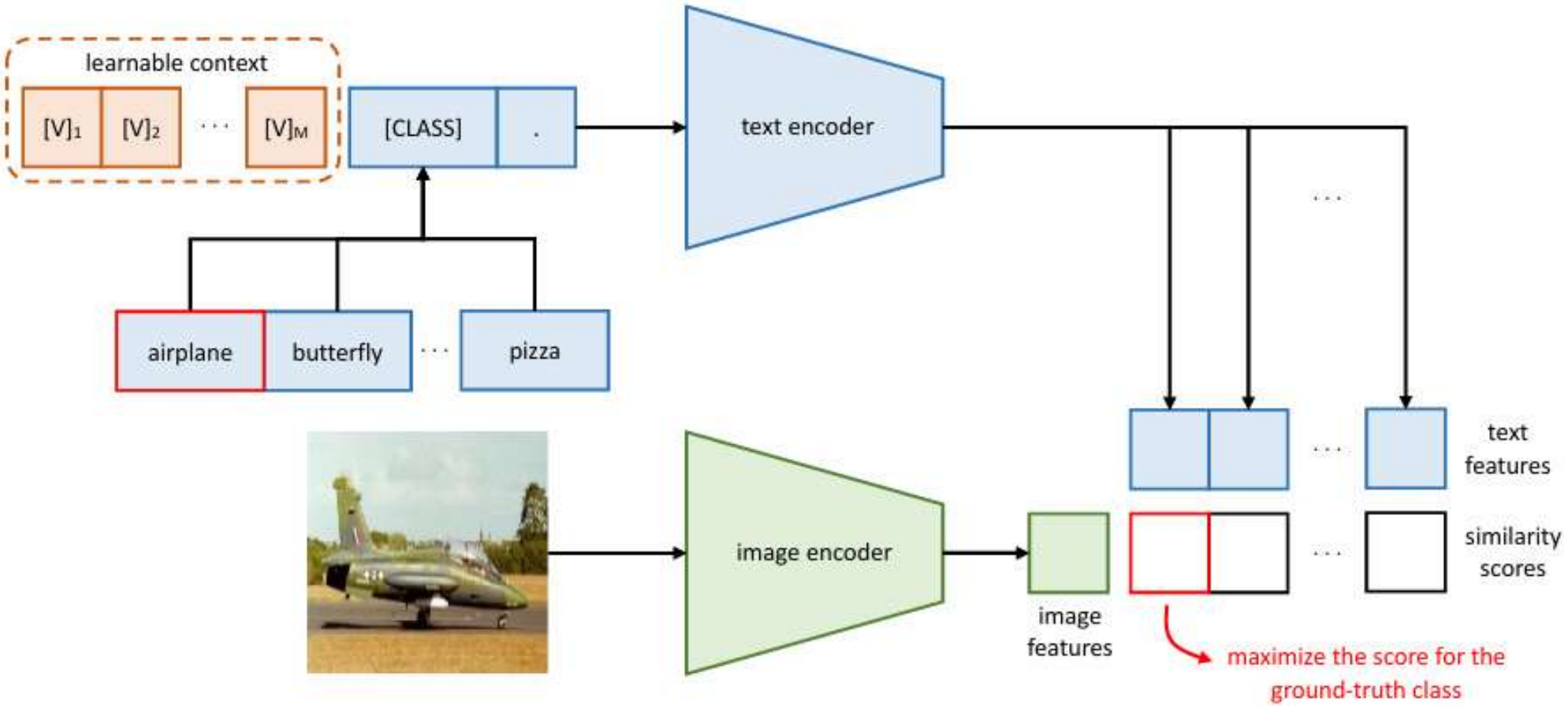


Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and the other is class-specific context, which learns for each class a specific set of context vectors.

(1) 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.

Background

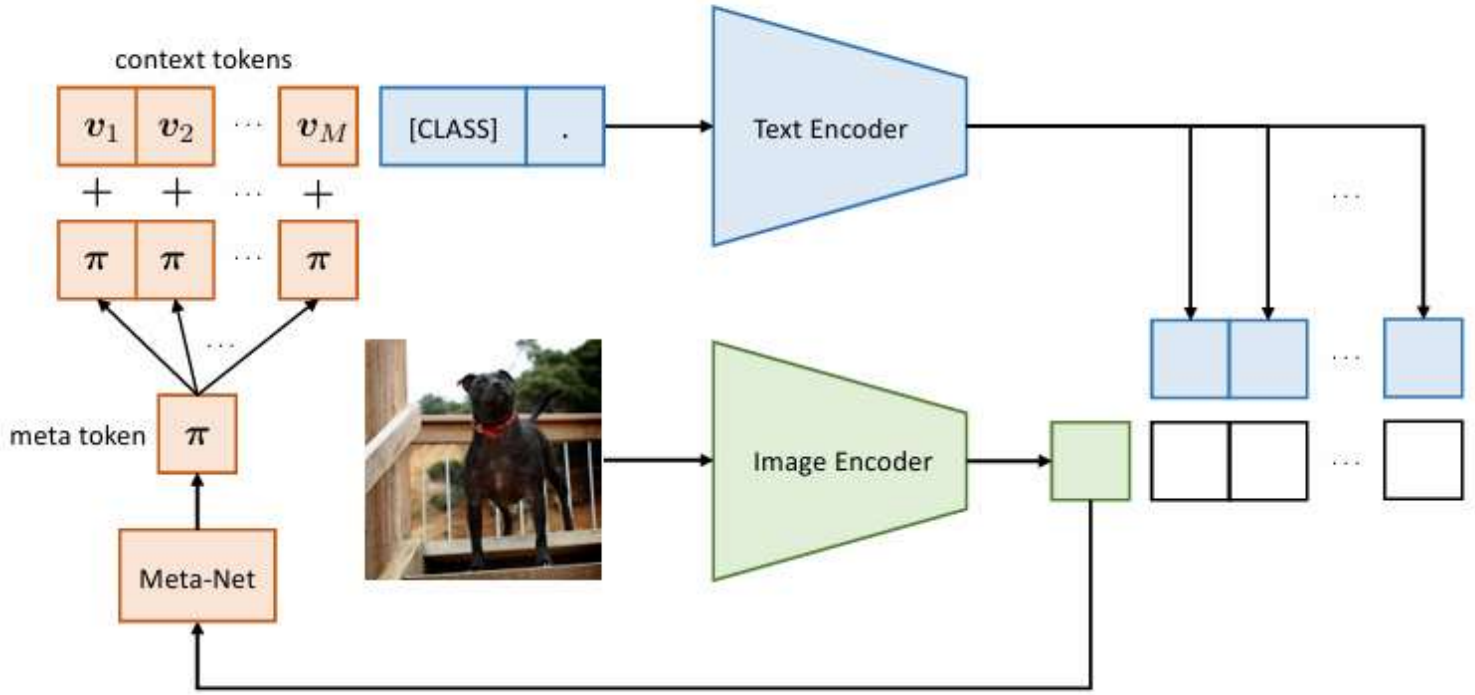


Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

(2) 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.

Background

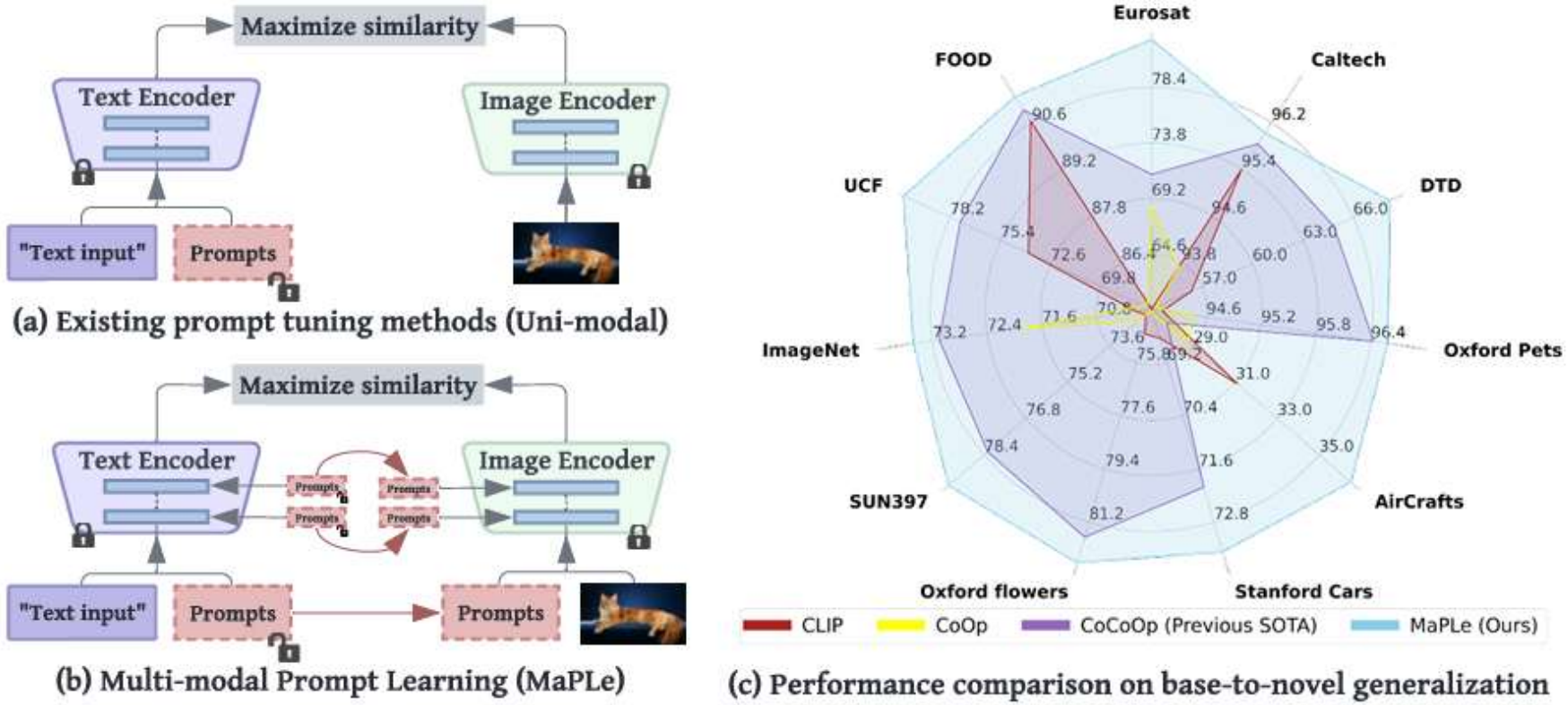


Figure 1. Comparison of MaPLe with standard prompt learning methods. **(a)** Existing methods adopt uni-modal prompting techniques to fine-tune CLIP representations as prompts are learned only in a single branch of CLIP (language or vision). **(b)** MaPLe introduces branch-aware hierarchical prompts that adapt both language and vision branches simultaneously for improved generalization. **(c)** MaPLe surpasses state-of-the-art methods on 11 diverse image recognition datasets for novel class generalization task.

(3) 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.

Background

①Are fine-tuned VLMs well-calibrated?

Expected Calibration Error (ECE)

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

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

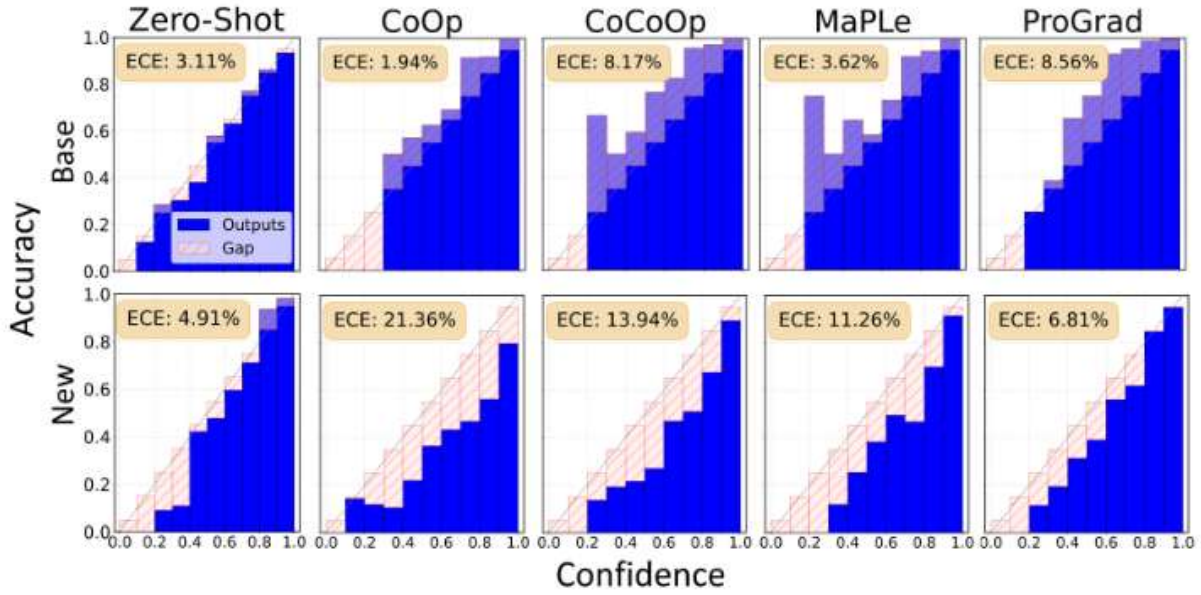


Figure 1. Reliability of fine-tuned CLIP (ViT-B/16) on the Flower102 dataset. ECE: Expected Calibration Error (lower is better). Miscalibration is depicted in pink for overconfidence and purple for underconfidence.

Background

②Can fine-tuned VLMs be calibrated?

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	-	-	-

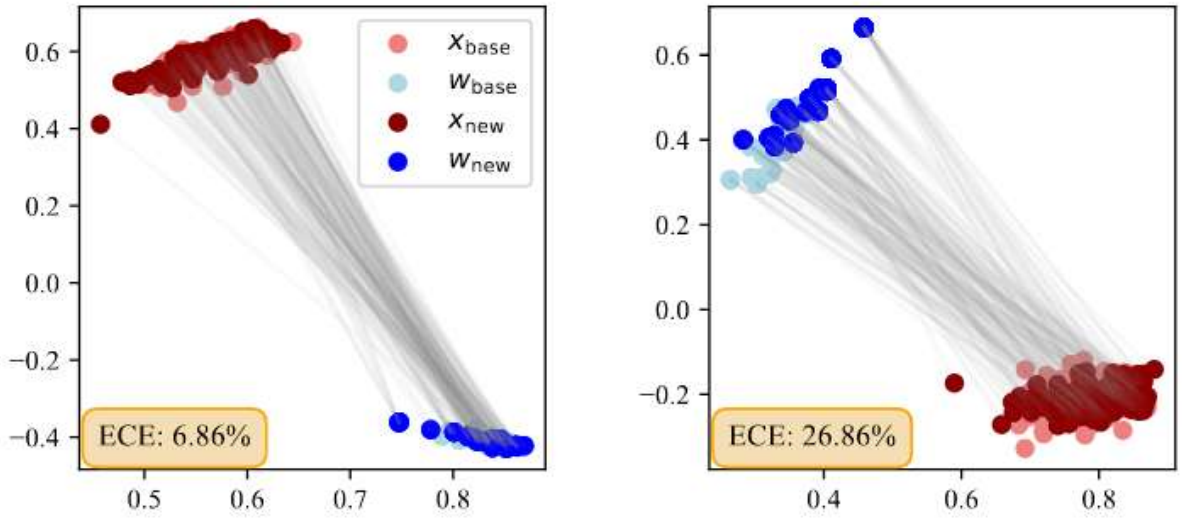
TS: Temperature Scaling(温度放缩)
DEN: Density-Ratio Calibration（密度比校准）
HB: Histogram Binning（直方图分桶）
IR: Isotonic Regression（等距回归）
MIR: Multi-Isotonic Regression（多等距回归）

IR：非参数方法，使用单调递增函数拟合预测概率与真实标签之间的关系
MIR:使用于多分类，针对每个类别进行单独校准

Finding
(1) Post-hoc calibration can remedy miscalibration in base classes.
(2) Post-hoc calibration on base classes can not transfer to new classes.

Background

Feature Space Analysis



(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

the deviation degree in the textual gap is crucial for open-vocabulary calibration

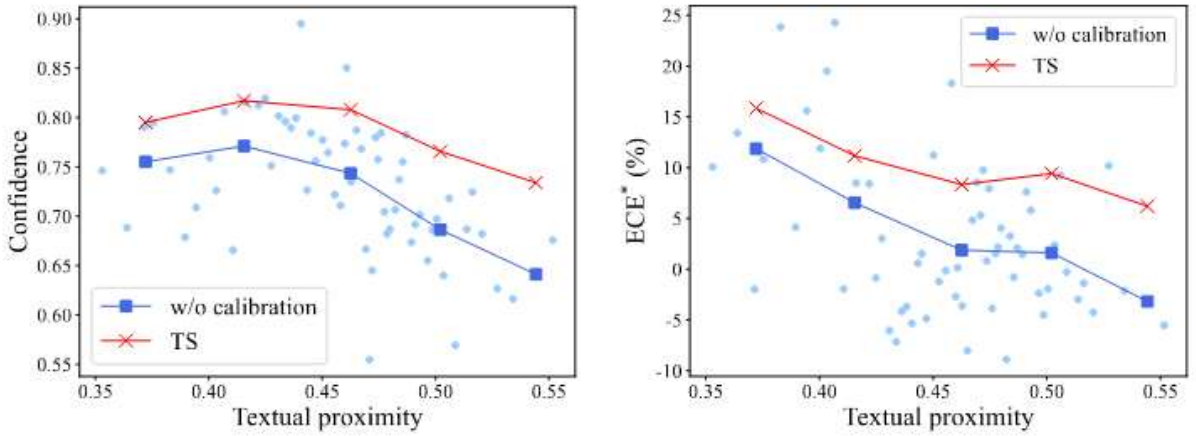
Motivation

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:

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

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

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



(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.

Lower proximity correlates with higher confidence and ECE

Textual deviation estimation

Ideally, the model is expected to give highly uncertain predictions for examples from novel classes, with relatively low accuracy.

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:

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

Calibrated inference

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

$$\hat{c} = argmax_c p(c|x)$$

Experiments

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.

	ECE(↓)		ACE(↓)		MCE(↓)		PIECE(↓)	
Method	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

ECE:将所有预测样本分成 M个置信度区间（bins），计算每个 bin 中预测置信度和真实准确率之间的差值的加权平均

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

ACE:但 bins 是根据置信度排序后均匀划分样本数量而非固定区间宽度

$$\mathrm{ACE} = \frac{1}{M} \sum_{m=1}^{M} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

MCE :衡量的是所有 bins 中校准误差的最大值

$$\mathrm{MCE} = \max_{m \in \{1,\ldots,M\}} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

PIECE:引入了预测区间的概念,评估实际标签是否落在预测的置信区间内

$$\mathrm{PIECE}_\alpha = \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(y_i \in \hat{I}_i^{(\alpha)}\right) - \alpha \right|$$

Experiments

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.

	ECE(↓)		ACE(↓)		MCE(↓)		PIECE(↓)	
Method	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

DAC improves open-vocabulary calibration in existing prompt tuning

Experiments

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.

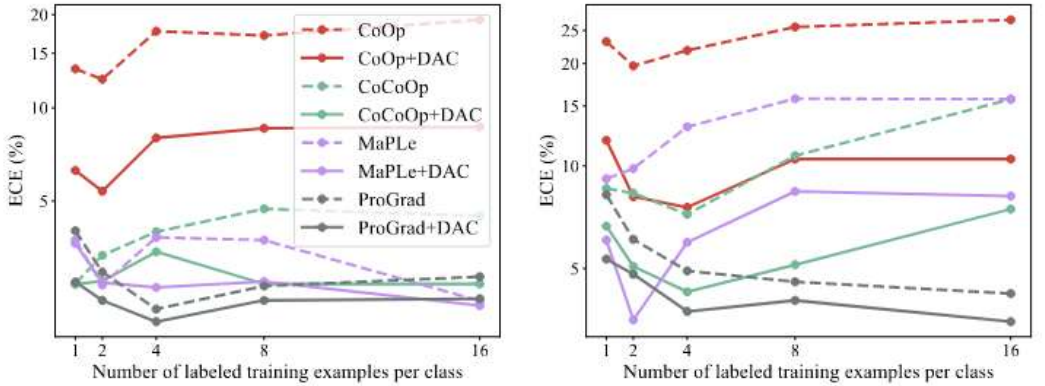
Method		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CoOp	Conf	0.00	19.24	18.86	15.87	20.42	33.28	30.30	37.84	41.60	18.57
	DAC	0.00	4.95	8.17	11.33	6.51	11.42	24.12	17.41	11.37	-0.94
	Δ	0.00	-14.29	-10.69	-4.54	-13.91	**-21.86**	-6.18	**-20.43**	**-30.23**	-19.51
MaPLe	Conf	0.00	0.00	-12.80	3.90	16.73	10.50	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

DAC significantly reduces calibration error, especially for high-confidence predictions

Experiments

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.

DAC is effective across various few-shot settings

Ablation results

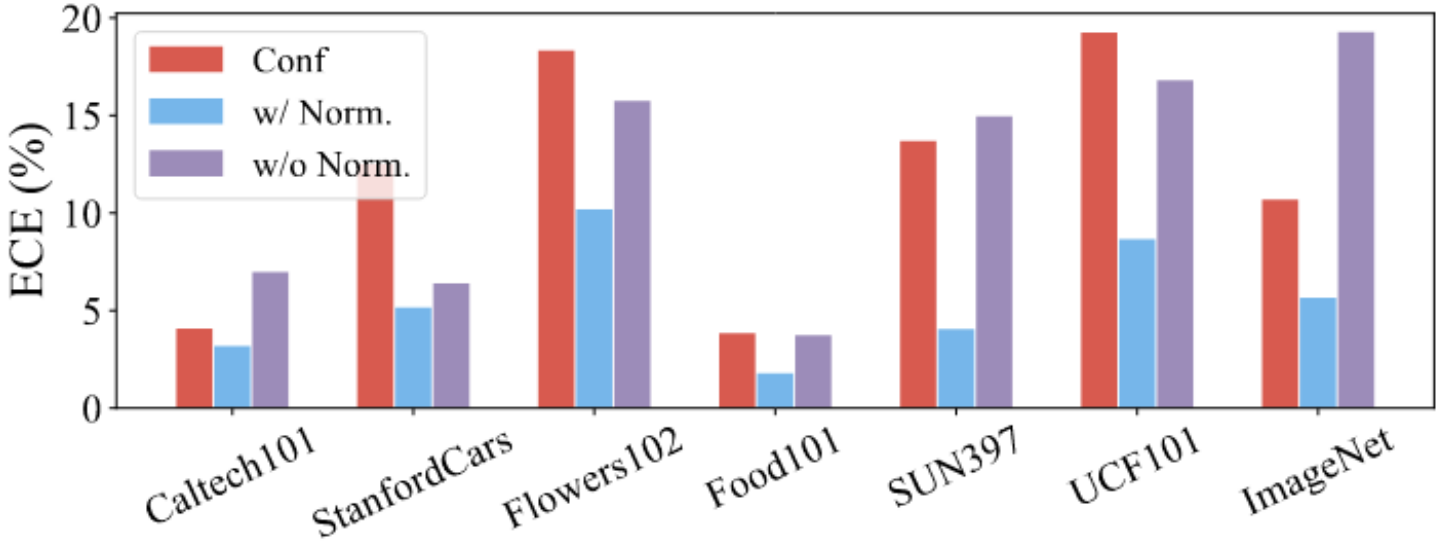


Figure 7. Ablation results of textual feature normalization with CoOp. We compare the effect of using normalization in the textual feature vs. without normalization.

Textual feature normalization is critical

Thanks