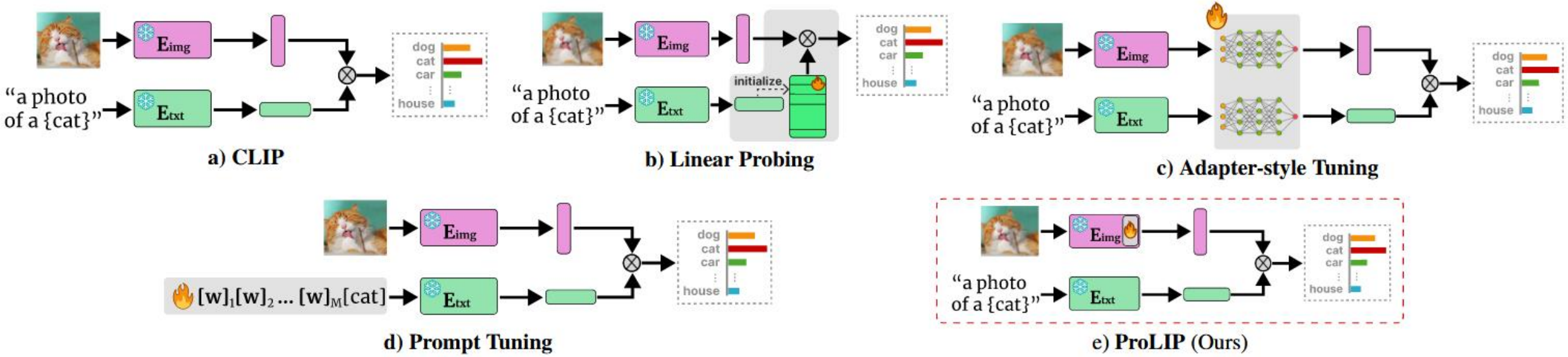
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

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

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.

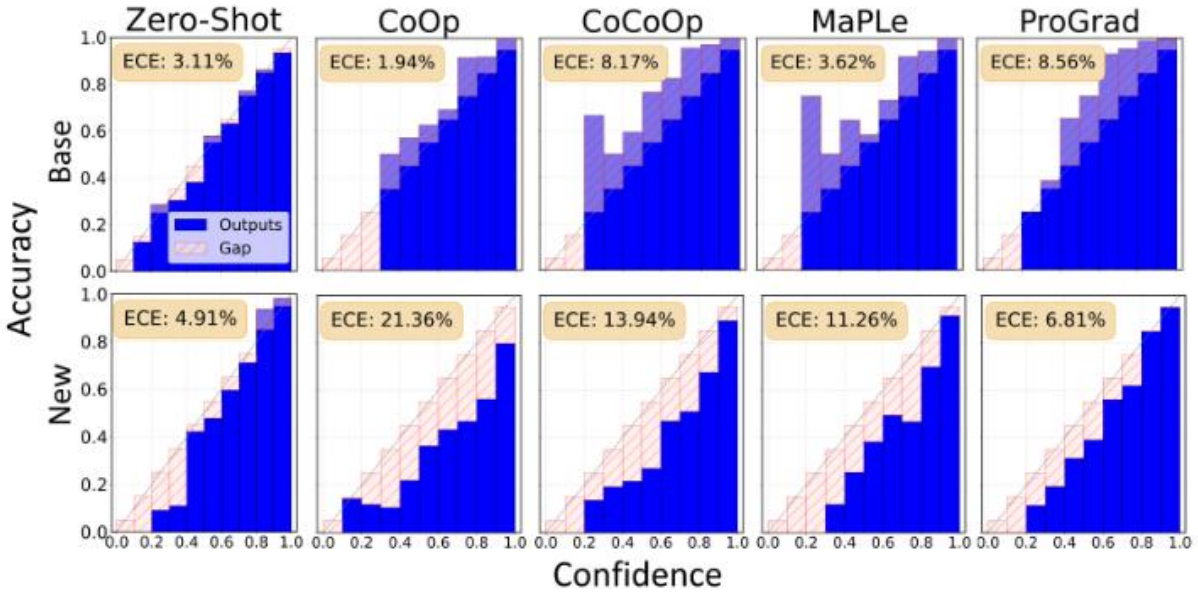


Figure 1. Reliability of fine-tuned CLIP (ViT-B/16) on the Flower102 dataset. ECE: Expected Calibration Error (lower is better). Miscalibration is depicted in pink for overconfidence and purple for underconfidence.

Background

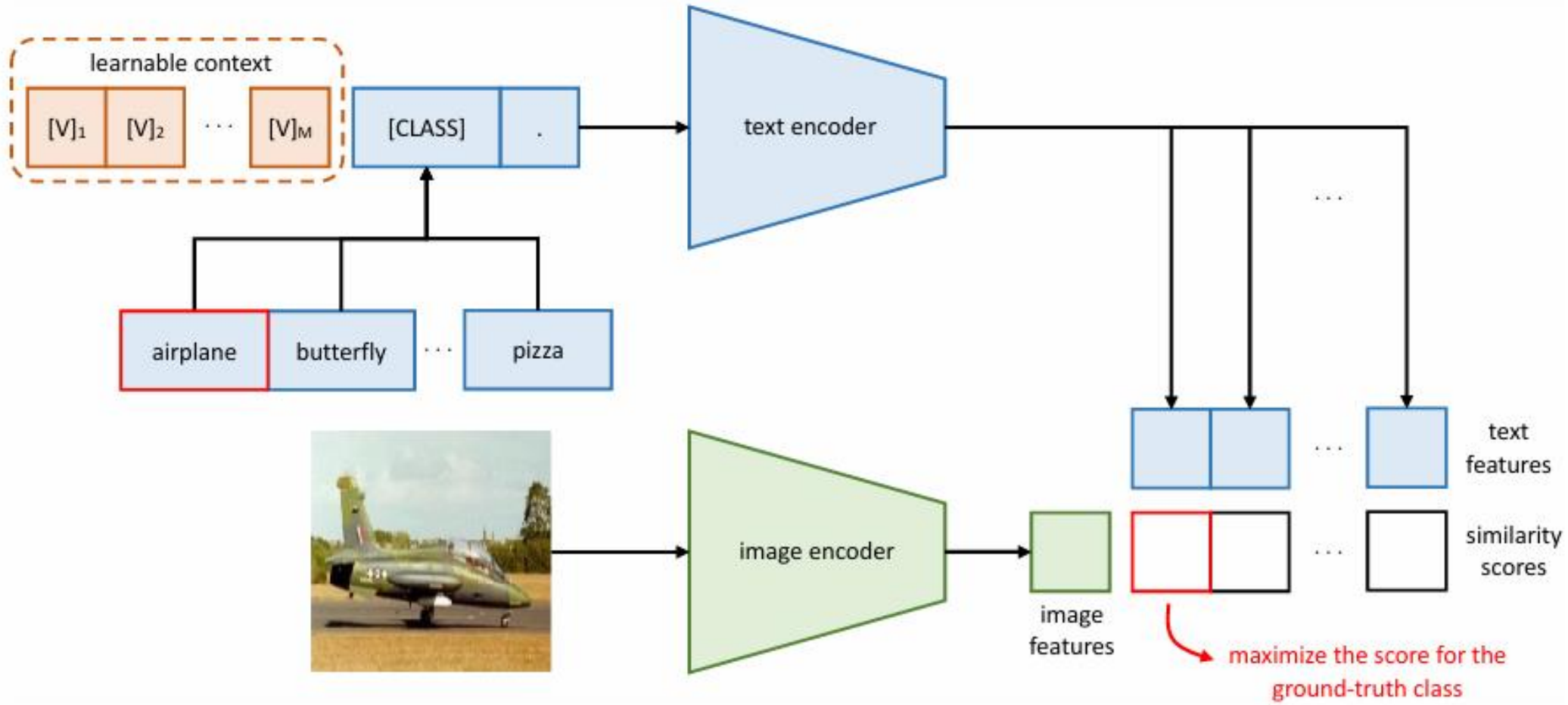
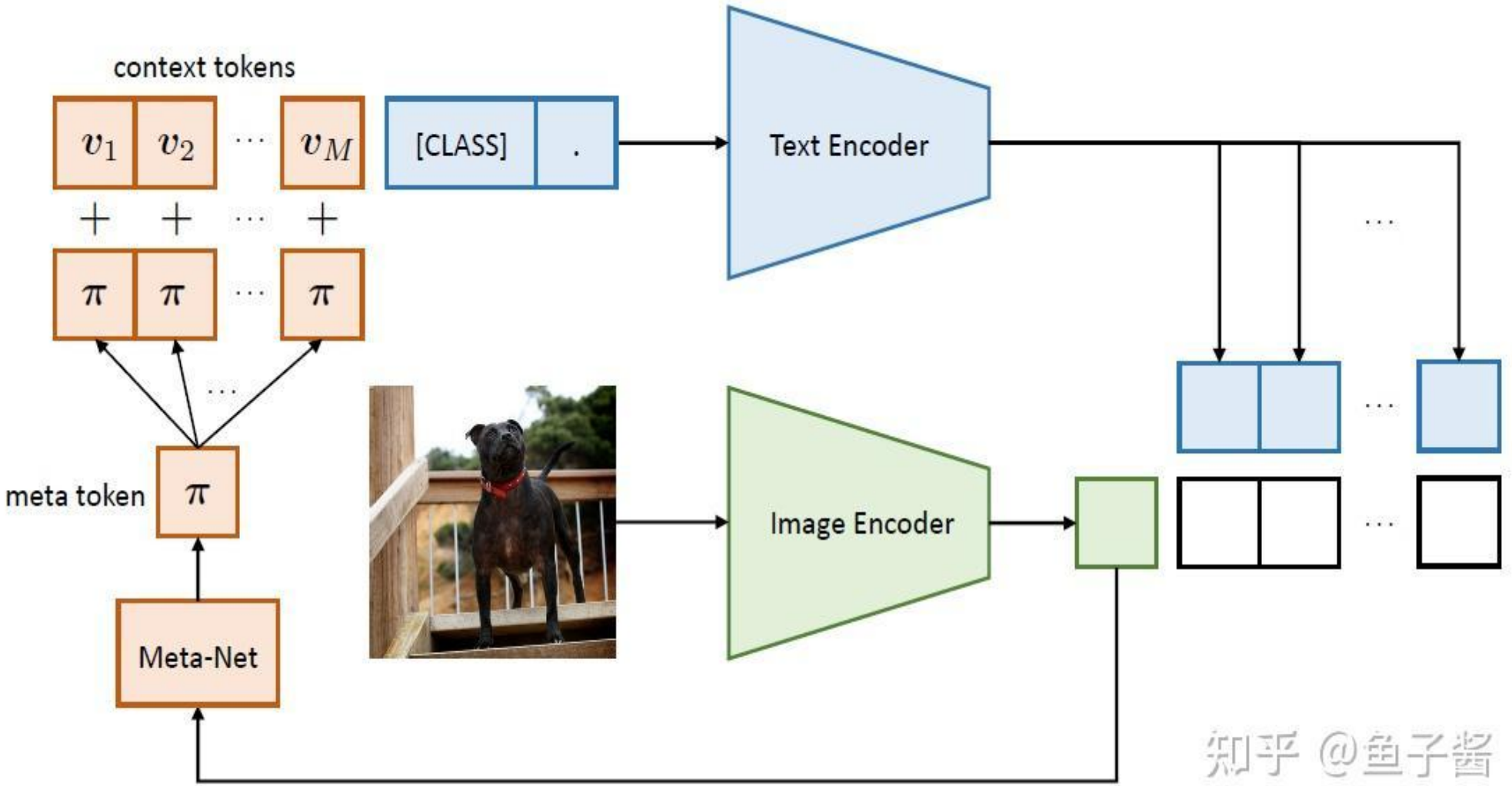


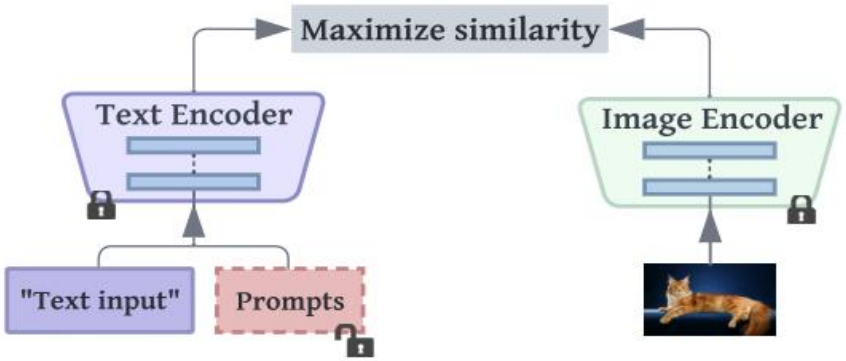
Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and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

Background

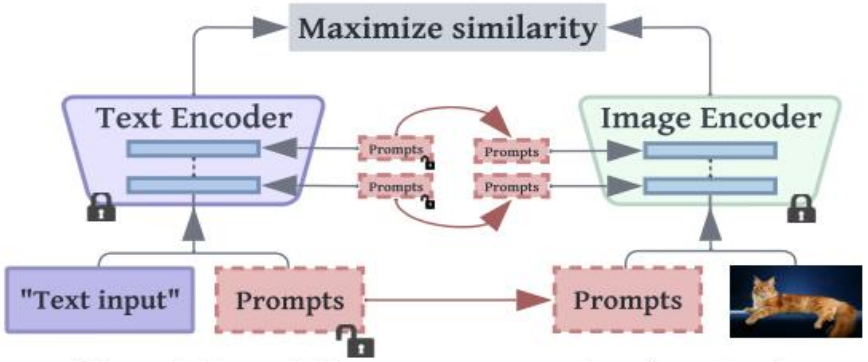


CoCoOp

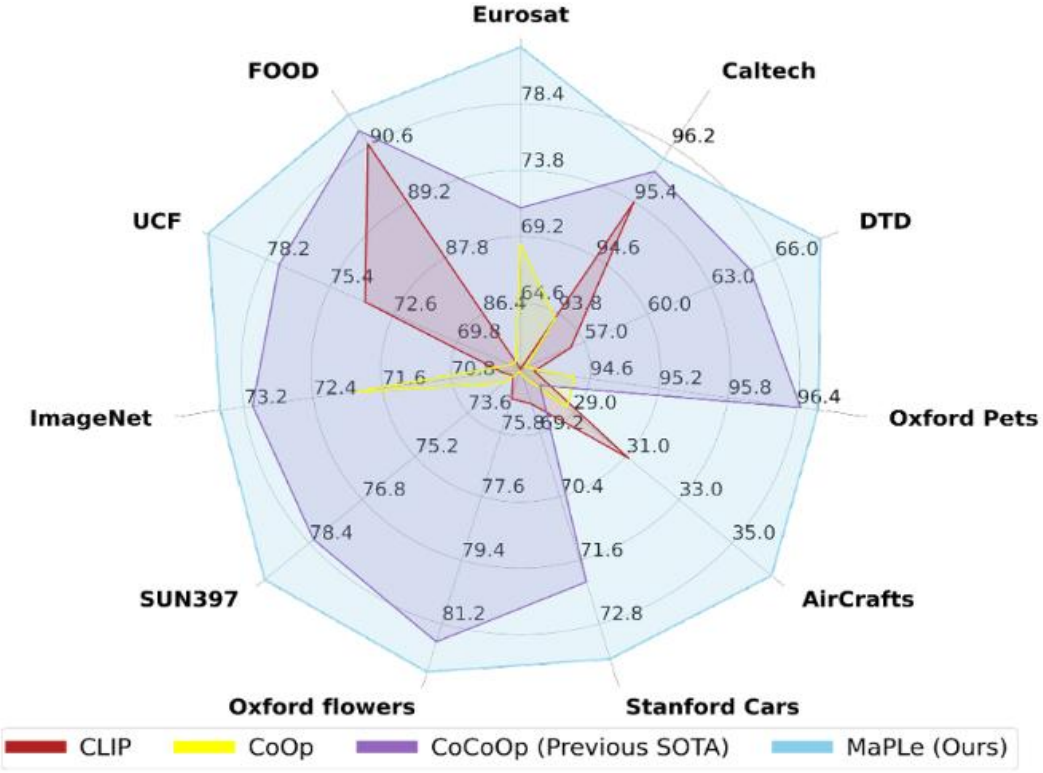
Background



(a) Existing prompt tuning methods (Uni-modal)



(b) Multi-modal Prompt Learning (MaPLe)

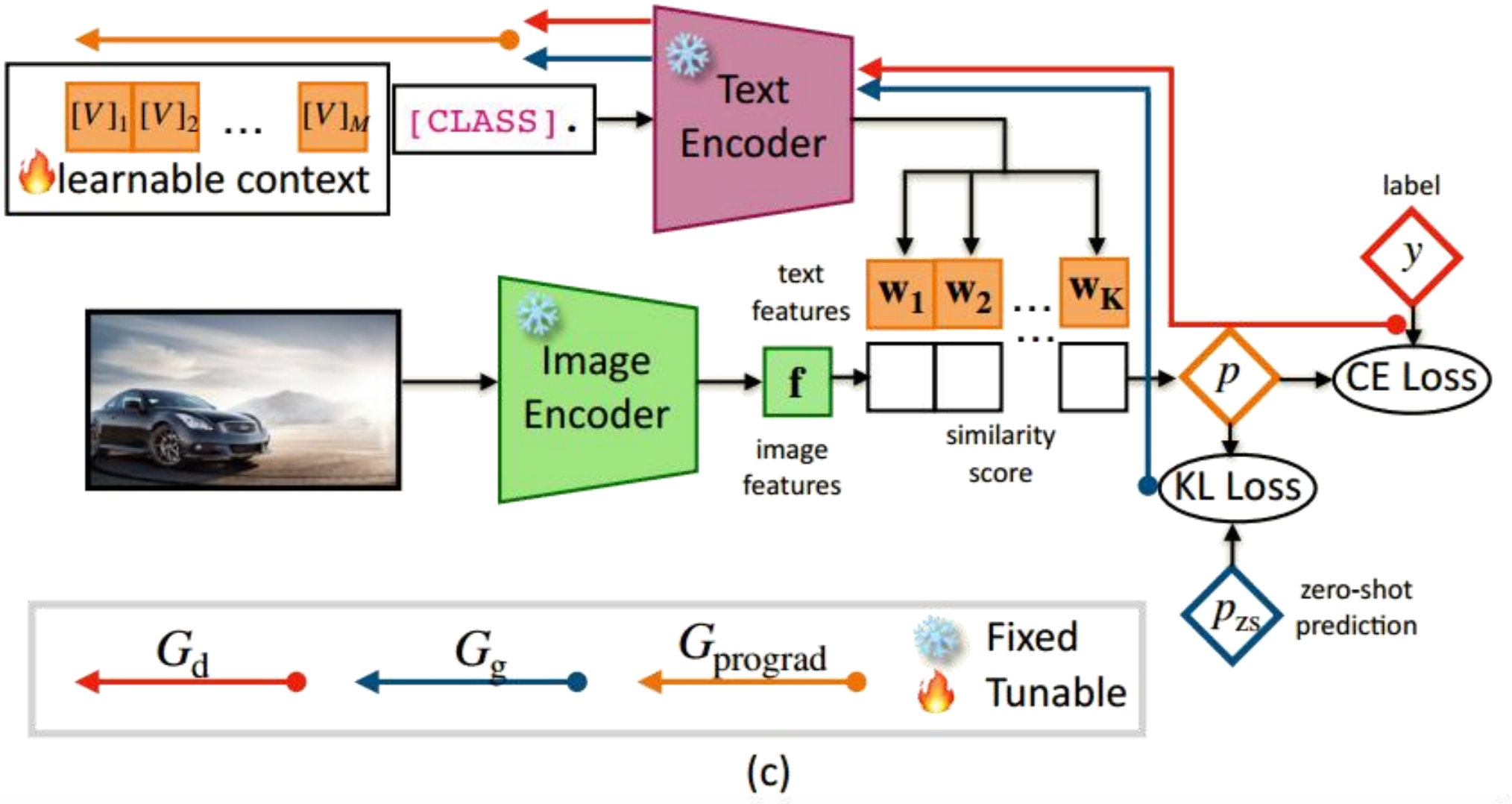


(c) Performance comparison on base-to-novel generalization

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.

MaPLe

Background



ProGrad

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.

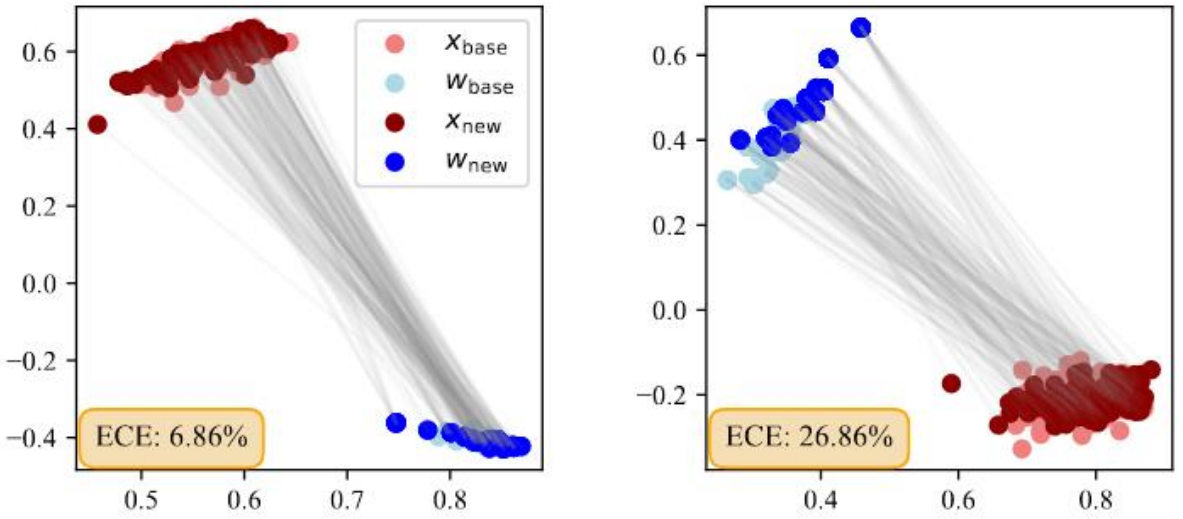
	ZS	Conf	TS	DEN	HB	IR	MIR
Base classes	3.58	4.82	1.94	0.73	4.23	2.09	0.82
New classes	2.09	1.59	3.90	3.86	-	-	-

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.

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

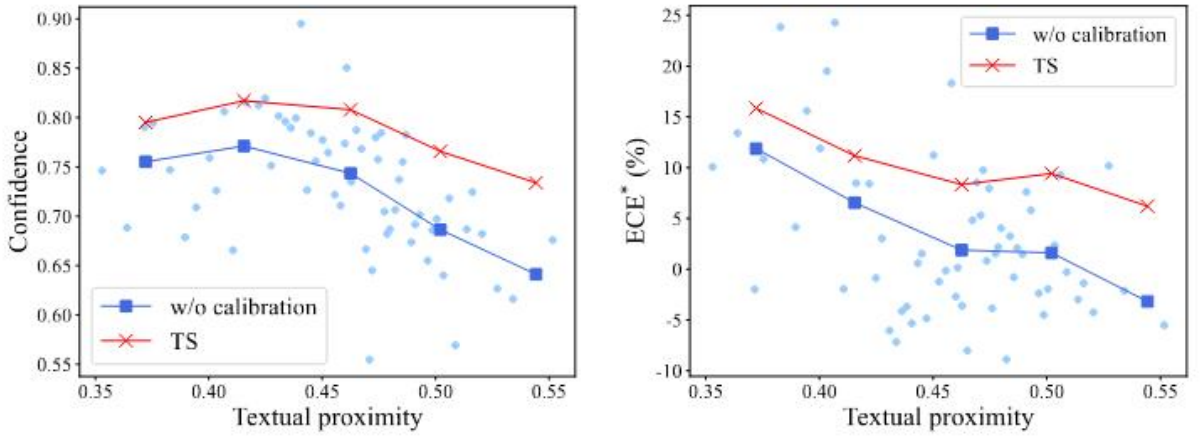
Motivation

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.

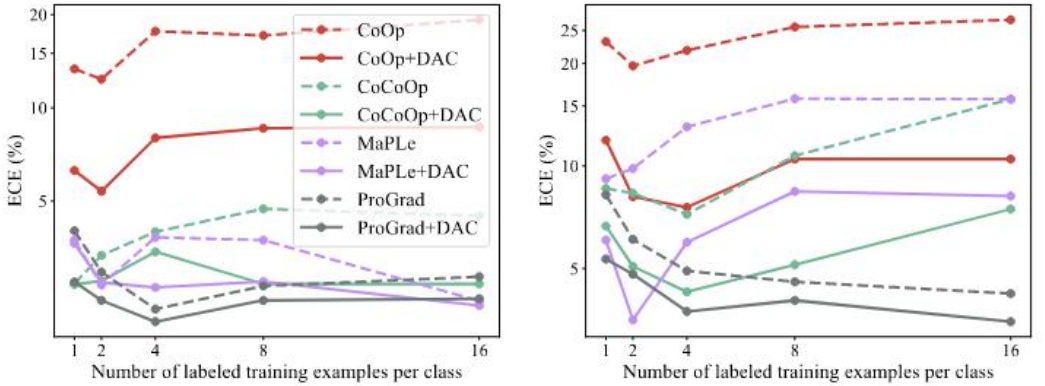
Method		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CoOp	Conf	0.00	19.24	18.86	15.87	20.42	33.28	30.30	37.84	41.60	18.57
	DAC	0.00	4.95	8.17	11.33	6.51	11.42	24.12	17.41	11.37	-0.94
	Δ	0.00	-14.29	-10.69	-4.54	-13.91	**-21.86**	-6.18	**-20.43**	**-30.23**	-19.51
MaPLe	Conf	0.00	0.00	-12.80	3.90	16.73	10.50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

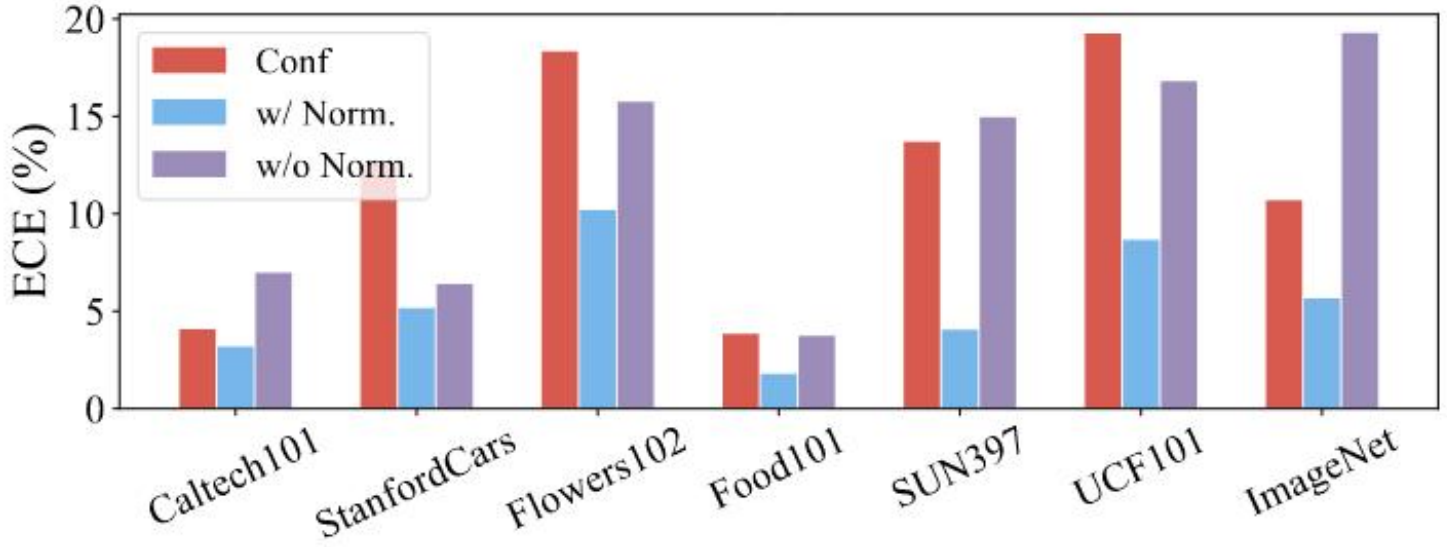


Figure 7. Ablation results of textual feature normalization with CoOp. We compare the effect of using normalization in the textual feature vs. without normalization.

Textual feature normalization is critical

Contrast-Aware Calibration for Fine-Tuned CLIP: Leveraging Image-Text Alignment

Song-Lin Lv [1 2] **Yu-Yang Chen** [1 2] **Zhi Zhou** [2] **Yu-Feng Li** [2 3] **Lan-Zhe Guo** [1 2]

arxiv 2025

Contrast Metric

Contrast is an indicator used to measure a model's ability to distinguish between positive and negative samples, which is widely used in contrastive learning.

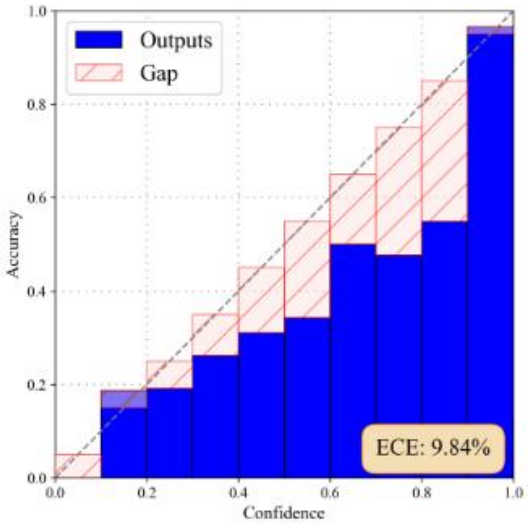
- **Positive Similarity.** For each sample i, the similarity score with its ground-truth label y_i is extracted as $s_i^+ = S[i, y_i]$, with the average positive similarity is defined as: Positive_Mean $= \frac{1}{N}\sum_{i=1}^{N} s_i^+$
- **Negative Similarity.** For each sample i, the maximum similarity score among incorrect labels is calculated as $s_i^- = max_{j \neq y_i} S[i, j]$, with the average negative similarity is defined as: Negative_Mean $= \frac{1}{N}\sum_{i=1}^{N} s_i^-$
- **Difference Calculation.** The contrast metric is calculated as the difference between the average of positive and negative similarities:

$$\text{Contrast} = \frac{1}{N}\sum_{i=1}^{N} s_i^+ - \frac{1}{N}\sum_{i=1}^{N} s_i^- \qquad (3)$$

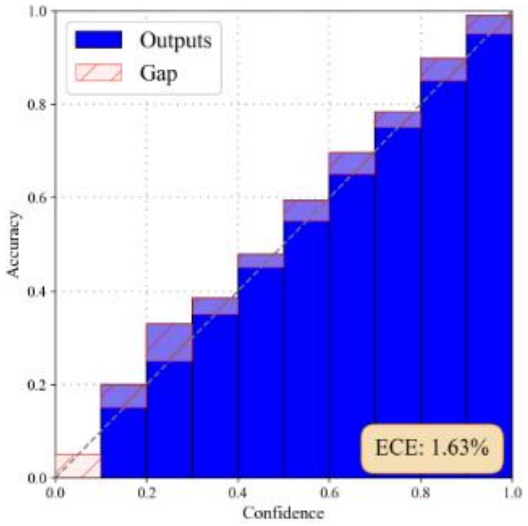
Motivation

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

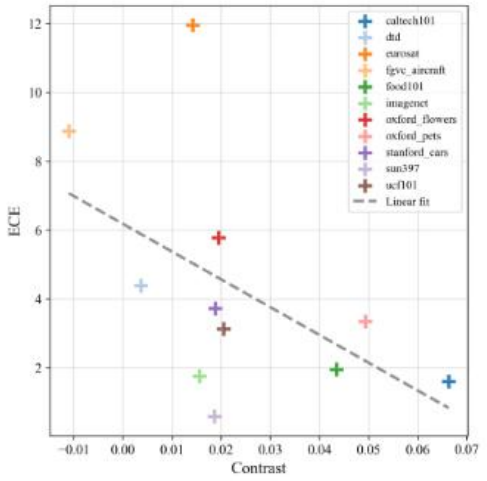
when the representations of fine-tuned VLMs deviate from the pre-trained image-text alignment, their class scores often become biased toward certain categories or exhibit similar scores across multiple classes, losing the pre-trained ability to discriminative intra-class and inter-class samples and causing miscalibration.



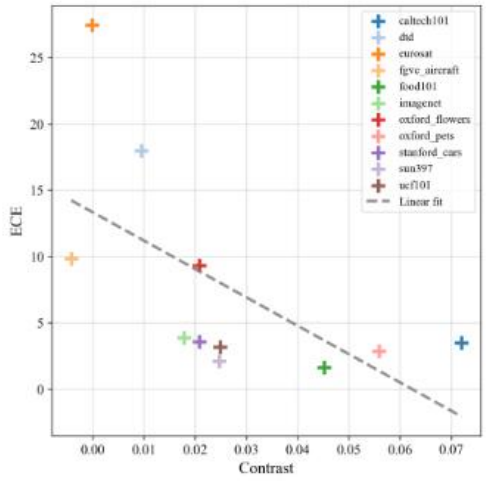
(a) ECE of FGVCAircraft



(b) ECE of Food101



(c) KgCoOp



(d) MaPLe

①Overconfidence caused by interclass similarity
②Underconfidence caused by intraclass variation

①contrast and ECE exhibit a negative correlation for unseen classes
②well-aligned VLMs typically exhibit better confidence calibration

Methods

$$z = \frac{1}{N}\sum_{i=1}^{N}|P_i - \hat{P}_i|, \tag{5}$$

where N denotes the total number of classes, and $\hat{P} = \{\hat{p}_i\}_{i=1}^{N}$ and $P = \{p_i\}_{i=1}^{N}$ represent the logits output by the original and fine-tuned CLIP, respectively. In particular, according to our analysis, the logits of contrastive learning-based VLMs are equivalent to the contrast metric, so z can serve as an indicator for measuring the confidence difference between the original and fine-tuned VLMs. To better leverage its negative correlation with ECE, we design the following function to transform z into CAW:

$$\gamma = \alpha \cdot e^{-kz}. \tag{6}$$

z: Due to **the negative correlation between ECE and contrast**,a range of [0, 1] and aligns with the required monotonicity

α: Since fine-tuned VLMs may be underconfidence and overconfident in various datasets, equipping CAW with the ability to **deal with underconfidence**

k: Since the text and image features in CLIPbased models **undergo normalization before computing the logits**, the L1 distance may be small,and enable the function to capture input variations more effectively

different datasets and fine-tuning methods require varying levels of calibration

$$\hat{\gamma} = \begin{cases} \gamma^2, & \text{if } \gamma < \lambda_1, \\ \gamma, & \text{if } \lambda_1 \leq \gamma \leq \lambda_2, \\ \gamma^2 & \text{if } \gamma > \lambda_2. \end{cases} \tag{7}$$

where $\hat{\gamma}$ represents the final calibration weight, γ represents the output of CAW, and λ_1 and λ_2 represent the boundary points of the interval for shrinking or amplifying γ. Through the following two designed modules, CAC achieves more flexible confidence calibration compared to CAW:

Given an input image i, we first collect the CAC scores of this image, denoted as $\hat{\gamma}_i$, which is then used to calculate the rectified logits as follows:

$$L_i^{CAC} = \hat{\gamma}_i * \tau * logits_i \tag{8}$$

Experiments

Table 1. Average calibration performance across 11 datasets. "Conf" represents the original performance on open-vocabulary classes with existing tuning methods. ↓ indicates smaller values are better. **Bold** numbers are significantly superior results.

Method	ECE(↓)			ACE(↓)			MCE(↓)			PIECE(↓)		
	Conf	DAC	CAC	Conf	DAC	CAC	Conf	DAC	CAC	Conf	DAC	CAC
CoCoOp	5.44	5.70	**4.24**	5.35	5.60	**4.22**	1.38	1.40	**1.20**	7.35	8.06	**6.83**
KgCoOp	3.98	4.11	**3.85**	3.93	4.09	**3.78**	1.08	1.18	**1.10**	6.45	6.62	**6.39**
MaPLe	7.80	5.91	**5.35**	7.77	5.93	**5.30**	2.08	1.62	**1.61**	9.53	8.19	**7.69**
ProGrad	5.04	6.13	**4.04**	4.95	6.18	**4.05**	1.47	1.53	**1.24**	7.41	8.20	**6.75**
PromptSRC	4.29	4.55	**3.47**	4.24	4.41	**3.40**	1.16	1.17	**1.03**	6.70	6.82	**6.12**

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 4. Average calibration performance across 11 datasets in 5 prompt learning methods. "w/o" indicates the results of CAC without the inclusion of that specific component.

Method	**Conf**	**DAC**	**w/o k**	**w/o α**	**w/o EXP**	**w/o piecewise**	**CAC**
CoCoOp	5.44	5.70	9.08	9.73	11.15	4.72	4.24
KgCoOp	3.98	4.11	7.52	8.12	26.74	3.97	3.85
MaPLe	7.80	5.91	11.32	11.25	28.13	6.57	5.35
ProGrad	5.04	6.13	8.92	11.84	28.85	4.44	4.04
PromptSRC	4.29	4.55	7.80	9.41	23.87	3.84	3.47

Experiments

Table 3. Average calibration performance of different k of CAC across 11 datasets.

Method	Conf	10	15	20	25
CoCoOp	5.44	4.82	**4.24**	6.41	11.09
KgCoOp	3.98	3.82	3.85	**3.66**	5.16
MaPLe	7.80	6.82	**5.35**	8.14	12.52
ProGrad	5.04	4.57	**4.04**	8.20	12.74
PromptSRC	4.29	3.93	**3.47**	5.33	8.31

Table 5. Average calibration performance of different α of CAC across 11 datasets.

Method	Conf	1.00	1.05	1.10	1.20
CoCoOp	5.44	7.87	5.88	**4.24**	4.84
KgCoOp	3.98	5.86	**3.81**	3.85	4.73
MaPLe	7.80	9.13	6.93	**5.35**	5.85
ProGrad	5.04	10.38	7.09	**4.04**	4.29
PromptSRC	4.29	7.87	5.00	**3.47**	3.90

Table 6. The impact of different piecewise function thresholds on CAC confidence calibration.

Method	λ_1		λ_2		CAC
	0.85	0.95	0.95	1.05	
CoCoOp	5.63	4.84	4.63	4.57	4.24
KgCoOp	4.22	3.55	3.72	3.59	3.85
MaPLe	7.88	5.92	5.71	5.69	5.35
ProGrad	5.19	5.27	4.71	4.65	4.04
PromptSRC	4.41	4.12	3.83	3.69	3.47

Thanks