TEMPERING CONFIDENCE: TEMPERATURE SCALING FOR CALIBRATING TEST-TIME PROMPT TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time prompt tuning (TPT) enables adaptation on a single test instance, achieving improved accuracy but often sacrificing calibration performance. Most existing calibration methods introduce additional regularization terms to promote dispersion across text embeddings and reduce calibration error, yet these methods often suffer from a drop in accuracy. Motivated by the well-calibrated nature of zero-shot predictions, we propose `CoTS`, a simple yet effective post-hoc calibration method that preserves accuracy. Specifically, `CoTS` applies temperature scaling to minimize the confidence gap between adapted and zero-shot predictions. To fully exploit the potential of multiple augmentations during adaptation, we introduce a weak-strong ensemble strategy that further boosts accuracy. We then apply `CoTS` to this ensemble, termed `E-CoTS`, to maintain its well-calibrated property. Extensive experiments on diverse datasets and backbones show that our approaches effectively mitigate miscalibration without compromising primary accuracy. For instance, `E-CoTS` reduces the average expected calibration error of TPT from 11.90% to 5.38% on ImageNet variants, while even increasing accuracy from 60.74% to 62.95%. Moreover, when integrated with existing calibration methods, `E-CoTS` usually enhances both accuracy and calibration simultaneously. Code is available at `https://anonymous.4open.science/r/CoTS/`.

1 INTRODUCTION

Vision-language models (VLMs), such as CLIP [55] and ALIGN [27], are pre-trained to align visual and textual features in a shared embedding space, enabling strong zero-shot performance. To efficiently adapt VLMs to diverse downstream tasks, pioneering works [80, 79] replace hand-crafted prompts with learnable continuous vectors optimized on training data. For unsupervised settings, earlier methods typically require access to multiple samples from either the entire test set [24, 66, 39] or streaming data [44, 73]. In contrast, episodic test-time adaptation (TTA) methods [61, 12, 60] perform prompt tuning using a single test instance, which is more challenging and has received increasing attention. A classic approach, test-time prompt tuning (TPT) [61] adapts CLIP during inference by minimizing the marginal entropy over multiple augmented predictions.

Although TPT [61] improves accuracy over zero-shot CLIP, it is prone to generating overconfident predictions, raising reliability concerns for VLMs in safety-critical applications [31, 52, 33, 62]. A pioneering work, C-TPT [75], shows that well-calibrated prompts tend to produce text embeddings with broader class-wise dispersion and thus introduces a distance-dispersion regularization term to push prompt embeddings away from the class centroid. Building on this insight, subsequent works [58, 1, 13] further propose various regularization terms to encourage angular separation between textual features. Although these regularization-based methods achieve lower expected calibration error (ECE) than the original TPT [61], we find they suffer from a noticeable accuracy drop, as shown in Fig. 1. We argue that reduced ECE is less meaningful when it comes at the expense of accuracy, especially as the underlying calibration mechanisms remain poorly interpretable.

Alternatively, SaLS [48] presents the only post-hoc calibration method that preserves accuracy by adjusting logits to remain within the zero-shot range after inference, but its calibration improvement remains limited. It is well known that CLIP [55] is generally well-calibrated [46, 15], with a relatively small gap between accuracy and confidence. Taking a closer look at the accuracy of CLIP and TPT, we find that prompt tuning using a single test instance does not substantially alter accuracy, suggesting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

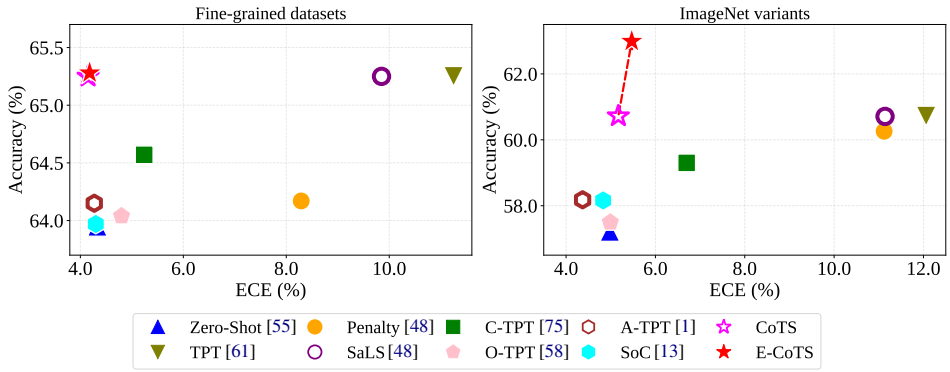


Figure 1: Average accuracy (%) and expected calibration error (ECE, %) of zero-shot CLIP, TPT, and various calibration methods on fine-grained datasets and ImageNet variants using ViT-B/16.

that the increased miscalibration stems from a shift in confidence. To validate this hypothesis, we conduct a pilot study that assigns zero-shot confidence to TPT predictions. We find that this straightforward replacement significantly reduces calibration error, surpassing [48, 75] while falling short of [58, 13]. However, this naive confidence replacement introduces a critical inconsistency: the predicted class comes from TPT's softmax distribution, whereas the confidence comes from the zero-shot distribution, potentially leading to unstable decisions.

Based on these insights, we propose **Co**nfidence-based **T**emperature **S**caling (`CoTS`), a simple yet effective calibration method that follows the classic idea of temperature scaling [16]. Specifically, `CoTS` optimizes a learnable temperature parameter to align TPT-predicted confidence with zero-shot confidence, thereby mitigating calibration error while retaining accuracy. Furthermore, inspired by prior works [10, 6] that ensemble augmented views to enhance prediction robustness, we introduce a new weak-strong ensemble strategy to further boost accuracy. In this strategy, we retain the weakly augmented view and balance its predictions with those from selected strong augmentations via a weighted trade-off. Empirically, we find that this ensemble strategy indeed boosts accuracy but may disrupt the model's well-calibrated properties. To mitigate this issue, we further integrate the ensemble strategy with `CoTS`, termed `E-CoTS`, achieving a better trade-off between accuracy and calibration. We validate the proposed methods on both fine-grained datasets and ImageNet variants.

To summarize, our contributions are as follows:

- We propose `CoTS`, a novel post-hoc confidence-based calibration method that uses temperature scaling to align TPT-predicted confidence with zero-shot confidence.
- We devise a simple weak-strong ensemble strategy integrated with `CoTS`, dubbed `E-CoTS`, that achieves reliable classification against TPT with superior accuracy and lower calibration error.
- Experiments across diverse datasets and backbones show that our methods reduce ECE to near the well-calibrated zero-shot baseline, while maintaining or even improving the adapted model's classification accuracy. Both `CoTS` and `E-CoTS` are plug-and-play and usually complement existing regularization-based calibration approaches, improving either accuracy or ECE.

2 Preliminaries and Motivation

CLIP-based zero-shot classification. CLIP [55] consists of a visual encoder $f_v(\cdot)$ and a text encoder $f_t(\cdot)$. For a K-class classification task with label space $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$, let x denote an input image and $y \in \mathcal{Y}$ its ground-truth label. The visual encoder maps x to an image embedding $v = f_v(x) \in \mathbb{R}^d$, while each class label $y_k \in \mathcal{Y}$ is converted into a textual prompt c_k (*e.g.*, using the template "a photo of a [CLASS]") and then encoded into a text embedding $t_k = f_t(c_k)$. Both the image and text embeddings are l_2-normalized, *i.e.*, $\|v\| = \|t_k\| = 1$. Then, the k-th values of the logit vector $\mathbf{z}$ and softmax probability vector $\mathbf{p}(x)$ corresponding to class y_k are defined as follows:

$$\mathbf{z}_k = (v^T \cdot t_k)/\tau_{\text{clip}}, \quad \mathbf{p}_k(x) = \mathbf{p}(y = k|x) = \frac{\exp(\mathbf{z}_k)}{\sum_{j=1}^{K} \exp(\mathbf{z}_j)}, \tag{1}$$

where the scaling factor $\tau_{\text{clip}} = 0.01$ is learned during pre-training [55].

Test-time prompt tuning (TPT). TPT [61] adapts CLIP [55] to a single test image by optimizing text prompts at test time, following the test-time adaptation paradigm [65, 78, 40]. Besides the weak augmented view $\mathcal{A}_1(x)$, TPT generates $(N-1)$ randomly augmented views of the test image using AugMix [20], and optimizes the input text prompts by minimizing the marginal entropy loss:

$$\mathcal{L}_{\text{TPT}} = -\sum_{k=1}^{K} \bar{\mathbf{p}}_k(x) \log \bar{\mathbf{p}}_k(x), \quad \bar{\mathbf{p}}_k(x) = \frac{1}{\rho N} \sum_{i=1}^{N} \mathbb{1}[\mathbf{H}(\mathbf{p}(\mathcal{A}_i(x))) \leq \gamma] \, \mathbf{p}_k(\mathcal{A}_i(x)). \tag{2}$$

Here $\bar{\mathbf{p}}_k(x)$ is the average softmax probability over selected augmentations that correspond to small entropy values, γ is the threshold corresponding to a cutoff percentile ρ (defaulting to 0.1 [61]), and $\mathbf{H}(\mathbf{p}(\mathcal{A}_i(x)))$ measures the self-entropy of the prediction on the i-th augmented view.

Temperature scaling. Temperature Scaling (TS) [16] is a classic post-hoc calibration technique [54] that adjusts model confidence without altering its predictions. It introduces a single temperature parameter $\tau > 0$ to rescale the logits before applying the softmax:

$$\hat{\mathbf{p}}_k(x;\tau) = \frac{\exp(\mathbf{z}_k/\tau)}{\sum_{j=1}^{K} \exp(\mathbf{z}_j/\tau)}, \quad k \in [1, \ldots, K], \tag{3}$$

where $\mathbf{z}_k$ is the logit for class k. A larger τ softens the distribution (reducing overconfidence), while a smaller τ sharpens it. The optimal temperature τ^* is typically obtained by minimizing the negative log-likelihood (NLL) on a labeled validation set $V = \{x_i, y_i\}_{i=1}^{|V|}$:

$$\tau^* = \arg\min_{\tau} -\sum_{i=1}^{|V|} \log \hat{\mathbf{p}}_{y_i}(x_i;\tau). \tag{4}$$

However, no labeled validation set is available in the episodic test-time adaptation problem, motivating alternative strategies for determining the temperature.

2.1 A Critical Examination of Prior Calibration Methods

Although TPT [61] improves accuracy compared to zero-shot CLIP [55], recent studies have shown that it produces overconfident predictions [48, 75]. To address this, a line of work has focused on calibrating TPT [75, 58, 17, 1, 13]. These methods introduce regularization terms during prompt optimization to encourage greater dispersion among text embeddings, *e.g.*, distance-based dispersion [75], orthogonal constraints [58], angular diversity [1], and semantic-aware separation [13]. Alternatively, SaLS [48] identifies the increase in logit range as the cause of miscalibration and adjusts logits to remain within the zero-shot range. To investigate whether these methods truly work as expected, we examine their ECE values and accuracies alongside those of CLIP and TPT in Fig. 1.

Observation ①: Calibration methods reduce TPT's calibration error but do not significantly outperform zero-shot CLIP. As depicted in Fig. 1, zero-shot predictions maintain low ECE values, confirming that CLIP models are generally well-calibrated [46, 15]. While TPT [61] clearly boosts accuracy, it leads to a substantial increase in ECE. Existing calibration methods significantly reduce the ECE of TPT, but most remain comparable to or slightly worse than zero-shot CLIP; only SoC [13] marginally outperforms it. The only post-hoc approach, SaLS [48], achieves notably weaker ECE reduction than its regularization-based counterparts.

Observation ②: Calibration methods improve accuracy over CLIP but often degrade the accuracy of TPT. As depicted in Fig. 1, all regularization-based methods decrease the accuracy of TPT. Specifically, the accuracy of O-TPT [58] and SoC [13] degrades substantially, even approaching that of zero-shot CLIP. In contrast, the post-hoc approach SaLS [48] fully preserves TPT's accuracy.

Taken together, zero-shot CLIP serves as a strong calibration baseline, while TPT provides a clear accuracy gain. Existing regularization-based methods trade accuracy for calibration, whereas SaLS [48] preserves accuracy but yields limited calibration gains. This reveals an unresolved gap: *no existing method preserves TPT's accuracy advantage while reducing calibration error to zero-shot level.*

2.2 Confidence Replacement: A Pilot Study

Comparing the accuracy of zero-shot CLIP and TPT, we find that adaptation using a single test instance typically yields limited accuracy gains, suggesting that TPT's miscalibration stems from a shift in confidence rather than from a change in the predicted class distribution. To test this hypothesis, we conduct a simple experiment: for each instance, we retain TPT's predicted class but replace its confidence (maximum predicted probability) with the corresponding zero-shot confidence, and then measure ECE. We refer to this procedure as Confidence Replacement (CoR).

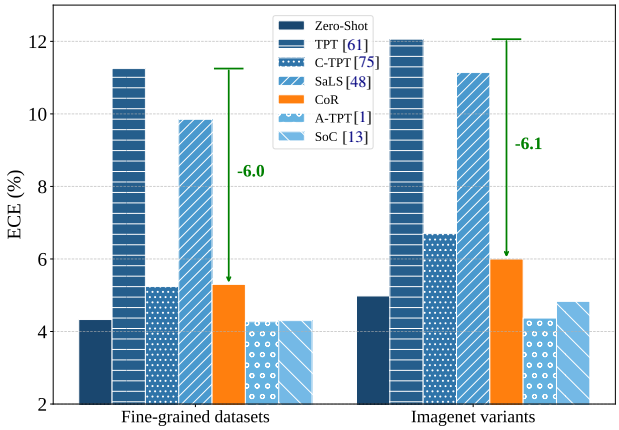


Figure 2: Comparison of average ECE (%) between CoR and existing calibration methods on fine-grained datasets and ImageNet variants using ViT-B/16.

As shown in Fig. 2, CoR substantially reduces ECE, outperforming SaLS [48] and C-TPT [75]. However, it remains inferior to methods such as SoC [13]. We attribute this to a fundamental inconsistency: the predicted class is derived from the TPT distribution, whereas the confidence comes from the zero-shot distribution, which may lead to incoherent decisions. Nevertheless, this experiment confirms two key findings: (1) the miscalibration of TPT is largely caused by confidence inflation, and (2) zero-shot confidence provides a useful reference for calibration. This reference remains effective even on datasets with relatively high zero-shot ECE or large accuracy gaps between zero-shot and TPT (Table 1 and Table 2). These observations motivate a principled approach that aligns TPT confidence with zero-shot confidence while maintaining a consistent probability distribution.

3 Methodology

Section 3.1 describes the proposed confidence-based temperature scaling method, while Section 3.2 then presents its enhanced variant under a new weak-strong ensemble strategy.

3.1 Confidence-based Temperature Scaling

Based on the findings in Section 2.2, we propose **Co**nfidence-based **T**emperature **S**caling (CoTS), a post-hoc calibration method that follows the classic idea of temperature scaling [16]. Rather than fitting the temperature by minimizing NLL on labeled validation data, CoTS leverages the zero-shot predictions as a label-free target. Specifically, CoTS learns a temperature τ to align TPT-predicted confidence with zero-shot confidence, thereby reducing calibration error while retaining accuracy.

Let $\mathbf{z}^{\texttt{tpt}}(x)$ and $\mathbf{z}^{\texttt{zs}}(x)$ denote the logits of the TPT-adapted model and the zero-shot model for the test image x, respectively. For a single augmented view, the objective is to minimize the squared difference between the temperature-scaled TPT confidence and the zero-shot confidence: $\left(\max_k \hat{\mathbf{p}}_k^{\texttt{tpt}}(x;\tau) - \max_k \mathbf{p}_k^{\texttt{zs}}(x)\right)^2$, where $\hat{\mathbf{p}}^{\texttt{tpt}}$ and $\mathbf{p}^{\texttt{zs}}$ are the corresponding softmax probabilities. To fully exploit the multiple augmentations available in TPT and reduce the sensitivity of relying on the weak view itself, we optimize over the set of selected strongly augmented views:

$$\tau^* = \arg\min_{\tau} \frac{1}{|\Omega|} \sum_{i\in\Omega} \left(\max_k \hat{\mathbf{p}}_k^{\texttt{tpt}}(\mathcal{A}_i(x);\tau) - \max_k \mathbf{p}_k^{\texttt{zs}}(\mathcal{A}_i(x)) \right)^2, \tag{5}$$

where $\Omega = \{i \mid i > 1,\ \mathbf{H}(\mathbf{p}^{\texttt{tpt}}(\mathcal{A}_i(x))) \leq \gamma'\}$ denotes the set of strongly augmented views selected by having low entropy after adaptation under the threshold γ', excluding the weakly augmented view. Once the optimal parameter τ^* is obtained, the calibrated probability is computed as $\hat{\mathbf{p}}^{\texttt{tpt}}(\mathcal{A}_1(x);\tau^*)$.

Compared to existing regularization-based calibration methods [75, 58, 1, 13], which modify the prompt optimization objective and may interfere with discriminative performance, CoTS operates entirely after adaptation and thus avoids any accuracy degradation.

3.2 Ensemble with Confidence-Based Temperature Scaling

While CoTS reduces calibration error without degrading the accuracy of TPT, it mainly preserves rather than improves the adapted prediction. We wonder whether we can further improve accuracy while maintaining good calibration, a goal that is especially important since existing calibration methods (*e.g.*, O-TPT [58], SoC [13]) often suffer from a notable accuracy degradation. Orthogonal to these calibration methods, ensembling over augmented views provides an effective way to boost accuracy during CLIP adaptation [10, 60, 6], and ZERO [10] theoretically shows that marginalizing predictions across augmented views can reduce the error bound compared with single-view inference.

Inspired by these insights, we propose a weak-strong ensemble strategy that further exploits the predictions from selected strongly augmented views to boost accuracy. Unlike prior work that uses majority voting [10] or uniform averaging [6] over augmented views, we balance the predictions from weak and strong augmentations using an adaptive weight:

$$\mathbf{p}_{\text{ens}} = \alpha \cdot \mathbf{p}^{\text{tpt}}(\mathcal{A}_1(x)) + (1-\alpha) \cdot \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathbf{p}^{\text{tpt}}(\mathcal{A}_i(x)), \tag{6}$$

where $\alpha = \frac{1}{K(K-1)} \sum_{i \neq j} t_i^T \cdot t_j$ is the average pairwise cosine similarity among the K normalized text embeddings $\{t_i\}_{i=1}^K$. Intuitively, a higher α indicates greater inter-class similarity in the text embedding space, which makes entropy-based view selection less discriminative (*i.e.*, the average of strong views becomes relatively unreliable). In such cases, the weakly augmented view (original image) is more reliable, and the ensemble assigns it a larger weight. Conversely, when text embeddings are well-separated (α is small), the entropy-filtered strongly augmented views provide more reliable complementary information. Note that α depends only on the text embeddings and thus varies across tasks and prompts, providing a task-adaptive weighting without per-sample tuning.

Through a preliminary study, we find that directly applying this ensemble on top of calibrated models indeed enhances accuracy but may disrupt their well-calibrated properties. To address this, we integrate the proposed confidence-based temperature scaling into the ensemble strategy. Specifically, we substitute each probability vector in Eq. (6) with its temperature-calibrated version:

$$\hat{\mathbf{p}}_{\text{ens}} = \alpha \cdot \hat{\mathbf{p}}^{\text{tpt}}(\mathcal{A}_1(x); \tau^*) + (1-\alpha) \cdot \frac{1}{|\Omega|} \sum_{i \in \Omega} \hat{\mathbf{p}}^{\text{tpt}}(\mathcal{A}_i(x); \tau^*). \tag{7}$$

The combined approach, termed E-CoTS, achieves a better trade-off between accuracy and calibration: the ensemble component enhances classification performance, while temperature scaling reduces calibration error. As an accuracy-enhancing extension, E-CoTS is designed to examine whether CoTS remains effective when the underlying predictor is strengthened through ensembling. Both methods are post-hoc, and their pseudocode is provided in Algorithm 1.

Algorithm 1: PyTorch-style code for CoTS and E-CoTS

```
# x:  input image (C, H, W); f_clip:  pretrained CLIP; A:  augmentation
# N: views nums; ρ:  selection ratio; T: steps; ensemble:  False/ True
def E-CoTS (x, f_clip, A, N, ρ, T):
    # step 1:  augment and select confident views
    views = A (x, num_views=N)                                   // (N, C, H, W)
    logits_zs = f_clip(views)                                    // (N, K)
    idx_zs = select_confident_views(logits_zs, top=ρ)
    # step 2:  test-time prompt tuning
    f_tpt = run_tpt_optimization(f_clip, views[idx_zs], steps=T)
    logits_tpt = f_tpt(views)
    # step 3:  confidence-based temperature scaling (CoTS, Eq.(5))
    idx_tpt = select_confident_views(logits_tpt[1:], top=ρ)
    τ = optimize_temperature(logits_tpt[idx_tpt], logits_zs[idx_tpt])
    p̂^tpt = (logits_tpt / τ).softmax(dim=1)
    if ensemble: # step 4:  weak-strong ensemble (E-CoTS, Eq.(7))
        α = compute_average_similarity()
        p̂_ecots = α · p̂^tpt[0] + (1 − α) · p̂^tpt[idx_tpt].mean(dim=0)
        return p̂_ecots.argmax(), p̂_ecots.max()
    else:
        p̂_cots = p̂^tpt[0]
        return p̂_cots.argmax(), p̂_cots.max()
```

Table 1: Performance comparison of different methods on fine-grained datasets using ViT-B/16. Accuracy (%) and ECE (%) are reported. In the Average column, green indicates performance improvement over TPT, while red indicates performance degradation.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Average
Accuracy	Zero-Shot [55]	66.72	44.33	67.32	83.66	62.58	23.85	88.20	93.96	65.19	42.05	65.56	63.95
	TPT [61]	68.91	47.12	68.64	84.65	65.50	23.27	87.25	94.08	68.20	42.91	66.58	65.19
	C-TPT [75]	68.45	45.17	69.58	83.14	64.53	24.10	88.20	93.75	65.05	42.45	65.80	64.57$_{-0.63}$
	Penalty [48]	68.84	45.80	68.36	84.43	65.44	22.75	82.94	93.27	67.09	41.20	66.19	64.21$_{-0.98}$
	SaLS [48]	68.91	47.12	68.67	84.65	65.49	23.29	87.25	94.08	68.21	42.91	66.57	65.20$_{+0.00}$
	O-TPT [58]	67.30	45.49	69.03	82.78	63.12	23.65	88.13	93.54	63.82	42.44	65.24	64.05$_{-1.14}$
	A-TPT [1]	67.92	45.17	69.79	83.06	63.20	23.80	87.89	93.24	64.03	42.05	65.20	64.12$_{-1.07}$
	SoC [13]	67.57	42.79	68.01	83.60	62.37	24.01	88.30	94.01	65.47	42.04	65.31	63.95$_{-1.24}$
	CoTS	68.91	47.12	68.64	84.65	65.50	23.27	87.25	94.08	68.20	42.91	66.58	65.19$_{+0.00}$
	E-CoTS	69.49	47.32	68.25	84.54	65.67	23.59	86.94	94.13	68.25	41.68	67.57	65.22$_{+0.03}$
ECE	Zero-Shot [55]	1.98	8.24	2.73	2.08	2.24	5.53	4.45	5.91	2.95	7.10	4.40	4.33
	TPT [61]	10.56	20.96	13.59	4.31	11.23	16.96	5.50	4.36	11.57	20.15	5.08	11.30
	C-TPT [75]	5.14	12.94	5.16	3.39	5.10	4.16	1.63	4.41	2.49	11.66	1.47	5.23$_{-6.07}$
	Penalty [48]	10.38	14.88	12.08	2.67	10.98	15.65	1.84	4.89	9.14	4.62	3.94	8.28$_{-3.02}$
	SaLS [48]	9.72	18.61	11.93	4.35	11.01	15.76	5.12	4.37	10.66	13.83	3.94	9.94$_{-1.36}$
	O-TPT [58]	2.03	7.66	3.68	4.28	8.42	3.76	2.13	4.53	2.70	11.52	1.81	4.77$_{-6.52}$
	A-TPT [1]	2.38	8.49	4.23	3.19	4.17	6.25	2.18	5.30	2.63	7.11	1.68	4.33$_{-6.97}$
	SoC [13]	3.71	7.63	2.98	2.84	2.84	5.41	1.83	5.80	2.71	7.08	4.63	4.31$_{-6.98}$
	CoTS	3.39	6.67	2.48	1.85	3.51	5.21	2.91	6.13	3.11	7.58	4.29	4.28$_{-7.01}$
	E-CoTS	3.00	6.94	4.10	1.29	2.59	5.73	3.43	4.74	3.50	8.23	3.84	4.31$_{-6.98}$

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Datasets. To comprehensively evaluate the performance of CoTS and E-CoTS, we first use a fine-grained benchmark consisting of eleven image classification datasets, including ImageNet [7], DTD [5], Flowers102 [50], Food101 [3], SUN397 [72], Aircraft [45], OxfordPets [53], Caltech101 [11], UCF101 [63], EuroSAT [19], and StanfordCars [34]. To further evaluate robustness under natural distribution shifts, we conduct experiments on four ImageNet variants, including ImageNet-A (natural adversarial examples) [22], ImageNet-V (re-collected images) [57], ImageNet-R (artistic renditions) [21], and ImageNet-K (sketch-style images with domain shifts) [69]. In the episodic test-time adaptation setting [9], each data instance is adapted independently of the others.

Baselines. We compare CoTS with representative calibration methods for fine-tuned CLIP [55]. These include regularization-based methods that encourage textual dispersion, such as C-TPT [75], O-TPT [58], A-TPT [1], and SoC [13]; logit-adjustment methods including ZS-Norm, Penalty, and SaLS [48]. Additionally, we evaluate whether our advanced E-CoTS can serve as a plug-and-play module on top of existing calibration methods [75, 58, 1, 13]. To examine generality beyond TPT methods, we further apply CoTS to other episodic TTA methods, including TTL [26] and TPS [64].

Implementation details. We use ViT-B/16 [55] as the backbone and follow the standard TPT [61] protocol. The initial text prompt "a photo of a [CLASS]" is optimized for one step using AdamW [43] with a learning rate of 0.005. For each test sample, we generate 64 augmented views using AugMix [20] and select confident views following TPT [61]. For temperature learning, we select the top 10% (ρ=0.1) most confident samples, excluding the weak augmentation (original image), and train the temperature parameter for 50 steps. We implement all baselines using their official code and report the average results over three random seeds [59]. *We provide additional results in the Appendix (*e.g.*, for ResNet-50, SigLIP [77], TDA [29], PLIP [25], standard deviations, and other metrics).*

4.2 MAIN RESULTS

Performance on fine-grained datasets. We evaluate the performance on fine-grained datasets in Table 1. CoTS preserves TPT's average accuracy while substantially improving calibration, reducing average ECE from 11.30% to 4.28%, which is the best among all methods. Compared to SaLS [48], the only post-hoc calibration baseline, CoTS achieves much lower ECE. Furthermore, E-CoTS improves the average accuracy and maintains a competitive ECE, demonstrating a better balance between classification performance and calibration.

Table 2: Performance comparison of different methods on ImageNet variants using ViT-B/16. Accuracy (%) and ECE (%) are reported.

Method	Accuracy (%)					ECE (%)				
	-A	-V	-R	-Sk	Average	-A	-V	-R	-Sk	Average
Zero-Shot [55]	47.83	60.94	73.99	46.10	57.22	8.34	3.18	3.55	4.87	4.98
TPT [61]	54.63	63.45	77.05	47.82	60.74	15.20	11.88	4.84	15.69	11.90
C-TPT [75]	51.19	62.58	75.79	47.43	59.25$_{-1.49}$	8.18	6.55	1.74	10.20	6.67$_{-5.24}$
Penalty [48]	53.07	63.40	76.68	47.79	60.24$_{-0.50}$	13.33	11.64	3.60	15.40	10.99$_{-0.91}$
SaLS [48]	54.63	63.44	77.05	47.82	60.74$_{+0.00}$	13.95	10.81	3.87	15.29	10.98$_{-0.92}$
O-TPT [58]	48.12	61.35	73.98	46.59	57.51$_{-3.23}$	6.68	3.45	3.98	5.76	4.97$_{-6.94}$
A-TPT [1]	49.16	61.77	74.99	46.89	58.20$_{-2.54}$	5.55	2.85	4.09	4.87	4.34$_{-7.56}$
SoC [13]	49.57	61.36	75.11	46.45	58.12$_{-2.62}$	8.15	3.65	4.50	3.16	4.87$_{-7.04}$
CoTS	54.63	63.45	77.05	47.82	60.74$_{+0.00}$	7.85	4.37	5.09	3.87	5.30$_{-6.61}$
E-CoTS	60.72	64.49	77.88	48.72	62.95$_{+2.22}$	8.71	3.77	2.94	6.11	5.38$_{-6.52}$

Table 3: Compatibility of CoTS, and E-CoTS with different base methods using ViT-B/16. Average accuracy (%) and ECE (%) are reported on fine-grained datasets and ImageNet variants, respectively.

(%)	Method	Fine-grained datasets					ImageNet variants				
		C-TPT [75]	Penalty [48]	O-TPT [58]	A-TPT [1]	SoC [13]	C-TPT [75]	Penalty [48]	O-TPT [58]	A-TPT [1]	SoC [13]
ACC	base	64.57	64.21	64.05	64.12	63.95	59.25	60.24	57.51	58.20	58.12
	+ E-CoTS	64.64	65.25	64.93	65.07	65.01	62.74	62.65	62.01	62.32	61.84
ECE	base	5.23	8.28	4.77	4.33	4.31	6.67	10.99	4.97	4.34	4.87
	+ SaLS [48]	4.97	9.04	4.56	4.25	4.26	6.17	11.06	4.38	4.80	5.02
	+ CoTS	4.53	3.90	4.55	4.38	4.44	4.75	5.14	4.14	4.43	4.37
	+ E-CoTS	4.16	3.86	4.30	3.85	4.20	5.14	5.39	4.55	4.87	4.58

Performance on ImageNet variants. The results on ImageNet variants are shown in Table 2. It is clear that regularization-based methods (O-TPT [58], A-TPT [1], SoC [13]) reduce ECE but at the cost of accuracy. In contrast, CoTS preserves TPT's accuracy while more than halving the average ECE. Moreover, E-CoTS achieves the best accuracy of 62.95% with a competitive ECE of 5.38%.

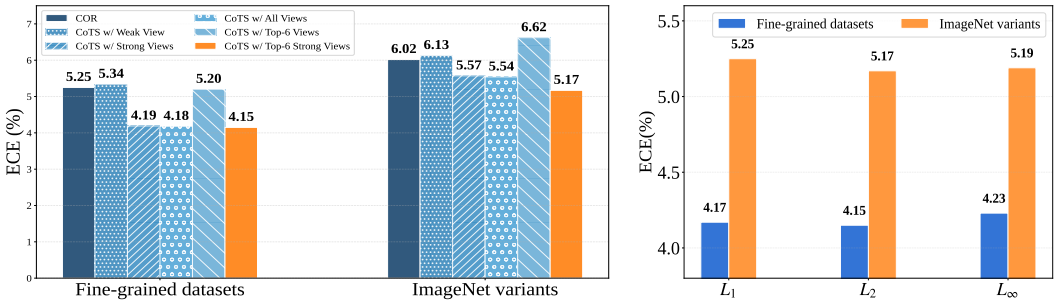
Validity of zero-shot confidence as a calibration reference. Although CoTS uses zero-shot confidence as a reference, it does not require the zero-shot model to be perfectly calibrated. As shown in Table 1 and Table 2, CoTS remains effective on datasets with high zero-shot ECE, such as DTD and EuroSAT, and large accuracy gaps between zero-shot and TPT, such as UCF101 and SUN397. This demonstrates that zero-shot confidence serves as a useful reference rather than an oracle target.

Plug-and-play compatibility with calibration methods. Table 3 evaluates the integration of CoTS and E-CoTS with existing calibration methods. Overall, E-CoTS improves accuracy across all base methods, while both CoTS and E-CoTS generally reduce ECE, particularly for Penalty [48] and C-TPT [75]. This demonstrates their plug-and-play compatibility with existing methods.

4.3 Analysis

We conduct sensitivity analyses across both methodological components and experimental configurations. At the component level, we investigate the choice of augmented views and confidence alignment loss in CoTS, as well as the ensemble weight in E-CoTS. We further evaluate the robustness of our methods under different prompt initializations, episodic TTA frameworks, and visual backbones. A component-level comparison with ensemble-only is provided in Appendix D.6.

View selection in CoTS. We analyze how different augmented view selections affect temperature learning (Fig. 3a). Using only the weak view yields limited improvement and even higher ECE than CoR, suggesting insufficient confidence variation for reliable temperature estimation. In contrast, using strong augmented views substantially reduces ECE, indicating that strong augmentations offer more informative confidence variations. However, selecting the Top-6 views indiscriminately remains



(a) View selection for CoTS. (b) Loss function for CoTS.

Figure 3: Ablation studies on key components (*i.e.*, view selection and loss function) of CoTS with ViT-B/16. Average ECE (%) is reported on fine-grained datasets and ImageNet variants.

Table 4: Performance comparison of different methods initialized with CoOp [80] using ViT-B/16. We report average accuracy (%) and ECE (%) on fine-grained datasets and ImageNet variants, and evaluate the effect of integrating CoTS and E-CoTS.

(%)	Method	Fine-grained datasets					ImageNet variants				
		TPT [61]	C-TPT [75]	O-TPT [58]	A-TPT [1]	SoC [13]	TPT [61]	C-TPT [75]	O-TPT [58]	A-TPT [1]	SoC [13]
ACC	base	65.67	65.12	64.58	64.68	64.23	63.08	61.51	59.16	60.48	61.45
	+ E-CoTS	65.47	65.70	65.60	65.54	65.19	64.45	64.15	63.36	63.88	64.24
ECE	base	17.46	10.56	8.43	9.29	5.61	19.95	15.38	9.92	12.20	10.22
	+ SALS [48]	13.83	6.76	5.38	6.05	5.44	16.00	9.94	5.46	7.26	8.13
	+ CoTS	5.61	5.22	5.18	5.11	5.23	5.06	4.62	4.68	4.65	4.44
	+ E-CoTS	6.10	5.37	5.30	5.12	4.98	6.44	5.26	4.26	4.80	4.80

suboptimal, as they may include weak or overly confident views with limited calibration signals. Overall, CoTS with the Top-6 strong views achieves the lowest ECE on both dataset groups.

Loss function in CoTS. We further compare different loss functions for confidence alignment in Fig. 3b. The results show that all three losses achieve comparable calibration performance, indicating that CoTS is not highly sensitive to the specific loss formulation. Among them, the $\mathcal{L}_2$ loss obtains the lowest ECE on both fine-grained datasets and ImageNet variants. Therefore, we adopt $\mathcal{L}_2$ as the default objective for temperature learning in our experiments.

Ensemble weight α in E-CoTS. We study the influence of the ensemble weight α in E-CoTS, with results on ImageNet variants shown in Fig. 4. Fixed weights generally underperform the proposed adaptive strategy across a range of tested values, while consistent results on fine-grained datasets are provided in Appendix D.2. The adaptive strategy achieves a favorable balance between accuracy and calibration, avoiding the need for manual weight selection.

Compatibility with CoOp initialization. We evaluate our method with CoOp [80] initialization in Table 4. Compared to existing regularization-based methods, CoTS consistently achieves much lower ECE across both dataset groups. Meanwhile, E-CoTS generally improves accuracy over the corresponding base methods, showing that the weak-strong ensemble remains beneficial under CoOp [80] initialization. Overall, our method supports stronger initialization and provides stronger calibration performance. Sensitivity to hand-crafted prompts is reported in Appendix D.1.

Generalization to other episodic TTA methods. We further evaluate CoTS and E-CoTS on other episodic TTA methods. The results in Table 5 show that our method generalizes well beyond TPT-based approaches. In particular, CoTS consistently brings large reductions in ECE on both datasets, substantially outperforming the corresponding base models and also achieving stronger calibration than SaLS [48]. Furthermore, E-CoTS maintains comparable, and sometimes improved accuracy over the base methods. These results indicate that our approach is not limited to a specific adaptation framework, but can serve as an effective calibration strategy for other episodic TTA methods as well.

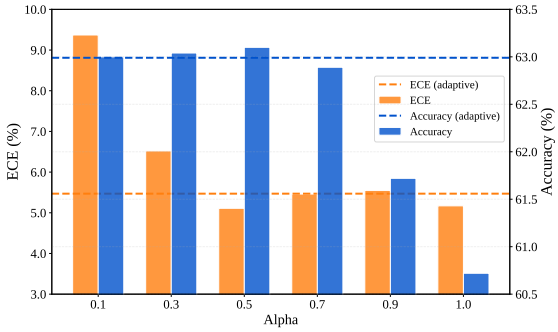


Figure 4: Effect of ensemble weight α on ImageNet variants using ViT-B/16. Accuracy (%) and ECE (%) are reported

Table 5: Evaluation of SaLS [48], CoTS, and E-CoTS on TTL [26] and TPS [64] using ViT-B/16. Average accuracy (%) and ECE (%) on fine-grained datasets and ImageNet variants are reported.

(%)	Method	Fine-grained datasets		ImageNet variants	
		TTL [26]	TPS [64]	TTL [26]	TPS [64]
ACC	base	64.48	64.99	62.46	61.69
	+ E-CoTS	64.23	65.11	62.39	61.71
ECE	base	28.60	16.79	34.30	27.16
	+ SaLS [48]	22.15	13.69	28.94	23.98
	+ CoTS	9.82	4.87	9.42	6.67
	+ E-CoTS	9.42	5.02	8.66	6.65

Table 6: Performance using ResNet-50. Average accuracy (%) and ECE (%) are reported on fine-grained datasets and ImageNet variants.

Dataset	Metric	Zero-Shot [55]	TPT [61]	C-TPT [75]	Penalty [48]	SaLS [48]	O-TPT [58]	A-TPT [1]	SoC [13]	CoTS	E-CoTS
Fine-grained datasets	ACC	56.04	58.06	57.69	57.52	58.06	57.36	57.30	56.19	58.06	57.72
	ECE	5.38	11.31	6.44	8.41	9.31	5.67	4.55	5.50	5.04	5.04
ImageNet variants	ACC	40.70	43.83	42.75	43.66	43.83	40.82	41.94	41.61	43.83	45.39
	ECE	7.09	17.55	13.14	17.02	15.37	6.46	7.76	7.52	6.87	7.24

Generalization to ResNet-50. As shown in Table 6, CoTS preserves TPT's average accuracy while substantially reducing ECE on both dataset groups. Moreover, E-CoTS improves accuracy on ImageNet variants while maintaining low ECE. These results demonstrate that our approach generalizes beyond ViT-B/16. Full results are provided in Appendix C.2.

Table 7: Average Brier score (BS, %) [4] and class-wise ECE (CECE, %) [35] using ViT-B/16. Lower values indicate better calibration.

Dataset	Metric	Zero-Shot [55]	TPT [61]	C-TPT [75]	Penalty [48]	SaLS [48]	O-TPT [58]	A-TPT [1]	SoC [13]	CoTS	E-CoTS
Fine-grained datasets	BS	14.19	16.41	14.84	15.86	15.84	14.58	14.48	14.20	14.44	14.25
	CECE	1.03	1.05	1.08	0.98	0.99	1.09	1.02	1.02	0.96	0.95
ImageNet variants	BS	17.16	19.75	18.13	19.74	19.30	17.03	17.20	17.16	17.59	17.54
	CECE	0.15	0.14	0.14	0.14	0.14	0.15	0.15	0.15	0.14	0.13

Evaluation with additional calibration metrics. We further evaluate calibration using the Brier score [4] and class-wise ECE [35]. As shown in Table 7, both CoTS and E-CoTS substantially reduce the average BS of TPT across both dataset groups. Full results are provided in Appendix C.7.

5 Conclusion

This paper studies the calibration problem in test-time prompt tuning from the perspective of confidence correction. We observe that existing calibration-oriented methods often reduce ECE at the cost of classification accuracy. To address this trade-off, we propose CoTS, a post-hoc confidence-based temperature scaling method that uses zero-shot CLIP predictions as a calibration anchor. By only rescaling logits, CoTS improves calibration performance without changing the discriminative benefits of TPT. We further introduce a weak-strong ensemble strategy and integrate it into CoTS, resulting in E-CoTS, which improves accuracy while maintaining reliable calibration. Extensive experiments across multiple datasets and backbones validate that our methods effectively mitigate miscalibration without compromising primary accuracy.

AI USE STATEMENT

Generative AI tools were used to polish the writing and improve the clarity of the manuscript. All AI-assisted edits were reviewed and revised by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

We provide implementation details, hyperparameter settings, and method pseudocode in the main paper. Additional experiments, standard deviations, calibration metrics, and sensitivity analyses are included in the Appendix. We also release an anonymized code repository at https://anonymous.4open.science/r/CoTS/.

REFERENCES

[1] Shihab Aaqil Ahamed, Udaya Sampath K. Perera Miriya Thanthrige, Ranga Rodrigo, and Muhammad Haris Khan. A-TPT: Angular diversity calibration properties for test-time prompt tuning of vision-language models. In *Proc. ICLR*, 2026.

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proc. ICML*, pp. 1613–1622, 2015.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proc. ECCV*, pp. 446–461, 2014.

[4] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proc. CVPR*, pp. 3606–3613, 2014.

[6] Konstantinos M. Dafnis and Dimitris N. Metaxas. Test-time spectrum-aware latent steering for zero-shot generalization in vision-language models. In *Proc. NeurIPS*, pp. 151169–151194, 2025.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, pp. 248–255, 2009.

[8] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *Proc. ICCV*, pp. 6889–6899, 2021.

[9] Hao Dong, Lijun Sheng, Jian Liang, Ran He, Eleni Chatzi, and Olga Fink. Adapting vision-language models without labels: A comprehensive survey. *arXiv preprint arXiv:2508.05547*, 2025.

[10] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Proc. NeurIPS*, pp. 129062–129093, 2024.

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proc. CVPR Workshops*, pp. 178–178, 2004.

[12] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proc. ICCV*, pp. 2704–2714, 2023.

[13] Leo Fillioux, Omprakash Chakraborty, Ismail Ben Ayed, Paul-Henry Cournède, Stergios Christodoulidis, Maria Vakalopoulou, and Jose Dolz. Soc: Semantic orthogonal calibration for test-time prompt tuning. In *Proc. CVPR*, 2026.

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, pp. 1050–1059, 2016.

[15] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers? In *Proc. ICLR*, 2023.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proc. ICML*, pp. 1321–1330, 2017.

[17] Jisu Han and Wonjun Hwang. D-tpt: Dimensional entropy maximization for calibrating test-time prompt tuning in vision-language models. *arXiv preprint arXiv:2510.09473*, 2025.

[18] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024.

[19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS*, pp. 204–207, 2018.

[20] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Proc. ICLR*, 2020.

[21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. ICCV*, pp. 8340–8349, 2021.

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proc. CVPR*, pp. 15262–15271, 2021.

[23] Dapeng Hu, Jian Liang, Xinchao Wang, and Chuan-Sheng Foo. Pseudo-calibration: Improving predictive uncertainty estimation in unsupervised domain adaptation. In *Proc. ICML*, pp. 19304–19326, 2024.

[24] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[25] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[26] Raza Imam, Hanan Gani, Muhammad Huzaifa, and Karthik Nandakumar. Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models. In *Proc. WACV*, pp. 5449–5459, 2025.

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*, pp. 4904–4916, 2021.

[28] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proc. ECCV*, pp. 709–727, 2022.

[29] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proc. CVPR*, pp. 14162–14171, 2024.

[30] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

[31] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proc. CVPR*, pp. 14829–14838, 2022.

[32] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proc. CVPR*, pp. 19113–19122, 2023.

[33] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proc. CVPR*, pp. 14766–14776, 2025.

[34] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proc. ICCV Workshops*, pp. 554–561, 2013.

[35] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Proc. NeurIPS*, pp. 12316–12326, 2019.

[36] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, pp. 6405–6416, 2017.

[37] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proc. EMNLP*, pp. 3045–3059, 2021.

[38] Vladislav Lialin, Vijeta Deshpande, Xiaowei Yao, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

[39] Jian Liang, Lijun Sheng, Zhengbo Wang, Ran He, and Tieniu Tan. Realistic unsupervised clip fine-tuning with universal entropy optimization. In *Proc. ICML*, pp. 29667–29681, 2024.

[40] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *Int. J. Comput. Vis.*, 133(1):31–64, 2025.

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. ICCV*, pp. 2980–2988, 2017.

[42] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proc. CVPR*, pp. 80–88, 2022.

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2019.

[44] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: test-time prompt adaptation for vision-language models. In *Proc. NeurIPS*, pp. 65252–65264, 2023.

[45] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[46] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Proc. NeurIPS*, pp. 15682–15694, 2021.

[47] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Proc. NeurIPS*, pp. 4694–4703, 2019.

[48] Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *Proc. ECCV*, pp. 147–165, 2024.

[49] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proc. AAAI*, pp. 2901–2907, 2015.

[50] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proc. ICVGIP*, pp. 722–729, 2008.

[51] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proc. CVPR Workshops*, 2019.

[52] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proc. CVPR*, pp. 14760–14769, 2024.

[53] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proc. CVPR*, pp. 3498–3505, 2012.

[54] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pp. 8748–8763, 2021.

[56] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.

[57] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proc. ICML*, pp. 5389–5400, 2019.

[58] Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *Proc. CVPR*, pp. 19942–19951, 2025.

[59] Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. In *Proc. NeurIPS D&B Track*, 2025.

[60] Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proc. CVPR*, pp. 29958–29967, 2025.

[61] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Proc. NeurIPS*, pp. 14274–14289, 2022.

[62] Julio Silva-Rodríguez, Fereshteh Shakeri, Houda Bahig, Jose Dolz, and Ismail Ben Ayed. Few-shot, now for real: Medical vlms adaptation without balanced sets or validation. In *Proc. MICCAI*, pp. 237–247, 2025.

[63] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[64] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *Proc. WACV*, pp. 825–835, 2025.

[65] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proc. ICML*, pp. 9229–9248, 2020.

[66] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *Proc. ICML*, pp. 33816–33832, 2023.

[67] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: improved calibration and predictive uncertainty for deep neural networks. In *Proc. NeurIPS*, pp. 13911–13922, 2019.

[68] Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023.

[69] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. In *Proc. NeurIPS*, pp. 10506–10518, 2019.

[70] Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned clip. In *Proc. ICML*, pp. 51734–51754, 2024.

[71] Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Understanding and mitigating miscalibration in prompt tuning for vision-language models. In *Proc. ICML*, pp. 63467–63489, 2025.

[72] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, pp. 3485–3492, 2010.

[73] Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Cheems Wang, and Cees G. M. Snoek. Dynaprompt: Dynamic test-time prompt tuning. In *Proc. ICLR*, 2025.

[74] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proc. CVPR*, pp. 6757–6767, 2023.

[75] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A. Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *Proc. ICLR*, 2024.

[76] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proc. CVPR*, pp. 23783–23793, 2024.

[77] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. ICCV*, pp. 11941–11952, 2023.

[78] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: test time robustness via adaptation and augmentation. In *Proc. NeurIPS*, pp. 38629–38642, 2022.

[79] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proc. CVPR*, pp. 16816–16825, 2022.

[80] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022.

APPENDIX CONTENTS

A Limitations and Broader Impact **16**

A.1 Limitations 16

A.2 Broader Impact 16

B Related Work **16**

C Additional Experimental Results **17**

C.1 Computational Efficiency 17

C.2 Results with ResNet-50 17

C.3 Results with Online TTA Framework 17

C.4 Results with SigLIP 17

C.5 Results on Medical Datasets 18

C.6 Standard Deviations 18

C.7 Calibration Metrics 18

D Additional Analysis **21**

D.1 Sensitivity to Hand-crafted Prompt Initialization 21

D.2 Ensemble Weight alpha in E-CoTS 21

D.3 Similarity-based Analysis 21

D.4 Step-wise Analysis 22

D.5 Sensitivity to K and T_0 22

D.6 Comparison between Ensemble-Only and E-CoTS 22

D.7 Comparison with ZERO and Our Ensemble Strategy 22

E Reliability Diagram **27**

A Limitations and Broader Impact

A.1 Limitations

While our methods effectively improve calibration for test-time prompt tuning, they are currently evaluated on classification tasks and introduce additional computational overhead. Extending confidence-oriented temperature scaling to other vision-language tasks, such as semantic segmentation, object detection, or adversarial robustness, remains to be explored.

A.2 Broader Impact

This work aims to improve the reliability of vision-language models by mitigating the miscalibration issue in test-time prompt tuning. The proposed methods achieve a favorable balance between accuracy and calibration error. Moreover, our approaches can be readily integrated with existing calibration techniques, highlighting their potential as general plug-and-play modules for test-time adaptation. We hope this work provides useful insights for future research on calibration in vision-language models and encourages the development of more reliable foundation models.

B Related Work

Prompt Tuning in Vision-Language Models. Vision-Language Models (VLMs) such as CLIP [55] and ALIGN [27] learn a shared embedding space that aligns visual features with corresponding textual descriptions, enabling zero-shot transfer to a variety of downstream classification tasks. To efficiently adapt VLMs to diverse downstream tasks, many parameter-efficient fine-tuning (PEFT) methods [38, 18] have proven effective by optimizing only a small subset of parameters while keeping the rest of the model frozen. One widely used PEFT technique is prompt tuning [37, 28], which adds trainable tokens either to the input or to intermediate layers. In particular, CoOp [80] pioneers prompt learning for VLMs by replacing fixed templates with learnable continuous vectors. Subsequently, CoCoOp [79] incorporates visual cues to generate instance-specific prompts, while MaPLE [32] further enhances context tuning by applying learnable prompts to both the image and text encoders. To avoid relying on annotated downstream training data during adaptation, several methods [24, 66, 39] explore unsupervised prompt learning with pre-trained VLMs. TPT [61] focuses on an interesting unsupervised scenario, where the model is adapted using only a single instance at test time. Such a challenging paradigm has gained increasing attention [26, 10, 76, 64, 60], and our work also studies this problem but focuses more on calibration performance.

Calibration of Deep Neural Networks. In real-world safety-critical decision systems, classification networks must not only be accurate but also indicate when they are likely to be incorrect. Confidence calibration [16, 51] formalizes this by measuring how well a model's predicted confidence aligns with its actual accuracy. A recent survey [68] categorizes existing calibration methods into post-hoc, regularization, and uncertainty estimation approaches. Post-hoc calibration methods [16, 35, 8] adjust predictions post-training without modifying parameters. A popular example is temperature scaling (TS) [16], which scales logits using a temperature parameter optimized on a validation set.Later methods [35, 8] extend TS to improve robustness under limited validation data, multi-label tasks, and distribution shifts. By contrast, regularization-based calibration methods [16, 42] always incorporate additional regularization objectives during model training. These include explicit penalties on over-confident predictions [16, 42] and implicit techniques (*e.g.*, label smoothing [47], mixup [67], focal loss [41]) that improve calibration and generalization via softened targets. Moreover, uncertainty estimation methods alleviate miscalibration by introducing randomness via Bayesian neural networks [2], ensembles [36], and Monte Carlo dropout [14]. Most existing post-hoc calibration methods require a labeled validation set. One recent approach [23] uses mixup synthesis over unlabeled samples to generate pseudo-labeled data for temperature scaling, whereas our method operates using only a single unlabeled instance. We further provide a hybrid calibration strategy that combines temperature scaling with an output-level ensemble.

Calibration of Vision-Language Models. Although prompt tuning methods are primarily developed to improve accuracy, recent works [70, 71, 48, 75] have also examined the calibration performance of VLMs, particularly CLIP [55]. Specifically, existing few-shot prompt tuning methods are found to lead to a calibration trade-off between base and new classes [70], *e.g.*, CoOp [80] causes overconfidence

on new classes, whereas KgCoOp [74] leads to underconfidence on base classes. Meanwhile, a classic unsupervised prompt tuning method, TPT [61], improves accuracy but incurs a high calibration error after test-time adaptation. Since our method belongs to the unsupervised category, we primarily review prior work within this branch of the literature. C-TPT [75] finds that well-calibrated prompts yield text embeddings with broader class-wise dispersion and, on this basis, proposes a dispersion-based regularization term to encourage embeddings to move away from the class centroid. To fully exploit the textual feature space, several recent methods (*i.e.*, O-TPT [58], A-TPT [1], and SoC [13]) encourage angular separation between textual features to promote greater dispersion. Alternatively, D-TPT [17] identifies and mitigates the influence of dominant feature dimensions to improve calibration performance. Unlike these regularization-based methods, we follow TS [16] and propose a simple yet efficient post-hoc calibration method. The most closely related work to ours is SaLS [48], which adjusts logits within the zero-shot range to maintain reliable confidence. Both methods are built on the observation that zero-shot predictions are relatively well-calibrated. However, unlike SaLS [48], our method learns a temperature parameter to bridge the confidence gap before and after adaptation.

C Additional Experimental Results

C.1 Computational Efficiency

We measure the computation time of different methods using a single 48GB GPU (NVIDIA RTX A6000). As shown in Table 8, integrating `CoTS` delivers effective calibration improvement without substantially increasing overall inference cost. These results show that `CoTS` offers an efficient calibration strategy for test-time prompt tuning.

C.2 Results with ResNet-50

We further evaluate different methods with ResNet-50 to verify the robustness of our approach across architectures. The results in fine-grained datasets and ImageNet variants are shown in Table 9 and Table 10.

C.3 Results with Online TTA Framework

To assess the applicability of our approach beyond the standard episodic setting, we integrate `CoTS` and `E-CoTS` with TDA [29], an online TTA framework. As shown in Table 11 and Table 12, `CoTS` substantially improves calibration, while `E-CoTS` achieves a favorable accuracy–calibration trade-off.

C.4 Results with SigLIP

We also investigate whether the proposed methods remain effective with a different vision-language model by conducting experiments with SigLIP [77]. Table 13 and Table 14 show that `CoTS` and `E-CoTS` consistently provide competitive accuracy and improved calibration.

Table 8: Computation time (s) of different methods on DTD and ImageNet-A using ViT-B/16. We report the average inference time per sample.

Dataset	TPT [61]	C-TPT [75]	O-TPT [58]	A-TPT [1]	SoC [13]	SaLS [48]	`CoTS`
DTD	0.192	0.193	0.206	0.207	0.206	+0.000	+0.053
ImageNet-A	0.223	0.224	0.258	0.258	0.256	+0.000	+0.054

Table 9: Performance comparison on fine-grained datasets using ResNet-50. Accuracy (%) and ECE (%) are reported.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg.
Accuracy	Zero-Shot [55]	58.17	40.37	61.67	73.94	58.84	15.75	83.54	85.88	58.84	23.67	55.78	56.04
	TPT [61]	60.68	41.51	62.57	74.99	61.36	17.69	84.43	87.87	60.67	28.42	58.47	58.06
	C-TPT [75]	60.47	41.25	65.03	74.84	60.93	17.21	83.66	87.37	60.25	27.18	56.38	57.69
	Penalty [48]	60.64	41.31	62.66	74.70	61.25	17.36	83.94	87.56	60.08	25.67	57.56	57.52
	SaLS [48]	60.69	41.51	62.57	74.99	61.36	17.68	84.43	87.86	60.67	28.42	58.47	58.06
	O-TPT [58]	58.99	41.65	65.58	74.64	59.61	17.03	83.19	87.07	59.81	27.73	55.64	57.36
	A-TPT [1]	59.91	41.86	64.20	74.29	59.31	16.00	82.68	86.88	59.81	29.14	56.17	57.30
	SoC [13]	59.93	40.37	61.90	73.82	59.04	15.67	83.28	85.84	58.90	23.64	55.68	56.19
	CoTS	60.68	41.51	62.57	74.99	61.36	17.69	84.43	87.87	60.67	28.42	58.47	**58.06**
	E-CoTS	60.90	41.33	60.71	74.27	61.33	17.73	84.12	87.69	60.47	27.45	58.96	57.72
ECE	Zero-Shot [55]	2.01	9.04	3.03	2.71	3.81	6.30	5.70	4.41	3.00	14.76	4.40	5.38
	TPT [61]	11.37	25.80	13.63	5.21	9.17	15.66	3.94	3.74	11.03	21.10	3.76	11.31
	C-TPT [75]	6.75	21.65	3.81	1.76	3.01	10.69	2.79	2.61	3.02	13.37	1.33	6.44
	Penalty [48]	11.29	16.96	11.94	3.34	8.98	13.39	3.25	3.66	10.08	6.92	2.65	8.41
	SaLS [48]	9.91	21.66	11.99	3.93	8.66	15.29	3.07	4.02	8.98	12.32	2.54	9.31
	O-TPT [58]	3.11	16.63	2.40	1.28	6.57	8.18	3.35	2.85	2.35	13.44	2.26	5.67
	A-TPT [1]	2.32	15.35	2.71	1.55	4.22	8.26	2.72	4.16	2.63	4.71	1.43	4.55
	SoC [13]	3.02	8.95	3.34	3.63	3.89	6.36	5.61	4.53	2.37	14.79	4.02	5.50
	CoTS	3.74	7.97	2.62	3.74	5.24	4.88	5.29	5.84	2.79	7.33	6.05	5.04
	E-CoTS	2.36	9.47	6.80	2.18	3.36	5.40	3.89	4.40	3.67	8.54	5.42	**5.04**

Table 10: Performance comparison of different methods on ImageNet variants using ResNet-50. Accuracy (%) and ECE (%) are reported.

Method	Accuracy (%)					ECE (%)				
	-A	-V	-R	-Sk	Avg.	-A	-V	-R	-Sk	Avg.
Zero-Shot [55]	21.87	51.45	56.12	33.36	40.70	21.24	3.03	0.98	3.11	7.09
TPT [61]	26.51	54.63	59.04	35.13	43.83	30.96	13.94	10.47	14.84	17.55
C-TPT [75]	24.27	54.21	57.77	34.75	42.75	26.61	9.35	5.51	11.08	13.14
Penalty [48]	26.12	54.58	58.80	35.12	43.66	29.85	13.80	9.73	14.70	17.02
SaLS [48]	26.52	54.63	59.04	35.13	43.83	28.10	12.27	7.79	13.32	15.37
O-TPT [58]	21.59	52.37	56.11	33.20	40.82	19.23	2.38	1.52	2.70	**6.46**
A-TPT [1]	22.56	53.26	57.61	34.32	41.94	20.28	4.11	0.94	5.70	7.76
SoC [13]	22.60	53.28	56.85	33.72	41.61	19.38	4.24	2.47	3.97	7.52
CoTS	26.51	54.63	59.04	35.13	43.83	17.52	4.55	3.07	2.33	6.87
E-CoTS	31.80	55.20	58.90	35.65	**45.39**	17.75	4.60	1.72	4.89	7.24

C.5 Results on Medical Datasets

We further evaluate CoTS and E-CoTS on the COVID [56] and Kather [30] medical datasets using the domain-specific PLIP [25] backbone (Table 26), where E-CoTS consistently reduces ECE over CoTS while retaining comparable accuracy.

C.6 Standard Deviations

We report standard deviations of accuracy and ECE across three seeds in Table 15 and Table 16. On fine-grained datasets and ImageNet, CoTS achieves consistently low variance, indicating stable performance. On ImageNet variants, both CoTS and E-CoTS remain robust across seeds, with E-CoTS maintaining strong calibration despite minor variability. Overall, the results confirm that the proposed methods are reliable and stable while improving calibration under different runs.

C.7 Calibration Metrics

Calibration error. To quantify model calibration [49, 16], we primarily adopt the Expected Calibration Error (ECE) [49], which measures the discrepancy between predicted probabilities and observed frequencies. Specifically, the predictions are partitioned into M confidence bins $\{B_1, B_2, \ldots, B_M\}$

Table 11: Performance comparison of different methods integrated with an online TTA framework (TDA [29]) on fine-grained datasets using ViT-B/16. Accuracy (%) and ECE (%) are reported.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Average
Accuracy	TDA [29]	68.27	46.04	69.75	83.68	65.01	24.06	88.36	94.04	67.99	51.04	66.55	65.65
	C-TPT [75]	69.15	47.22	70.69	83.42	65.52	24.30	88.69	94.08	67.64	47.67	67.01	65.62
	Penalty [48]	68.26	46.04	69.79	83.67	65.09	24.12	88.36	94.00	67.99	51.04	66.57	65.90
	SaLS [48]	68.27	46.04	69.75	83.68	65.01	24.06	88.36	94.04	67.99	51.04	66.55	65.65
	O-TPT [58]	68.49	48.11	70.77	83.15	64.93	24.09	88.58	94.20	66.98	47.06	66.56	65.44
	A-TPT [1]	68.83	48.11	71.09	83.22	65.08	23.97	88.55	93.43	67.01	51.04	66.53	65.80
	SoC [13]	68.59	45.21	69.71	83.66	64.82	24.24	88.61	94.20	68.17	51.04	66.55	65.62
	CoTS	68.27	46.04	69.75	83.68	65.01	24.06	88.36	94.04	67.99	51.04	66.55	65.65
	E-CoTS	70.25	47.16	69.35	84.58	66.96	25.02	88.80	94.24	69.13	51.52	68.85	**66.56**
ECE	TDA [29]	5.27	17.02	7.76	2.21	5.78	13.98	2.81	3.32	5.88	7.05	1.95	6.78
	C-TPT [75]	8.80	20.09	11.35	1.68	7.11	12.54	2.04	3.14	5.41	12.32	4.17	7.98
	Penalty [48]	5.29	16.99	7.71	2.23	5.75	13.94	2.86	3.28	5.92	7.08	2.04	6.64
	SaLS [48]	3.98	13.36	5.59	1.48	4.33	12.10	3.38	3.72	4.78	6.75	1.90	5.74
	O-TPT [58]	5.80	15.10	9.71	1.92	3.85	11.70	1.79	3.20	4.60	13.42	3.63	6.89
	A-TPT [1]	5.43	14.88	10.21	1.47	7.58	14.46	1.26	3.16	4.63	7.05	3.97	6.87
	SoC [13]	5.35	16.31	8.95	1.48	5.08	13.85	1.43	3.64	6.08	7.02	1.85	6.57
	CoTS	3.47	7.13	3.09	2.54	4.97	4.72	5.02	6.40	3.37	6.66	5.61	4.95
	E-CoTS	3.05	6.26	2.77	3.02	5.20	3.80	3.27	5.28	3.13	6.78	7.02	**4.65**

Table 12: Performance comparison of different methods integrated with an online TTA framework (TDA [29]) on ImageNet variants using ViT-B/16. Accuracy (%) and ECE (%) are reported.

Method	Accuracy (%)					ECE (%)				
	-A	-V	-R	-Sk	Average	-A	-V	-R	-Sk	Average
TDA [29]	50.03	61.87	75.17	48.84	58.98	11.96	10.06	1.62	10.10	8.44
C-TPT [75]	52.92	63.32	76.60	49.21	60.51	12.03	13.06	1.93	14.59	10.40
Penalty [48]	50.03	61.89	75.17	48.83	58.98	11.98	10.07	1.62	10.14	8.45
SaLS [48]	50.03	61.87	75.17	48.84	58.98	10.22	8.11	2.23	7.95	7.13
O-TPT [58]	50.65	62.27	75.22	48.98	59.28	10.57	10.48	1.94	10.92	8.48
A-TPT [1]	51.44	62.57	75.95	49.14	59.77	9.45	10.01	1.84	10.43	7.93
SoC [13]	51.45	62.29	75.96	48.73	59.61	12.03	9.29	2.58	9.07	8.24
CoTS	50.03	61.87	75.17	48.84	58.98	6.37	3.40	5.43	2.22	**4.36**
E-CoTS	59.63	64.28	78.36	51.16	**63.36**	8.18	3.60	6.24	1.89	4.98

based on the confidence values, where B_m contains all samples whose predicted confidence falls into the m-th interval. The ECE is then computed as a weighted average of the absolute difference between accuracy and confidence within each bin:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{S} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|, \tag{8}$$

where S is the total number of samples, $|B_m|$ denotes the number of samples in bin B_m, and $\text{acc}(B_m)$ and $\text{conf}(B_m)$ represent the fraction of correctly predicted samples and the mean predicted confidence for the m-th bin, respectively. ECE provides a scalar measure of miscalibration: smaller values indicate better alignment between predicted confidences and empirical accuracies.

Brier score. We also report the Brier Score (BS) [4] to evaluate the quality of probabilistic predictions. Different from ECE, which measures the confidence-accuracy discrepancy after binning, the BS directly computes the squared error between the predicted probability distribution and the one-hot ground-truth label. For a multi-class classification problem with K classes, it is formulated as:

$$\text{BS} = \frac{1}{S} \sum_{i=1}^{S} \sum_{c=1}^{K} \left(p_{i,c} - y_{i,c} \right)^2, \tag{9}$$

where S denotes the total number of samples, $p_{i,c}$ represents the predicted probability of sample i for class c, and $y_{i,c}$ is the corresponding one-hot ground-truth label. A lower BS indicates that the predicted probability distribution is closer to the true label distribution. The BS values of different methods with ViT-B/16 are shown in Table 17 and Table 18. Our methods achieve competitive Brier scores across datasets.

Table 13: Performance comparison of different methods with SigLIP [77] on ImageNet and fine-grained datasets. Accuracy (%) and ECE (%) are reported.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Average
Accuracy	Zero-Shot [55]	75.69	62.94	84.41	87.30	69.64	40.65	93.21	97.93	70.82	41.35	90.70	74.06
	TPT [61]	76.46	64.60	84.53	87.60	70.00	40.95	92.75	98.13	71.19	40.68	91.28	74.38
	C-TPT [75]	76.11	63.18	84.25	87.35	69.70	40.29	92.89	97.97	70.61	40.69	90.67	73.97
	Penalty [48]	76.45	64.07	84.33	87.45	69.91	40.38	92.72	96.19	70.90	41.83	91.11	74.12
	SaLS [48]	76.46	64.60	84.53	87.60	70.00	40.95	92.75	98.13	71.19	40.68	91.28	74.38
	O-TPT [58]	75.69	62.41	84.33	87.32	69.56	39.78	92.86	97.93	70.18	40.68	90.69	73.77
	A-TPT[1]	75.91	62.41	84.29	87.31	69.59	40.32	92.89	98.01	70.55	40.60	90.69	73.87
	SoC[13]	75.97	63.00	84.33	87.35	69.68	40.65	93.13	97.93	70.82	41.35	90.75	74.09
	CoTS	76.46	64.60	84.53	87.60	70.00	40.95	92.75	98.13	71.19	40.68	91.28	**74.38**
	E-CoTS	77.86	65.19	84.78	87.73	70.04	39.60	92.72	98.22	71.64	38.74	91.38	74.35
ECE	Zero-Shot [55]	4.54	6.29	3.83	2.04	3.89	7.80	2.54	1.53	6.94	15.99	0.69	5.10
	TPT [61]	6.57	11.70	4.08	3.34	7.04	12.90	1.87	0.81	9.52	19.90	1.87	7.24
	C-TPT [75]	5.81	9.40	4.15	2.57	5.35	11.48	2.12	1.30	8.35	18.39	0.91	6.35
	Penalty [48]	6.49	9.68	4.21	2.69	6.90	12.90	2.60	1.63	8.30	14.35	1.43	6.47
	SaLS [48]	6.29	10.43	4.12	3.05	6.63	12.61	1.87	0.82	9.07	18.95	1.75	6.87
	O-TPT [58]	4.96	8.76	3.87	2.42	4.63	11.02	2.18	1.38	8.11	18.51	0.74	6.05
	A-TPT [1]	5.35	8.36	3.99	2.48	4.87	11.28	2.16	1.29	8.23	17.84	0.83	6.06
	SoC [13]	5.30	6.30	3.89	2.30	4.36	7.83	2.33	1.21	7.40	16.04	1.03	5.27
	CoTS	4.18	4.67	4.18	2.20	3.68	7.55	2.27	1.37	6.65	15.12	1.18	**4.82**
	E-CoTS	4.22	5.22	4.67	1.95	3.95	6.64	2.34	1.44	6.26	16.27	0.84	4.89

Table 14: Performance comparison of different methods with SigLIP [77] on ImageNet variants. Accuracy (%) and ECE (%) are reported.

Method	Accuracy (%)					ECE (%)				
	-A	-V	-R	-Sk	Average	-A	-V	-R	-Sk	Average
Zero-Shot [55]	45.07	68.43	89.31	66.59	67.35	15.90	7.26	1.42	7.43	8.00
TPT [61]	46.73	69.23	89.89	67.08	68.23	17.25	9.45	0.68	10.37	9.43
C-TPT[75]	45.84	68.94	89.57	66.88	67.81	17.08	8.65	0.88	9.57	9.04
Penalty [48]	46.01	69.19	89.70	67.08	68.00	16.98	9.40	0.78	10.28	9.36
SaLS [48]	46.73	69.23	89.89	67.08	68.23	16.69	9.11	0.86	9.98	9.16
O-TPT [58]	45.25	68.41	89.34	66.56	67.39	16.54	7.79	1.05	7.99	8.34
A-TPT [1]	45.45	68.78	89.46	66.80	67.62	16.71	8.10	0.96	8.69	8.62
SoC [13]	45.72	68.75	89.51	66.85	67.71	16.39	8.05	1.02	8.39	8.46
CoTS	46.73	69.23	89.89	67.08	68.23	13.91	6.72	1.82	7.14	7.40
E-CoTS	56.35	70.70	90.95	67.44	**71.36**	10.64	6.42	2.44	7.43	**6.73**

Class-wise ECE. We additionally report the Class-wise Expected Calibration Error (CECE) [35], which evaluates calibration across all classes rather than only the predicted class. It is computed as:

$$\text{CECE} = \sum_{m=1}^{M} \sum_{c=1}^{K} \frac{|B_{m,c}|}{SK} \left|\text{acc}_c(B_{m,c}) - \text{conf}_c(B_{m,c})\right|, \tag{10}$$

where $B_{m,c}$ denotes the m-th confidence bin for class c. Lower CECE indicates better class-wise calibration. The results are reported in Table 19 and Table 20.

Adaptive ECE. We also report Adaptive Expected Calibration Error (AECE) [51], which uses adaptive bins with approximately equal numbers of samples to reduce the sensitivity to fixed-width binning. It is defined as:

$$\text{AECE} = \sum_{r=1}^{R} \sum_{c=1}^{K} \frac{1}{RK} \left|\text{acc}_c(B_{r,c}) - \text{conf}_c(B_{r,c})\right|. \tag{11}$$

Lower AECE indicates better calibration. The results are reported in Table 21 and Table 22.

Table 15: Standard deviations across three seeds on fine-grained datasets and ImageNet using ViT-B/16. Accuracy(%) and ECE(%) are reported in percentage points.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars
Accuracy	TPT [61]	0.04	0.11	0.08	0.02	0.06	0.16	0.11	0.22	0.27	0.09	0.08
	C-TPT [75]	0.01	0.15	0.12	0.03	0.11	0.09	0.02	0.10	0.05	0.02	0.11
	Penalty [48]	0.02	0.10	0.23	0.01	0.06	0.35	0.19	0.18	0.15	0.06	0.07
	SaLS [48]	0.04	0.11	0.08	0.02	0.05	0.19	0.11	0.22	0.27	0.09	0.09
	O-TPT [58]	0.02	0.11	0.04	0.04	0.04	0.01	0.05	0.10	0.06	0.06	0.12
	A-TPT [1]	0.02	0.03	0.07	0.03	0.04	0.35	0.01	0.16	0.13	0.00	0.10
	SoC [13]	0.03	0.05	0.12	0.03	0.04	0.01	0.12	0.11	0.07	0.01	0.05
	CoTS	0.04	0.11	0.08	0.02	0.06	0.16	0.11	0.22	0.27	0.09	0.08
	E-CoTS	0.04	0.28	0.06	0.03	0.09	0.20	0.26	0.12	0.18	0.14	0.13
ECE	TPT [61]	0.07	0.10	0.07	0.04	0.14	0.12	0.10	0.12	0.38	0.04	0.07
	C-TPT [75]	0.04	0.17	0.10	0.04	0.15	0.07	0.10	0.12	0.22	0.26	0.35
	Penalty [48]	0.04	0.15	0.25	0.07	0.11	0.40	0.18	0.14	0.17	0.15	0.17
	SaLS [48]	0.06	0.18	0.11	0.01	0.03	0.11	0.09	0.06	0.36	0.38	0.08
	O-TPT [58]	0.02	0.14	0.14	0.04	0.04	0.13	0.07	0.10	0.07	0.10	0.15
	A-TPT [1]	0.02	0.25	0.13	0.03	0.08	0.31	0.10	0.16	0.08	0.00	0.10
	SoC [13]	0.06	0.02	0.04	0.02	0.02	0.05	0.16	0.09	0.06	0.01	0.05
	CoTS	0.08	0.61	0.12	0.03	0.19	0.28	0.07	0.22	0.29	0.27	0.07
	E-CoTS	0.03	0.86	0.26	0.00	0.26	0.13	0.09	0.16	0.36	0.13	0.15

Table 16: Standard deviations across three seeds on ImageNet variants using ViT-B/16. Accuracy(%) and ECE(%) are reported in percentage points.

Method	Accuracy (%)				ECE (%)			
	-A	-V	-R	-Sk	-A	-V	-R	-Sk
TPT [61]	0.07	0.14	0.04	0.03	0.15	0.24	0.03	0.05
C-TPT [75]	0.13	0.16	0.06	0.02	0.18	0.05	0.05	0.02
Penalty [48]	0.13	0.16	0.03	0.02	0.10	0.22	0.09	0.04
SALS [48]	0.07	0.14	0.04	0.03	0.04	0.24	0.13	0.03
O-TPT [58]	0.01	0.02	0.03	0.03	0.03	0.16	0.03	0.05
A-TPT [1]	0.04	0.19	0.06	0.04	0.14	0.18	0.07	0.06
SoC [13]	0.10	0.08	0.03	0.03	0.09	0.11	0.04	0.02
CoTS	0.07	0.14	0.04	0.03	0.29	0.18	0.05	0.08
E-CoTS	0.24	0.11	0.08	0.07	0.02	0.16	0.08	0.08

D Additional Analysis

D.1 Sensitivity to Hand-crafted Prompt Initialization

We evaluate the performance of different methods under multiple initial prompt templates on ImageNet variants. Table 23 reports ACC (%) and ECE (%) for each method, showing how CoTS and E-CoTS consistently maintain low ECE while achieving high accuracy across prompts.

D.2 Ensemble Weight α in E-CoTS

As shown in Fig. 5, the adaptive weighting strategy achieves a better balance between accuracy and calibration, avoiding the need to tune α for different settings.

D.3 Similarity-based Analysis

Fig. 6 shows the weak-strong accuracy ratio plotted against dataset-level similarity. This result demonstrates the effectiveness of the proposed weak-strong ensemble strategy.

Table 17: Brier scores (%) [4] of different methods on fine-grained datasets with ViT-B/16.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Average
Brier score	Zero-Shot [55]	16.44	18.45	14.22	9.92	18.33	14.77	7.47	6.11	14.55	19.36	16.44	14.19
	TPT [61]	18.20	24.82	16.83	10.14	20.15	18.49	8.78	5.76	16.01	24.81	16.54	16.41
	C-TPT [75]	16.88	21.12	13.95	10.29	19.36	14.40	7.20	5.86	15.17	22.36	16.61	14.84
	Penalty [48]	18.23	23.46	16.28	10.43	20.03	17.37	10.21	6.57	16.13	19.26	16.47	15.86
	SaLS [48]	17.94	23.36	16.23	9.98	19.90	17.61	8.60	5.83	15.75	22.66	16.34	15.84
	O-TPT [58]	16.16	19.06	13.38	10.56	19.46	14.11	7.26	5.71	15.37	22.73	16.57	14.58
	A-TPT [1]	16.47	19.71	13.45	10.27	19.00	15.04	7.44	6.34	15.65	19.36	16.53	14.48
	SoC [13]	16.63	18.61	13.84	10.10	18.79	14.79	7.00	6.02	14.45	19.37	16.59	14.20
	CoTS	16.61	19.51	14.24	9.73	18.34	14.14	7.43	6.62	14.36	21.62	16.22	14.44
	E-CoTS	16.30	19.72	14.36	9.60	17.99	14.30	7.45	5.58	14.03	21.43	15.96	14.25

Table 18: Brier scores (%) [4] of different methods on ImageNet variants with ViT-B/16.

(%)	Method	-A	-V	-R	-Sk	Average
Brier score	Zero-Shot [55]	20.68	17.58	12.45	17.91	17.16
	TPT [61]	23.95	20.10	12.82	22.12	19.75
	C-TPT [75]	21.64	18.56	12.47	19.83	18.13
	Penalty [48]	23.87	20.00	12.97	22.13	19.74
	SaLS [48]	23.34	19.71	12.50	21.68	19.30
	O-TPT [58]	20.32	17.41	12.40	17.98	17.03
	A-TPT [1]	20.51	17.69	12.35	18.25	17.20
	SoC [13]	20.63	17.66	12.25	18.09	17.16
	CoTS	21.40	17.97	12.67	18.33	17.59
	E-CoTS	21.93	17.75	12.10	18.39	17.54

D.4 Step-wise Analysis

Fig. 7 presents ECE and ACC of CoTS and E-CoTS over iteration steps on DTD and ImageNet-A. As the number of steps increases, ECE tends to converge while ACC remains stable.

D.5 Sensitivity to K and T_0

We further investigate the sensitivity to the number of selected strong views K and the temperature initialization T_0. Fig. 8 shows that CoTS is generally robust to variations in both hyperparameters. We set $K = 6$ and $T_0 = 1.0$ by default throughout the experiments.

D.6 Comparison between Ensemble-Only and E-CoTS

We further compare base methods, ensemble-only, and E-CoTS across different TPT-based calibration methods (Table 24). Ensemble-only often improves accuracy but can sometimes increase ECE. By integrating CoTS into the ensemble prediction, E-CoTS reduces ECE while retaining the accuracy gains from ensembling.

D.7 Comparison with ZERO and Our Ensemble Strategy

Table 25 compares ZERO [10] with our ensemble strategy under both zero-shot and TPT settings. Our ensemble achieves a better accuracy–calibration trade-off than ZERO, while E-CoTS further substantially reduces ECE and maintains competitive accuracy.

Table 19: Class-wise ECE (%) [35] of different methods on fine-grained datasets with ViT-B/16.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Average
CECE	Zero-Shot [55]	0.04	1.42	0.62	0.21	0.13	0.57	0.70	0.24	0.52	6.59	0.23	1.03
	TPT [61]	0.04	1.49	0.53	0.19	0.12	0.71	0.62	0.17	0.47	6.95	0.22	1.05
	C-TPT [75]	0.04	1.39	0.54	0.24	0.12	0.61	0.60	0.22	0.55	7.37	0.23	1.08
	Penalty [48]	0.04	1.43	0.54	0.19	0.12	0.66	0.93	0.23	0.49	5.96	0.22	0.98
	SaLS [48]	0.04	1.46	0.53	0.18	0.12	0.69	0.62	0.18	0.47	6.38	0.22	0.99
	O-TPT [58]	0.04	1.36	0.55	0.25	0.14	0.60	0.60	0.23	0.58	7.39	0.23	1.09
	A-TPT [1]	0.04	1.38	0.54	0.24	0.13	0.60	0.63	0.25	0.57	6.59	0.23	1.02
	SoC [13]	0.04	1.44	0.58	0.23	0.13	0.57	0.62	0.25	0.51	6.58	0.23	1.02
	CoTS	0.04	1.25	0.58	0.20	0.12	0.55	0.72	0.25	0.50	6.07	0.24	0.96
	E-CoTS	0.04	1.28	0.58	0.20	0.12	0.56	0.70	0.22	0.50	6.02	0.24	**0.95**

Table 20: Class-wise ECE (%) [35] of different methods on ImageNet variants with ViT-B/16.

(%)	Method	-A	-V	-R	-Sk	Average
CECE	Zero-Shot [55]	0.30	0.06	0.18	0.06	0.15
	TPT [61]	0.28	0.06	0.14	0.06	0.14
	C-TPT [75]	0.29	0.06	0.16	0.06	0.14
	Penalty [48]	0.28	0.06	0.14	0.06	0.14
	SaLS [48]	0.28	0.06	0.14	0.06	0.14
	O-TPT [58]	0.30	0.06	0.18	0.06	0.15
	A-TPT [1]	0.30	0.06	0.17	0.06	0.15
	SoC [13]	0.29	0.06	0.17	0.06	0.15
	CoTS	0.27	0.06	0.17	0.06	0.14
	E-CoTS	0.26	0.06	0.16	0.06	**0.13**

Table 21: Adaptive ECE (%) [51] of different methods on fine-grained datasets with ViT-B/16.

(%)	Method	ImgNet	DTD	Flowers	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Average
AECE	Zero-Shot [55]	0.03	1.33	0.48	0.18	0.10	0.55	0.54	0.17	0.41	6.57	0.12	0.95
	TPT [61]	0.02	1.37	0.45	0.15	0.08	0.66	0.54	0.13	0.38	6.93	0.13	0.98
	C-TPT [75]	0.02	1.24	0.43	0.19	0.09	0.59	0.48	0.15	0.43	7.33	0.13	1.01
	Penalty [48]	0.02	1.35	0.46	0.16	0.08	0.63	0.80	0.16	0.40	5.97	0.13	0.92
	SaLS [48]	0.02	1.35	0.46	0.15	0.08	0.64	0.55	0.14	0.38	6.40	0.13	0.94
	O-TPT [58]	0.03	1.28	0.43	0.20	0.09	0.60	0.48	0.15	0.46	7.34	0.14	1.02
	A-TPT [1]	0.03	1.28	0.43	0.20	0.10	0.57	0.49	0.17	0.46	6.57	0.13	0.95
	SoC [13]	0.03	1.36	0.46	0.18	0.09	0.55	0.49	0.18	0.41	6.56	0.12	0.95
	CoTS	0.03	1.29	0.48	0.17	0.09	0.56	0.60	0.18	0.41	6.11	0.12	**0.91**
	E-CoTS	0.03	1.30	0.49	0.17	0.09	0.56	0.60	0.17	0.41	6.05	0.13	0.91

Table 22: Adaptive ECE (%) [51] of different methods on ImageNet variants with ViT-B/16.

(%)	Method	-A	-V	-R	-Sk	Average
AECE	Zero-Shot [55]	0.24	0.03	0.16	0.05	0.12
	TPT [61]	0.21	0.03	0.12	0.05	0.10
	C-TPT [75]	0.23	0.03	0.14	0.05	0.11
	Penalty [48]	0.21	0.03	0.12	0.05	0.10
	SaLS [48]	0.21	0.03	0.12	0.05	**0.10**
	O-TPT [58]	0.25	0.03	0.16	0.05	0.12
	A-TPT [1]	0.25	0.03	0.15	0.05	0.12
	SoC [13]	0.24	0.03	0.14	0.05	0.11
	CoTS	0.22	0.03	0.14	0.05	0.11
	E-CoTS	0.21	0.03	0.14	0.05	0.11

Table 23: Performance comparison under different initial prompts on ImageNet variants. Accuracy (%) and ECE (%) are reported.

Prompt template	Metric (%)	Zero-Shot [55]	TPT [61]	SaLS [48]	A-TPT [1]	SoC [13]	CoTS	E-CoTS
"a_bad_photo_of_a"	ACC	58.60	61.61	61.61	57.58	59.03	61.61	63.66
	ECE	4.47	12.07	10.61	5.13	5.73	5.16	5.12
"a_black_and_white_photo_of_a"	ACC	55.46	59.30	59.30	56.84	57.09	59.30	60.69
	ECE	6.50	18.07	16.67	7.88	8.06	5.95	7.23
"a_blurry_photo_of_a"	ACC	57.72	61.11	61.11	58.35	59.10	61.11	62.82
	ECE	4.87	15.67	12.72	10.27	8.74	5.32	5.25
"a_good_photo_of_a"	ACC	57.67	61.34	61.34	58.48	59.47	61.34	63.21
	ECE	4.67	14.90	12.84	8.14	7.86	5.35	5.57
"a_high_contrast_photo_of_a"	ACC	56.35	60.20	60.20	57.21	57.79	60.20	61.59
	ECE	5.49	19.30	16.57	8.84	9.62	5.43	6.20
"a_low_contrast_photo_of_a"	ACC	57.63	61.28	61.28	56.57	58.28	61.28	62.82
	ECE	5.21	17.34	8.12	8.60	14.98	5.31	5.55
"a_photo_of_a_small"	ACC	56.25	60.10	60.10	58.88	58.15	60.10	61.67
	ECE	4.19	14.97	12.61	5.10	6.93	5.07	4.80

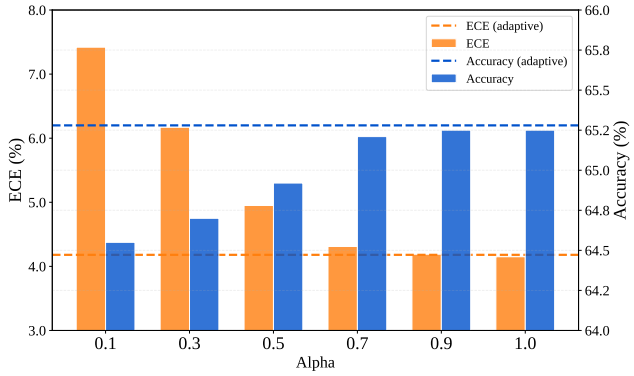


Figure 5: .Effect of the ensemble weight α on E−CoTS using ViT-B/16. Average accuracy (%) and ECE (%) on fine-grained datasets are reported.

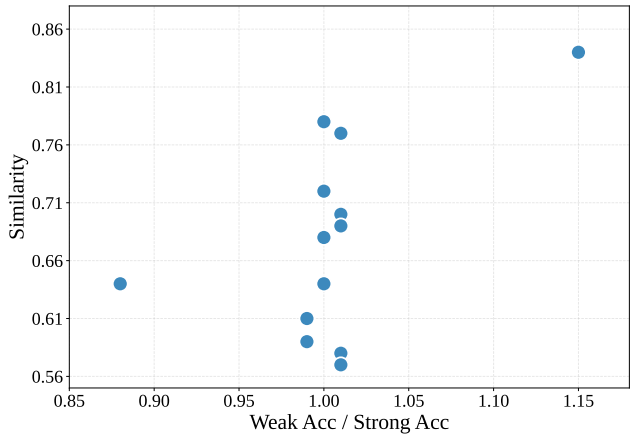
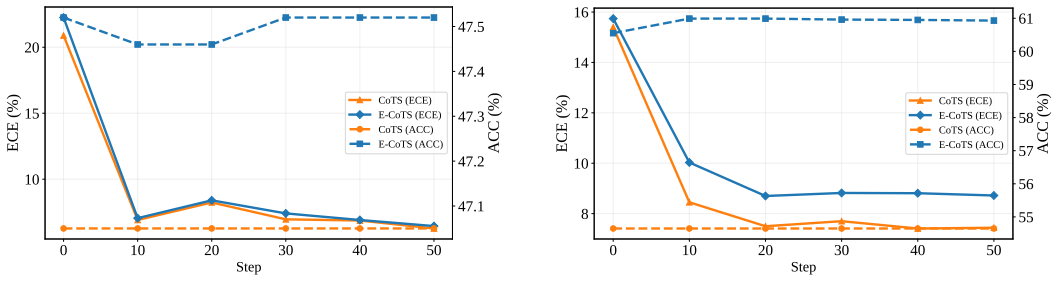


Figure 6: Weak-strong accuracy ratio versus dataset-level similarity across different datasets.



(a) DTD (b) ImageNet-A

Figure 7: ACC(%) and ECE(%) curves of CoTS and E−CoTS on DTD and ImageNet-A over iteration steps using ViT-B/16.

Table 24: Performance of base, direct ensemble, and E−CoTS across different TPT-based methods using ViT-B/16. Accuracy (%) and ECE (%) are reported on fine-grained datasets and ImageNet variants.

(%)	Method	Fine-grained datasets					ImageNet variants				
		TPT [61]	C-TPT [75]	O-TPT [58]	A-TPT [1]	SoC [13]	TPT [61]	C-TPT [75]	O-TPT [58]	A-TPT [1]	SoC [13]
ACC	base	65.25	64.57	64.04	64.15	63.96	60.72	59.26	57.51	58.18	58.17
	Ensemble	65.32	65.27	64.99	65.12	65.11	62.90	62.66	62.03	62.41	61.84
	E−CoTS	65.28	65.24	64.95	65.09	65.11	62.99	62.67	61.97	62.28	61.81
ECE	base	11.25	5.24	4.80	4.27	4.30	12.06	6.74	4.99	4.37	4.83
	Ensemble	12.28	5.35	4.90	4.00	4.12	13.36	7.86	5.36	5.38	4.08
	E−CoTS	4.18	4.19	4.29	3.86	4.18	5.47	5.18	4.51	4.84	4.66

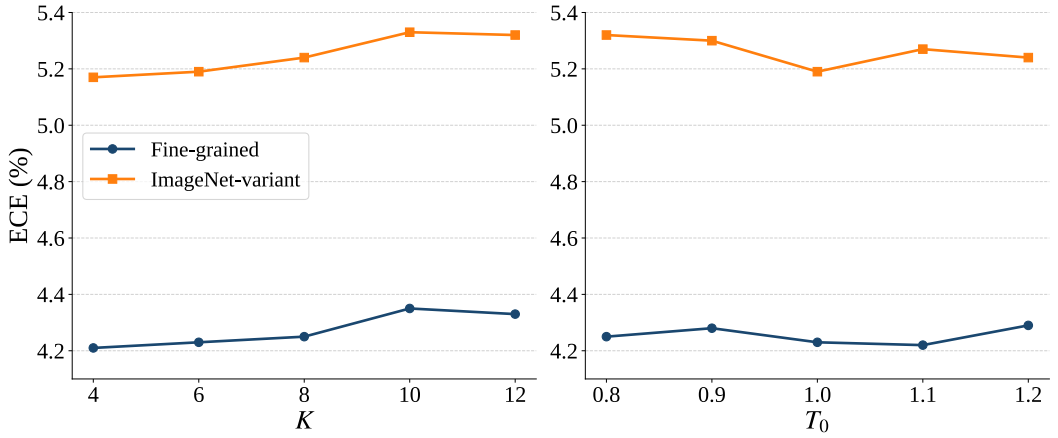


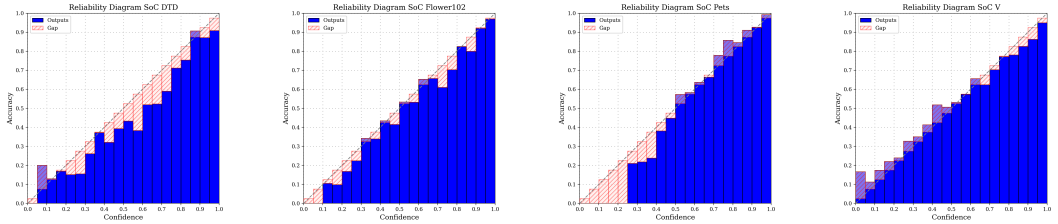
Figure 8: Sensitivity analysis of CoTS to the number of selected strong views K and the temperature initialization T_0. ECE (%) is reported on fine-grained datasets and ImageNet variants.

Table 25: Comparison of ZERO [10] and our ensemble strategy using ViT-B/16. Accuracy (%) and ECE (%) are reported on fine-grained datasets and ImageNet variants.

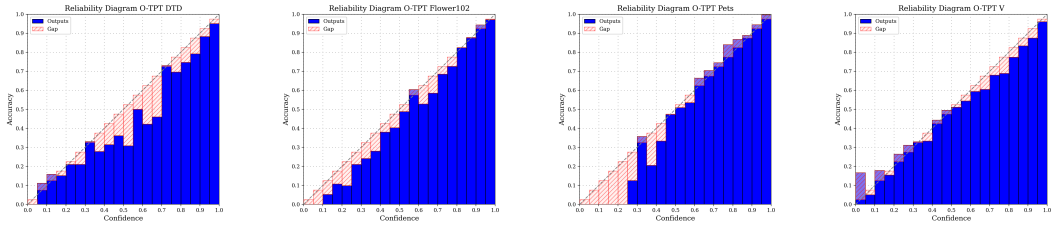
Method	Fine-grained datasets		ImageNet variants	
	ACC (%)	ECE (%)	ACC (%)	ECE (%)
Zero-Shot [55]	63.95	4.33	57.22	4.98
w/ ZERO [10]	64.41	18.24	**62.42**	19.54
w/ Ensemble (Ours)	**65.21**	**4.00**	61.86	**5.03**
TPT [61]	65.19	11.30	60.74	11.90
w/ ZERO [10]	64.65	27.26	**62.97**	28.34
w/ Ensemble (Ours)	**65.32**	12.28	62.90	13.36
w/ E-CoTS	65.22	**4.31**	62.95	**5.38**

Table 26: Performance of different test-time adaptation methods on the COVID [56] and Kather [30] medical datasets using PLIP [25]. Accuracy (%) and ECE (%) are reported, with the best result among adaptation methods highlighted in bold.

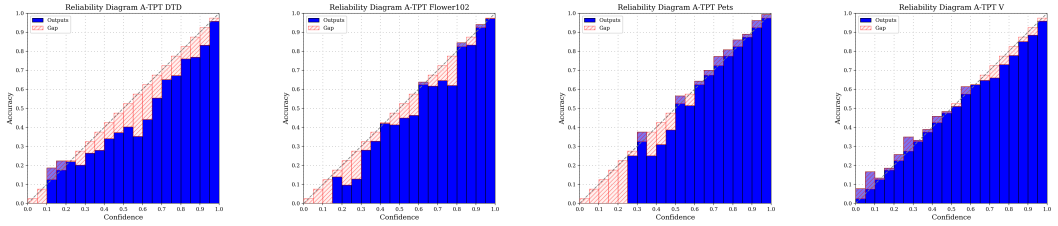
Dataset	Metric	Zero-Shot [55]	TPT [61]	C-TPT [75]	SaLS [48]	O-TPT [58]	A-TPT [1]	CoTS	E-CoTS
COVID [56]	ACC	73.86	75.68	74.28	75.68	74.88	73.81	75.68	**75.71**
	ECE	9.85	20.56	21.35	13.52	20.52	18.31	11.32	**11.03**
Kather [30]	ACC	56.57	59.21	**60.69**	59.21	59.26	58.67	59.21	59.07
	ECE	8.53	21.12	17.32	9.43	16.71	9.74	8.18	**7.79**



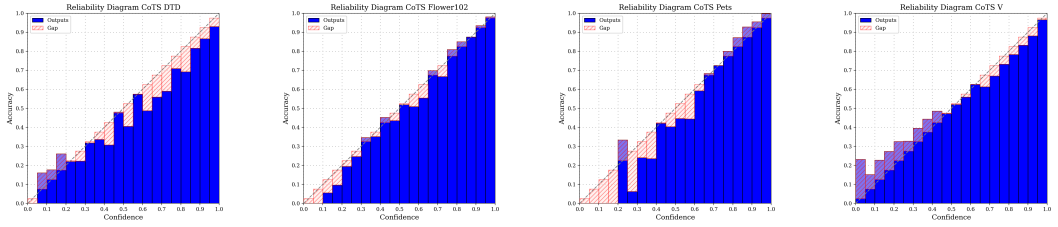
(a) SoC [13]: DTD (b) SoC [13]: Flowers (c) SoC [13]: Pets (d) SoC [13]: IN-V



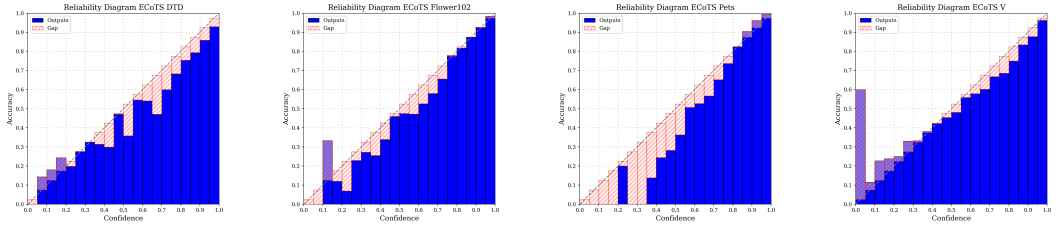
(e) O-TPT [58]: DTD (f) O-TPT [58]: Flowers (g) O-TPT [58]: Pets (h) O-TPT [58]: IN-V



(i) A-TPT [1]: DTD (j) A-TPT [1]: Flowers (k) A-TPT [1]: Pets (l) A-TPT [1]: IN-V



(m) CoTS: DTD (n) CoTS: Flowers (o) CoTS: Pets (p) CoTS: IN-V



(q) E−CoTS: DTD (r) E−CoTS: Flowers (s) E−CoTS: Pets (t) E−CoTS: IN-V

Figure 9: Reliability diagrams of SoC [13], O-TPT [58], A-TPT [1], CoTS , and E−CoTS on DTD, Flowers, Pets, and ImageNet-V using ViT-B/16.

E RELIABILITY DIAGRAM

As shown in Fig. 9, CoTS and E−CoTS achieve small accuracy-confidence gaps across datasets, indicating reliable calibration.