Improving Calibration in Test-Time Prompt Tuning for Vision-Language Models via Data-Free Flatness-Aware Prompt Pretraining

Hyeonseo Jang[1] Jaebyeong Jeon[1] Joong-Won Hwang[2] Kibok Lee[1]
[1]Yonsei University [2]ETRI
{jhyeonseo715, jaebyeong98, kibok}@yonsei.ac.kr jwhwang@etri.re.kr

Abstract

Test-time prompt tuning (TPT) has emerged as a promising technique for enhancing the adaptability of vision-language models by optimizing textual prompts using unlabeled test data. However, prior studies have observed that TPT often produces poorly calibrated models, raising concerns about the reliability of their predictions. Recent works address this issue by incorporating additional regularization terms that constrain model outputs, which improve calibration but often degrade performance. In this work, we reveal that these regularization strategies implicitly encourage optimization toward flatter minima, and that the sharpness of the loss landscape around adapted prompts is a key factor governing calibration quality. Motivated by this observation, we introduce Flatness-aware Prompt Pretraining (FPP), a simple yet effective pretraining framework for TPT that initializes prompts within flatter regions of the loss landscape prior to adaptation. We show that simply replacing the initialization in existing TPT pipelines—without modifying any other components—is sufficient to improve both calibration and performance. Notably, FPP requires no labeled data and incurs no additional computational costs during test-time tuning, making it highly practical for real-world deployment. The code is available at: https://github.com/YonseiML/fpp.

1. Introduction

Vision-language models such as CLIP [32] have recently achieved remarkable zero-shot performance across a wide range of downstream tasks [9, 32]. In these models, textual input templates play a crucial role in determining performance, motivating recent advances in prompt tuning methods that optimize prompt templates in a few-shot manner [15, 16, 44, 45]. However, these approaches often face limitations in real-world applications, as they cannot directly adapt to distribution shifts in target tasks [34, 37]. To address this limitation, test-time prompt tuning (TPT) [34] introduces an entropy minimization (EM) loss that enables prompt adaptation without requiring labeled data, inspiring extensive subsequent research [4, 33, 39, 41, 47].

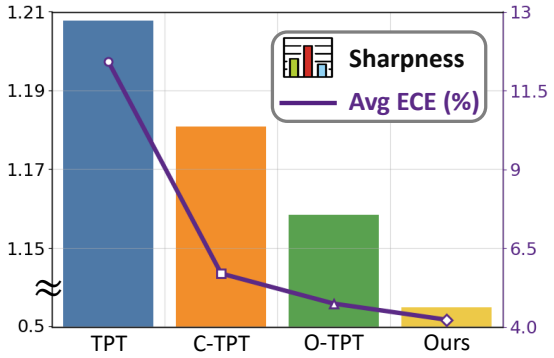

Figure 1. Applying regularization loss into TPT (C-TPT and O-TPT) encourages the prompt to converge toward a flatter loss landscape, leading to reduced Expected Calibration Error (ECE).

However, TPT-based methods have been observed to degrade model calibration, leading to a mismatch between prediction confidence and actual accuracy [1, 27, 33, 41]. Since VLMs are widely deployed in domains that require reliable uncertainty estimates, such as healthcare [23, 38] and autonomous systems [2, 14], ensuring proper calibration in TPT settings has emerged as a critical research direction. Recent studies primarily address this issue by regularizing TPT with additional loss terms [1, 33, 41], which improve calibration through geometric constraints that encourage broader dispersion of output text features. Although these approaches demonstrate empirical success, they offer limited insight into how such regularized optimization affects the learned input prompts, leaving the mechanisms underlying calibration improvement largely unexplored.

In this paper, we show that existing regularization strategies for TPT implicitly act as mechanisms that guide prompts toward flat minima in the loss landscape, which are known to improve generalization in neural networks [3, 13, 20, 46]. As illustrated in Fig. 1, prompts optimized via regularized TPT exhibit reduced sharpness in their surrounding loss landscapes, accompanied by a corresponding decrease in expected calibration error (ECE). Notably, we fur-

ther observe that adapted prompts lying in flat minima consistently achieve substantially lower calibration errors than those trapped in sharp minima, indicating that convergence to flatter regions of the loss landscape is associated with improved calibration. These findings highlight that prompt tuning within flat minima is crucial to achieve effective calibration. However, explicitly enforcing flatness through geometric regularization of output features often degrades performance. Furthermore, existing methods [1, 33, 41] struggle to adequately explore flatter regions of the loss landscape, as increasing the number of iterations also incurs the risk of overconfidence in incorrect predictions.

Building on these insights, we propose FPP, a **F**latness-aware **P**rompt **P**retraining framework that initializes prompts within flatter regions of the loss landscape before TPT. Our FPP employs two complementary objectives: (i) an alignment loss that encourages the learned prompts to generate text features consistent with those of the original prompts, and (ii) a flatness loss that reduces output sensitivity to random perturbations. By jointly optimizing these objectives, FPP generates new initial prompts with provably reduced sharpness for any differentiable loss function, facilitating efficient convergence toward flat minima even in a single step. Because FPP serves as a pretraining scheme for prompt initialization, it can be seamlessly integrated with existing TPT-based methods by replacing their initial prompts with our pretrained ones, without altering subsequent adaptation procedures. Moreover, FPP operates without relying on any external resources such as training or testing image samples, and incurs no extra computational cost during adaptation, making it practical for real-world deployment. Our contributions are summarized as follows:

- We identify a strong correlation between flat minima and calibration quality, and further reveal that existing calibration methods for TPT implicitly guide prompts toward flatter regions of the loss landscape.
- We introduce FPP, a flatness-aware prompt pretraining framework for TPT that initializes prompts in flatter regions of the loss landscape without relying on additional resources, enabling effective convergence to flat minima.
- Our extensive experiments demonstrate that simply initializing existing TPT-based methods with our pretrained prompts achieves state-of-the-art (SOTA) results in both calibration and performance in downstream tasks.

2. Related Work

Prompt Tuning. The zero-shot performance of CLIP [32] is highly sensitive to the phrasing of manually crafted input textual prompts, where even slight variations can lead to substantial changes in accuracy [45]. To address this issue, prompt tuning methods such as CoOp [45] and CoCoOp [44] replace fixed textual templates with learnable prompt vectors that can be trained through few-shot learning. Building on this idea, MaPLe [15] introduces a multimodal approach that generates visual prompts from text prompts via linear projections. Another line of work, TPT [34], extends the test-time adaptation (TTA) paradigm to prompt tuning by employing an EM objective, enabling unsupervised adaptation to distribution shifts in test data. DiffTPT [4] further enhances this approach by incorporating diffusion-generated images to improve accuracy. More recently, Self-TPT [47] proposed a method that further adapts the few-shot learned prompts to class names before test-time, thereby improving generalization across classes. Despite their impressive improvement in accuracy, these methods are designed without considering model calibration, often resulting in overconfident predictions [33, 41].

Calibration of Neural Networks. Calibration measures how well the predicted confidence of a model aligns with its actual accuracy [8]. Broadly, calibration approaches fall into two categories: post-hoc [8, 19, 31, 36] and train-time [10, 17, 40] methods. Post-hoc methods, such as temperature scaling [8] and Platt scaling [31], adjust a trained model using a held-out validation set to better match predicted probabilities with observed outcomes. While effective, these techniques depend on labeled datasets that closely resemble the target distribution, limiting their flexible application [22]. Train-time calibration, on the other hand, incorporates additional calibration objectives during model training to encourage predictions that better align with true probabilities. In the TTA setting, C-TPT [41] and O-TPT [33] follow this paradigm by incorporating an additional regularization loss term to improve calibration.

Flat Minima. The relationship between sharpness of the loss landscape and generalization has been widely explored, showing that flatter local minima improve model generalization [3, 13, 20, 46]. In this regard, SAM [5] and its subsequent studies [18, 21, 24, 48] introduce a new optimization objective that seeks flat minima by uniformly minimizing the loss within a neighborhood. Following them, some TTA methods [7, 29] incorporate the SAM objective into the existing EM loss, leveraging flatter loss landscapes to suppress the influence of noisy samples while continuously updating model parameters. However, despite extensive research on flat minima, their connection to calibration remains largely unexplored and somewhat controversial. For example, [25] report that some regularization techniques, such as data augmentation, improve calibration but often yield sharper minima. Conversely, CSAM [35] shows that the SAM objective acts as an implicit entropy regularizer and leads to improved calibration. However, these studies primarily focus on supervised settings and emphasize the effects of regularization on calibration rather than examining the intrinsic properties of flat minima themselves. In contrast, we demonstrate that even in the absence of explicit regularization, flat minima inherently contribute to better calibration.

3. Preliminaries

3.1. Prompt Tuning for CLIP at Test-Time

CLIP-Based Classification. Given an image X, a set of K class names $C = \{c_1, c_2, \ldots, c_K\}$, and a textual prompt θ, CLIP encodes them into a joint embedding space using an image encoder $f_I(\cdot)$ and a text encoder $f_T(\cdot)$. The image feature is represented as $v = f_I(X)$, while the text feature for class c_k is $t_k = f_T(c_k, \theta)$. Class scores are computed as cosine similarities $\cos(v, t_k)$, and subsequently converted into probabilities using a temperature-scaled softmax function with temperature τ. The predicted class is given by $\hat{c} = \arg\max_{c_k} P(c_k|\theta, v)$, with confidence $\hat{P} = \max_{c_k} P(c_k|\theta, v)$.

Test-Time Prompt Tuning (TPT). TPT [34] performs sample-specific adaptation by optimizing a separate prompt θ for each test pair (X, C) using an entropy minimization (EM) objective. In this setting, the class name set C is fixed, while the image sample X varies at each step, being provided sequentially. The EM loss is computed from a set of image features V, obtained from high-confidence augmented views of X, and is defined as:

$$L_{\text{ent}}(\theta, V) = -\sum_{k=1}^{K} P(c_k \mid \theta, V) \log P(c_k \mid \theta, V), \quad (1)$$

where $P(c_k \mid \theta, V)$ denotes the prediction probability for class c_k, averaged across the augmented views.

Regularized TPT. Motivated by the empirical observation that higher dispersion of text features is associated with lower calibration error, C-TPT [41] and O-TPT [33] incorporate an additional regularization term, L_{reg}, to explicitly encourage feature diversity:

$$L_{\text{reg}}^{\text{C-TPT}}(\theta) = -\frac{1}{K}\sum_{k=1}^{K} \|t_k - \mu\|_2, \quad (2)$$

$$L_{\text{reg}}^{\text{O-TPT}}(\theta) = \|TT^\top - I_K\|_2^2, \quad (3)$$

where $T = f_T(C; \theta) = [t_1, \ldots, t_K]^\top$ denotes the text feature matrix collecting the K text features, $\mu = \frac{1}{K}\sum_{k=1}^{K} t_k$ is the mean text feature vector, and I_K is the identity matrix. For brevity, we omit the dependence on C in the above equations, as it remains fixed for each dataset. The overall objective combines the EM loss with the regularization term, weighted by a hyperparameter λ:

$$L_{\text{total}}(\theta, V) = L_{\text{ent}}(\theta, V) + \lambda L_{\text{reg}}(\theta). \quad (4)$$

3.2. Sharpness-Aware Minimization (SAM)

In prior works [5, 18, 48], *sharpness of the loss landscape* is commonly quantified by the maximum loss difference between the current parameter w_t and its perturbed counterpart. Formally, this measure is defined as follows:

$$h(w_t, v) \triangleq \max_{\|\varepsilon\| \le \rho} L(w_t + \varepsilon, v) - L(w_t, v), \quad (5)$$

where ρ controls the magnitude of the perturbation.

To identify flat minima in the loss landscape and achieve better generalization, recent studies [5, 18, 24, 48] minimize this sharpness using a common two-step optimization procedure. Starting from the current parameter w_t, they first search for a nearby perturbation ε such that $w_t + \varepsilon$ yields a higher loss. The gradient is then computed at this perturbed point and used to update w_t. This procedure encourages the model to reduce loss not only at w_t but throughout its neighborhood, effectively guiding parameters toward flatter regions of the loss landscape [5].

A representative example is Sharpness-Aware Minimization (SAM) [5], which seeks the perturbation that maximizes the loss within a local neighborhood. Because exactly determining such a perturbation is computationally intractable [24], SAM approximates it using a first-order Taylor expansion, leading to the following update equation:

$$w_{t+1} = w_t - \eta \ \nabla_w L(w, v)|_{w_t + \hat{\varepsilon}_t}. \quad (6)$$

Here, $\hat{\varepsilon}_t = \rho \cdot \frac{\nabla_w L(w_t, v)}{|\nabla_w L(w_t, v)|}$ denotes the approximated perturbation. Following prior works [5, 18, 48], we adopt this perturbation form when computing the sharpness defined in Eq. (5) throughout our experiments.

4. Analysis of Regularized TPT through SAM

In this section, we show that promoting higher dispersion of text features via regularization guides prompts toward flat minima in TPT frameworks. We begin by revisiting the standard setup of TPT-based methods [4, 33, 34, 41]. In these approaches, the learned prompt is reinitialized to a predefined prompt θ_0^{zs} for every test sample. Thus, for a fixed set of class names C, adaptation always starts from the same text features, denoted as $f_T(C, \theta_0^{\text{zs}})$. Consequently, since the regularization terms in Eq. (2) and Eq. (3) depend solely on these features, their gradients $\nabla_\theta L_{\text{reg}}(\theta)\big|_{\theta_0^{\text{zs}}}$ remain identical across all test samples.

Because the methods we analyze perform a single-step update per sample [33, 41], accumulated effects such as momentum can be ignored. For analytical clarity, we adopt a simple gradient-based update rule and assume a learning rate set to one. Consequently, the update formulation derived from Eq. (4) is given by:

$$\begin{aligned}\theta_1^{\text{reg}} &= \theta_0^{\text{zs}} - \nabla_\theta L_{\text{ent}}(\theta, V)|_{\theta_0^{\text{zs}}} - \lambda \ \nabla_\theta L_{\text{reg}}(\theta)|_{\theta_0^{\text{zs}}} \\ &= \theta_0^{\text{zs}} - \nabla_\theta L_{\text{ent}}(\theta, V)|_{\theta_0^{\text{zs}}} - \varepsilon_{\text{reg}}, \end{aligned} \quad (7)$$

where ε_{reg} denotes the constant offset induced by the gradient of the regularization loss, and subscripts 0 and 1 denote prompts before and after the update, respectively.

By combining the two fixed terms in Eq. (7), we define $\theta_0^{\text{reg}} := \theta_0^{\text{zs}} - \varepsilon_{\text{reg}}$ as a new initialization for TPT. Under this definition, the regularized TPT becomes equivalent to the

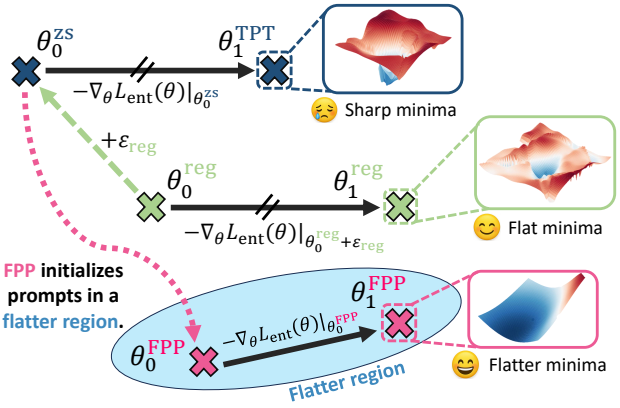

Figure 2. Regularized TPT can be interpreted as an optimization in which the initial prompt is shifted and fixed at θ_0^{reg}, while gradients of the EM losses are computed at a perturbed point θ_0^{zs}. This mechanism encourages convergence toward flat minima. In contrast, our FPP pretrains the initial prompt to reside in a flatter region, facilitating convergence to flat minima without perturbation.

original TPT formulation without the regularization term:

$$\begin{aligned}\theta_1^{\text{reg}} &= \theta_0^{\text{reg}} - \nabla_\theta L_{\text{ent}}(\theta, V)|_{\theta_0^{\text{zs}}} \\ &= \theta_0^{\text{reg}} - \nabla_\theta L_{\text{ent}}(\theta, V)|_{\theta_0^{\text{reg}}+\varepsilon_{\text{reg}}}\,. \end{aligned} \tag{8}$$

Here, the gradient of the EM loss is always computed at the perturbed point $\theta_0^{\text{reg}} + \varepsilon_{\text{reg}}$ rather than at the point where it is ultimately applied, θ_0^{reg}.

Notably, because ε_{reg} corresponds to the gradient of the regularization loss, the perturbed point $\theta_0^{\text{reg}} + \varepsilon_{\text{reg}}$ yields a higher regularization loss than the original point θ_0^{reg}. This encourages text features from different classes to cluster more closely. In other words, adding ε_{reg} can be interpreted as introducing a perturbation that produces more uniform prediction probabilities, which corresponds to an increase in the EM loss. The following theorem formalizes the connection between the regularization objective and the EM loss (see Appendix A for the full statement and proof):

Theorem 1 (Informal). *Let $\mathbb{S}^{D-1} = \{v \in \mathbb{R}^D : \|v\|_2 = 1\}$ be the unit $(D-1)$-sphere, where D is the dimension of the feature space. For an image feature v sampled from the uniform distribution on $\mathbb{S}^{D-1}$, the expected EM loss is*

$$\mathcal{H}(T) := \mathbb{E}_{v\sim \text{Unif}(\mathbb{S}^{D-1})}[L_{\text{ent}}(T, v)]\,.$$

Then, for constants $\alpha > 0$ and β, L_{reg} corresponding to Eq. (2) and Eq. (3) satisfy

$$\mathcal{H}(T) = \alpha L_{\text{reg}}(T) + \beta + O(D^{-3/2}).$$

Therefore, for sufficiently large D, increasing the regularization loss also leads to increase the expected EM loss.

Recall that, in regularized TPT, θ_0^{reg} is updated in the direction of a gradient computed at a perturbed point $\theta_0^{\text{reg}} + \varepsilon_{\text{reg}}$, as shown in Eq. (8). According to Theorem 1, this perturbed point is likely to yield a higher EM loss than the original point. This mechanism aligns with the core principle of the SAM in Eq. (6), where gradients are computed at perturbed points that yield higher loss within the neighborhood. Fig. 1 empirically confirms that the regularization term effectively reduces sharpness, guiding the prompts toward flat minima. Fig. 2 illustrates this mechanism.

5. Flatness-Aware Prompt Pretraining (FPP)

As discussed in Section 4, existing regularization strategies for TPT encourage prompts to converge toward flat minima. While they can improve calibration, the associated geometric constraints often distort output features, leading to performance degradation. Moreover, because these methods rely on a single-step update, they are fundamentally limited to effectively explore flat regions of the loss landscape. Simply increasing the number of update steps is not a viable solution, as it incurs additional computational costs during TTA and amplifies the risk of overconfidence.

Motivated by the correlation between the sharpness of the loss landscape and calibration, we consider an alternative perspective for improving calibration in TPT: rather than seeking flat minima around predefined prompts during TTA, we initialize the prompts directly within a flatter region, such that calibration becomes inherently better. Formally, we aim to initialize prompts θ that already exhibit the desired flatness properties prior to TTA. To achieve this, we propose a flatness loss, which encourages the learned prompts to lie in flatter regions of the loss landscape by penalizing variations in the outputs under small perturbations:

$$\mathcal{L}_{\text{flat}} = \text{dist}_{\cos}\big(f_T(C+\varepsilon_1;\ \theta+\varepsilon_2),\ f_T(C;\theta)\big), \tag{9}$$

where $\text{dist}_{\cos}$ denotes the cosine distance, and ε_1 and ε_2 are small random perturbations. Intuitively, enforcing output stability under such perturbations reduces the model's sensitivity to changes in θ, which leads to lower sharpness for any differentiable loss function (see Appendix B for a formal proof).

However, optimizing only $\mathcal{L}_{\text{flat}}$ can distort the original text features of θ_0^{zs}, thereby degrading zero-shot performance. Because the EM loss is highly sensitive to initial prediction probabilities [37], such distortions can lead to substantial performance drops after adaptation. To address this issue, we introduce an alignment loss that encourages the output text features of the learned prompt to remain close to those of θ_0^{zs} to preserve the semantic structure of the original prompt:

$$\mathcal{L}_{\text{align}} = \text{dist}_{\text{L2}}\big(f_T(C;\theta),\ f_T(C;\theta_0^{\text{zs}})\big), \tag{10}$$

Method	Venue	Metric	Air	Calt	Car	DTD	SAT	FLW	Food	Pets	SUN	UCF	Avg.
CLIP [32]	ICML 2021	Acc.	23.9	92.9	65.3	44.3	41.3	67.3	83.6	88.0	62.5	65.0	63.41
		ECE	5.11	5.50	4.25	8.50	7.40	3.00	2.39	4.37	2.53	3.59	4.67
		SCE	0.52	0.25	0.23	1.33	6.18	0.59	0.20	0.68	0.12	0.52	1.06
TPT [34]	NeurIPS 2022	Acc.	23.4	93.8	66.3	46.7	42.4	69.0	84.7	87.1	65.5	67.3	<u>64.62</u>
		ECE	16.8	4.51	5.16	21.2	21.5	13.5	3.98	5.77	11.3	13.0	11.67
		SCE	0.58	0.16	0.25	1.44	7.07	0.51	0.17	0.60	0.15	0.57	1.15
C-TPT [41]	ICLR 2024	Acc.	24.0	93.6	65.8	46.0	43.2	69.8	83.7	88.2	64.8	65.7	64.48
		ECE	4.36	4.24	1.59	11.9	13.2	5.04	3.43	1.90	5.04	2.54	5.32
		SCE	0.56	0.22	0.22	1.31	6.81	0.52	0.22	0.58	0.14	0.52	<u>1.11</u>
O-TPT [33]	CVPR 2025	Acc.	23.64	93.95	64.53	45.68	42.84	70.07	84.13	87.95	64.23	64.16	64.12
		ECE	3.68	3.80	1.78	7.88	12.98	3.87	1.46	1.90	4.93	2.34	<u>4.46</u>
		SCE	0.56	0.17	1.07	1.24	6.58	0.53	0.19	0.57	0.12	0.51	1.15
FPP (Ours)	—	Acc.	24.75	93.25	66.65	46.14	50.66	69.43	84.31	87.30	64.08	67.08	**65.37**
		ECE	7.26	5.85	2.00	7.52	5.19	2.67	1.92	2.63	3.22	3.04	**4.13**
		SCE	0.57	0.26	0.22	1.31	5.12	0.58	0.20	0.70	0.12	0.51	**0.96**

Table 1. Comparison of accuracy and calibration performance using the CLIP-ViT/B16 backbone within the TPT framework under a predefined hard prompt ("a photo of a"). The best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

where dist_{L2} denotes the L2 distance. We then combine the two losses with a scaling factor λ to form the final objective:

$$\mathcal{L}_{\text{FPP}} = \mathcal{L}_{\text{align}} + \lambda\, \mathcal{L}_{\text{flat}}. \tag{11}$$

Here, we set $\lambda = \gamma_1 + \frac{\gamma_2}{K}$, where $\gamma_1, \gamma_2 > 0$ are hyperparameters and K denotes the number of classes. This formulation downweights the flatness loss for larger class sets, where maintaining alignment with the original text features becomes more challenging. Notably, both losses depend only on the predefined prompt θ_0^{zs} and the class-name set C, enabling data-free pretraining of prompts that inherit the calibration advantages of flat minima. The resulting prompt can be used as an initialization for existing TPT-based methods, which we apply without any modification to their adaptation procedures.

6. Experiments

6.1. Experimental Setting

Implementation Details. We use the CLIP-ViT-B/16 architecture as the backbone. During the pretrain stage, we adopt the AdamW optimizer with a cosine learning rate scheduler. The initial learning rate is set to 0.01, and pretraining is performed for 1K iterations. For λ, we use $\gamma_1 = 1.0$ and $\gamma_2 = 0.15$. The perturbation terms ε_1 and ε_2 are drawn from zero-mean isotropic Gaussian distributions with variances of 0.02 and 0.005, respectively. Ablation studies for hyperparameter settings are provided in the Appendix C. For TTA, we follow the TPT [34] setting adopted in the baseline works [33, 41] without modification, and follow O-TPT [33] for other unspecified details. Following the baseline works [33, 41], we evaluate calibration using ECE and SCE, with detailed definitions provided in the Appendix E.

6.2. Main Results

Fine-Grained Classification. Tab. 1 presents the results for the fine-grained classification task, where the predefined prompt θ_0^{zs} is set as a hard textual template "a photo of a." with class names provided for each dataset. In terms of calibration errors, our method achieves state-of-the-art (SOTA) performance in both ECE and SCE, with a particularly large improvement in SCE. Moreover, our approach also achieves SOTA performance in accuracy, even surpassing TPT. This suggests that the sharpness of the loss landscape can also influence the post-adaptation accuracy. Notably, while C-TPT and O-TPT exhibit a trade-off between accuracy and calibration, our method is the first to achieve SOTA performance in both metrics simultaneously.

Different Predefined Prompts. Tab. 2 presents the results under the same fine-grained classification datasets as in Tab. 1, but with the predefined prompt θ_0^{zs} obtained from the supervised-trained prompt embeddings. Following the baseline [33], we employ the officially released checkpoints of CoOp [45] and MaPLe [15], and use them to train a new initial prompt θ according to Eq. (11). Our approach achieves SOTA performance in both accuracy and calibration, showing even larger improvements compared to the hard prompt setting in Tab. 1. These results indicate that our method can leverage prior knowledge more effectively, suggesting its strong compatibility with supervised approaches.

Different TPT Framework. Tab. 3 presents the results obtained using the Dynaprompt [39] framework, which accumulates updates during TTA instead of resetting the prompt for each sample as in TPT. Specifically, Dynaprompt maintains multiple prompts, adaptively selects suitable ones for each sample, and accumulates updates during TTA. Since the selected prompts are optimized using the same EM loss

Method	Metric	Air	Calt	Car	DTD	SAT	FLW	Food	Pets	SUN	UCF	Avg.
CoOp+TPT	Acc.	20.0	94.0	65.6	44.5	40.6	68.7	83.8	89.1	65.6	67.2	63.91
	ECE	29.6	3.65	6.63	34.8	31.3	19.9	9.66	7.40	20.8	19.9	18.36
CoOp+C-TPT	Acc.	19.2	93.9	63.1	45.0	40.7	69.0	83.7	89.3	65.1	66.6	63.56
	ECE	21.5	1.66	2.45	21.0	13.2	10.2	4.49	2.12	11.8	12.0	10.04
CoOp+O-TPT	Acc.	18.69	93.71	64.12	45.45	40.17	68.57	83.55	89.07	64.01	65.64	63.29
	ECE	16.82	0.92	2.85	16.02	13.76	6.81	3.59	1.92	7.23	9.16	7.91
CoOp+FPP (Ours)	Acc.	21.51	93.96	63.74	44.80	48.72	68.57	84.31	88.42	66.44	67.30	**64.78**
	ECE	9.34	3.49	5.33	11.79	9.69	6.00	1.21	2.00	3.00	4.88	**5.67**
MaPLe+TPT	Acc.	24.36	94.42	66.50	50.05	47.32	70.72	85.01	87.78	64.87	66.48	65.75
	ECE	10.58	2.38	4.14	11.80	9.42	11.63	1.78	1.79	8.47	7.41	6.94
MaPLe+O-TPT	Acc.	24.00	92.29	65.38	49.11	44.58	71.53	84.35	89.97	63.49	65.82	65.05
	ECE	6.41	3.49	3.61	4.90	7.92	4.35	1.49	3.97	2.78	2.22	4.11
MaPLe+FPP (Ours)	Acc.	24.06	93.47	66.27	48.94	52.19	69.35	85.17	91.55	68.29	70.84	**67.01**
	ECE	3.98	2.83	2.18	11.58	1.91	4.03	3.53	5.03	1.74	3.93	**4.07**

Table 2. Comparison of accuracy and calibration error using the CLIP-ViT/B16 backbone within the TPT framework under predefined prompts trained from CoOp and MAPLE. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Metric	Air	Calt	Car	DTD	SAT	FLW	Food	Pets	UCF	Avg.
DynaPrompt (Paper)	Acc.	24.33	94.32	67.65	47.96	42.28	69.95	85.42	88.28	68.72	65.43
DynaPrompt (Replication)	Acc.	22.68	94.16	66.88	47.87	35.91	69.67	84.92	87.71	68.17	64.22
	ECE	18.67	3.05	5.08	23.19	33.76	12.73	6.75	5.62	13.72	13.62
	SCE	0.73	0.15	0.21	1.53	8.53	0.52	0.18	0.57	0.49	1.43
DynaPrompt+C-TPT	Acc.	23.34	93.83	66.04	46.99	36.57	70.08	83.58	88.61	65.74	63.86
	ECE	10.21	2.51	2.17	15.87	16.94	4.35	2.21	2.43	5.33	6.89
	SCE	0.63	0.19	0.22	1.43	6.99	0.53	0.19	0.57	0.49	1.25
DynaPrompt+O-TPT	Acc.	22.41	93.67	65.58	45.80	36.32	68.74	83.39	88.55	65.69	63.35
	ECE	9.60	2.74	2.13	14.31	17.73	3.60	2.03	2.78	3.43	6.48
	SCE	0.61	0.19	0.23	1.43	7.04	0.55	0.20	0.58	0.51	1.26
DynaPrompt+FPP (Ours)	Acc.	24.96	92.33	66.26	46.10	49.09	69.63	84.51	88.36	66.09	**65.26**
	ECE	5.28	5.72	3.70	5.27	8.49	2.44	2.94	4.96	2.20	**4.56**
	SCE	0.56	0.28	0.23	1.33	5.72	0.59	0.21	0.70	0.53	**1.13**

Table 3. Comparison of accuracy and calibration error using the CLIP-ViT/B16 backbone within the DynaPrompt framework under a predefined hard prompt ("a photo of a"). The best and second-best results are highlighted in **bold** and underline, respectively.

as in TPT, we simply add the regularization terms from the baselines [33, 41] to apply them within this new framework. The predefined prompt θ_0^{zs} is set to "a photo of a" following the original setup, with class names provided for each dataset. We adjust the prompt buffer size to 12 and reproduce all experiments using their official codebase. Our method still achieves SOTA performance across all evaluation metrics, showing strong generalizability in different learning paradigms.

Natural Distribution Shifts. Tab. 4 presents results on natural distribution shifts [45], evaluating out-of-distribution (OOD) performance using widely adopted ImageNet variant datasets, where details provided in Appendix. In this setting, we fix λ as 1.25 across all datasets. Under this configuration, our method consistently achieves notable improvements in both ECE and SCE, with only a minor accuracy reduction of 0.42% compared to TPT. In contrast, O-TPT, which exhibits the best calibration among prior works, experiences an accuracy drop of more than 2%. These findings highlight that our method maintains strong effectiveness under OOD scenarios.

6.3. Experimental Analysis

Ablation Study. Tab. 5 summarizes the ablation results across fine-grained classification datasets using TPT, designed to assess the effectiveness of the two proposed losses, the $\mathcal{L}_{\text{align}}$ and the $\mathcal{L}_{\text{flat}}$, in the pretraining stage. The first block reports the default setting, where TPT is directly applied to the original initial prompt θ_0^{zs} without any pretraining. The second block shows that employing the the

Method	Metric	I	I-A	I-V2	I-R	I-S	OOD Avg.
CLIP [32]	Acc.	66.7	47.8	60.8	74.0	46.1	57.18
	ECE	2.12	8.61	3.01	3.58	4.95	5.04
	SCE	0.04	0.30	0.06	0.18	0.06	0.15
TPT [34]	Acc.	69.0	52.6	63.0	76.7	47.5	**59.95**
	ECE	10.6	16.4	11.1	4.36	16.1	11.99
	SCE	0.04	0.29	0.06	0.14	0.06	0.14
C-TPT [41]	Acc.	68.5	51.6	62.7	76.0	47.9	59.55
	ECE	3.15	8.16	6.23	1.54	7.35	5.82
	SCE	0.04	0.29	0.06	0.16	0.06	0.14
O-TPT [33]	Acc.	67.3	49.9	61.7	72.6	47.1	57.82
	ECE	1.97	7.22	3.97	1.46	6.87	4.88
	SCE	0.04	0.29	0.06	0.17	0.06	0.15
FPP (Ours)	Acc.	67.8	52.3	61.9	76.7	47.2	59.53
	ECE	2.97	5.38	3.45	7.07	2.62	**4.63**
	SCE	0.04	0.27	0.06	0.16	0.06	**0.12**

Table 4. Comparison of accuracy and calibration error in natural distribution shifts datasets.

Align Loss	Flat Loss		Zero-Shot	Acc.	ECE	SCE
	ε_1	ε_2				
–	–	–	63.41	64.62	11.67	1.15
–	✓	✓	4.65	4.16	–	–
✓	–	–	63.49	64.40	7.05	1.09
✓	–	✓	63.92	64.77	5.86	1.03
✓	✓	–	64.15	64.89	4.64	1.02
✓	✓	✓	**64.35**	**65.37**	**4.13**	**0.96**

Table 5. Ablation study under the fine-grained classification setting, evaluating the contribution of each component.

flatness loss $\mathcal{L}_{\text{flat}}$ without the alignment loss $\mathcal{L}_{\text{align}}$ leads to a collapse in the zero-shot accuracy of the learned prompt θ, making adaptation with the EM loss infeasible. Conversely, applying only the $\mathcal{L}_{\text{align}}$ preserves baseline zero-shot performance and achieves competitive results after TPT adaptation, but it does not ensure proper calibration. The fourth and fifth blocks ablate the effect of each perturbation component within the $\mathcal{L}_{\text{flat}}$, showing that both contribute to improvements in accuracy and calibration. Finally, the last row confirms that combining all components yields the best overall performance, indicating that each component contributes synergistically to performance improvement.

Class Name Dependency. In many TTA methods, task-specific information such as the set of class names C is known in advance and often utilized to design task-specific configurations [11, 42, 47]. However, this assumption may not hold in scenarios where such information is unavailable prior to test-time. To address this issue, we conduct experiments summarized in Tab 6, where pretraining is performed after modifying the original class names in two ways: (i) applying ImageNet class names, and (ii) replacing the original class name embeddings with Gaussian noise. Notably, our method consistently achieves SOTA performance in both cases, surpassing the baseline results reported in Tab. 1. This finding aligns with observations from [43], which indicate that text semantics are relatively easy to sample and that even random text embeddings can effectively preserve the underlying feature space. Overall, these results demonstrate that our approach is largely independent of explicit class-name supervision and exhibits strong generalization, even when class names are partially or entirely absent.

Class Source	Metric	Fine-Grained Avg.	Δ
ImageNet-1K	Acc.	64.97	-0.40
	ECE	4.40	+0.27
	SCE	0.97	+0.01
Gaussian Noise	Acc.	65.28	-0.09
	ECE	4.25	+0.12
	SCE	1.03	+0.07

Table 6. Evaluation of robustness to class name variation, reported as the average performance across fine-grained classification datasets. Δ denotes the performance difference relative to the default setting, which uses original class names during pretraining.

Flat Minima and Calibration. To further investigate the relationship between flat minima and calibration, we conduct the experiments shown in Fig. 3. For each test sample, we first optimize an individual prompt and then quantify the sharpness of the corresponding loss landscape. The resulting prompts are sorted in ascending order of sharpness and divided equally into three groups, representing low, medium, and high sharpness levels. Within each dataset, we compute the ECE and mean sharpness for each group, and then average these quantities across all fine-grained datasets to visualize their relationship. This procedure follows the baseline setup in [33], ensuring that each group contains an equal and sufficient number of samples for statistically reliable calibration estimates. Across different calibration methods, our findings consistently show that prompts converging to flatter minima exhibit lower calibration errors. These results suggest that, among the various local minima of the EM loss, those corresponding to flatter minima yield more reliable predictive probabilities.

Underconfidence and Overconfidence. To examine whether our approach can address both overconfidence and underconfidence, we present reliability diagrams in Fig. 4. As illustrated, O-TPT exhibits overconfidence on the texture dataset (DTD) and the remote-sensing dataset (EuroSAT), while showing underconfidence on the natural-domain dataset (SUN397). In contrast, our method effectively mitigates these issues across all cases, showing consistent improvements over both types of calibration error.

TPT with SAM Objective. In Section 4, we show that the optimization formulation of regularized TPT [33, 41] is inherently aligned with the core principle underlying the SAM mechanism. The key distinction, however, is that regularized TPT operates with a fixed perturbed point, defined as $\theta_0^{\text{zs}} = \theta_0^{\text{reg}} + \varepsilon_{\text{reg}}$, which we argue is key to enabling SAM in TPT. When SAM-based methods [5, 18] are directly applied to TPT, the algorithm simply seeks a perturbation direction that increases the EM loss, with-

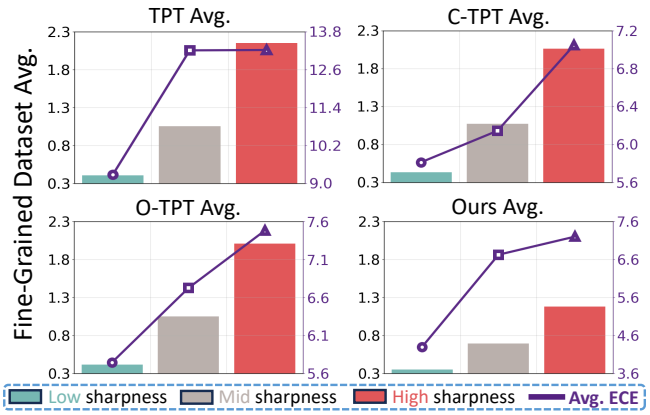

Figure 3. Relationship between sharpness of the loss landscape and calibration error. Across several datasets and methods, flatter minima consistently correspond to lower calibration errors.

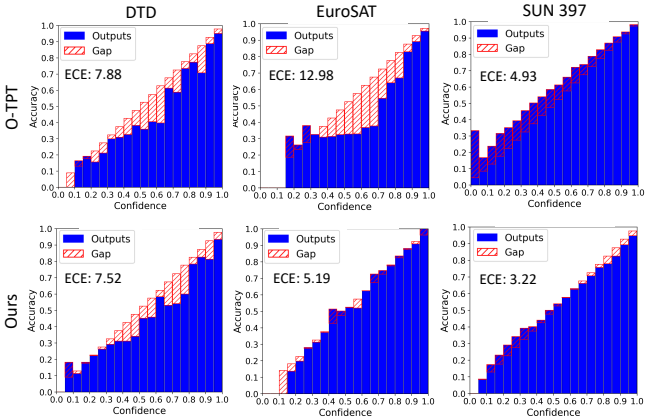

Figure 4. Reliability diagrams showing underconfidence and overconfidence. Our method demonstrates the ability to produce appropriate confidence values as needed.

Method	Acc.	ECE	SCE	Sharp.
TPT	64.62	11.67	1.15	1.23
TPT+ASAM [18]	64.18	9.78	1.12	1.12
TPT+SAM [5]	61.59	4.15	1.12	0.78
FPP (Ours)	**65.37**	**4.13**	**0.96**	**0.52**

Table 7. Evaluation of SAM-based methods under the TPT setting, showing reduced calibration error but degraded accuracy.

Method	Acc.	ECE	SCE	Sharp.
$L_{\text{ent}} \rightarrow L_{\text{ent}}$ (2-step)	64.67	18.90	1.27	1.31
$L_{\text{ent}} \rightarrow L_{\text{reg}}$ (2-step)	64.14	12.60	1.17	1.26
$L_{\text{reg}} \rightarrow L_{\text{ent}}$ (2-step)	64.29	11.78	1.19	1.22
$L_{\text{ent}} + L_{\text{reg}}$ (O-TPT)	64.12	4.46	1.15	1.16

Table 8. Evaluation of isolated effect of regularization loss in O-TPT, showing limited calibration benefit without SAM mechanism.

Method	Dataset	Pretrain	TTA
O-TPT	EuroSAT	–	9m 55s
	SUN397	–	81m 27s
FPP (Ours)	EuroSAT	0m 18s	**9m 11s**
	SUN397	10m 21s	**76m 50s**

Table 9. Comparison of computational costs in terms of execution time between our method and O-TPT.

out considering the prediction probabilities associated with the perturbed prompts. Such perturbations often distort text features, causing incorrect predictions after adaptation due to the high sensitivity of the EM loss to initial accuracy [37]. As shown in Table 7, although these approaches successfully reduce both sharpness and calibration error, they also lead to notable accuracy degradation. One could attempt to mitigate this issue by filtering out noisy test samples to avoid perturbations that destabilize initial predictions [7, 29]. However, this strategy is infeasible in TPT, where sample-specific adaptation is essential. In contrast, our method enables the prompt to converge to flat minima without relying on any form of perturbation.

Regularization Loss Without the SAM Effect. As noted earlier, when the regularization loss is combined with the EM loss, it follows the same optimization principle as SAM, and we argued that the resulting flat minima are the reason of improved calibration. At this point, one might question whether the regularization loss itself has an intrinsic influence on calibration, independent of the flat minima. To address this concern, we conduct experiments in which the two losses are optimized sequentially rather than jointly. This update scheme no longer aligns with the SAM formulation and, as shown in Tab. 8, it fails to produce any meaningful reduction in sharpness. Correspondingly, we observe no improvement in calibration, and the geometric constraint imposed by the regularization term even degrades accuracy. This finding highlights that it is the sharpness of the loss landscape—rather than the regularization term itself—that plays a direct role in achieving effective calibration.

Computational Costs. Tab. 9 summarizes the computation times for both the pretraining and TTA stages. Note that the proposed pretraining method does not rely on any test samples and is completely decoupled from the inference process. Consequently, our method only affects the pretraining stage, maintaining the same inference time as the original TPT, while the overall pretraining time remains negligible compared to the inference time. In contrast, O-TPT, in its practical implementation, incorporates a Householder transformation [12] within the regularization loss, introducing a $\mathcal{O}(|C|^3)$ computational complexity and causing a direct increase in inference latency.

7. Conclusion

In this paper, we reveal that prompts converging to flat minima consistently achieve superior calibration performance compared to those trapped in sharper minima. We further show that existing regularization methods implicitly function as mechanisms that guide prompts toward flat minima of the EM loss. Building on these insights, we propose FPP, a pretraining framework that positions prompts in flatter regions before adaptation, facilitating more effective convergence to flat minima. Notably, simply replacing the initial prompt in existing methods with our pretrained ones yields SOTA calibration and accuracy.

References

[1] Abhishek Basu, Fahad Shamshad, Ashshak Sharifdeen, Karthik Nandakumar, and Muhammad Haris Khan. Calibration-aware prompt learning for medical vision-language models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2025. 1, 2

[2] Arthur Bucker, Luis Figueredo, Sami Haddadin, Ashish Kapoor, Shuang Ma, Sai Vemprala, and Rogerio Bonatti. Latte: Language trajectory transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[3] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 1, 2

[4] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3

[5] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 7, 8

[6] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[7] Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 8

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 2

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2021. 1

[10] Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael Curtis Mozer, and Rebecca Roelofs. Soft calibration objectives for neural networks. In *Advances in Neural Information Processing Systems*, 2021. 2

[11] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[12] Linda Kaufman. The generalized householder transformation and sparse matrices. *Linear Algebra and its Applications*, 90:221–234, 1987. 5. 8

[13] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2

[14] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2023. 1, 2, 5

[16] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[17] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2805–2814. PMLR, 2018. 2

[18] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2021. 2, 3, 7, 8

[19] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. 2

[20] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2

[21] Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[22] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 80–88, 2022. 2

[23] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[24] Yong Liu, Siqi Mai, Minhao Cheng, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Random sharpness-aware minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[25] Israel Mason-Williams, Fredrik Ekholm, and Ferenc Huszar. Explicit regularisation, sharpness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*

Workshop on Scientific Methods for Understanding Deep Learning, 2024. 2

[26] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H.S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5

[27] Balamurali Murugesan, Julio Silva-Rodriguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[28] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 5

[29] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 8

[30] Jeremy Nixon, Michael W, Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 5

[31] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint*, 2021. 1, 2, 5, 7

[33] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19942–19951, 2025. 1, 2, 3, 5, 6, 7, 4

[34] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3, 5, 7

[35] Chengli Tan, Yubo Zhou, Haishan Ye, Guang Dai, Junmin Liu, Zengjie Song, Jiangshe Zhang, Zixiang Zhao, Yunda Hao, and Yong Xu. Towards understanding the calibration benefits of sharpness-aware minimization. *arXiv preprint arXiv:2505.23866*, 2025. 2

[36] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. 2

[37] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 4, 8

[38] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 1

[39] 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 *International Conference on Learning Representations (ICLR)*, 2025. 1, 5

[40] Hee Suk Yoon, Joshua Tian Jin Tee, Eunseop Yoon, Sunjae Yoon, Gwangsu Kim, Yingzhen Li, and Chang D. Yoo. ESD: Expected Squared Difference as a Tuning-free Trainable Calibration Measure. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[41] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 3, 5, 6, 7, 4

[42] Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 7

[43] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 7

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825, 2022. 1, 2

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 1–12, 2022. 1, 2, 5, 6, 4

[46] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, and Weinan E. Towards theoretically understanding why sgd generalizes better than adam in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[47] Yuhan Zhu, Guozhen Zhang, Chen Xu, Haocheng Shen, Xiaoxin Chen, Gangshan Wu, and Limin Wang. Efficient test-time prompt tuning for vision-language models. *arXiv preprint arXiv:2408.05775*, 2024. 1, 2, 7

[48] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

Improving Calibration in Test-Time Prompt Tuning for Vision-Language Models via Data-Free Flatness-Aware Prompt Pretraining

Supplementary Material

A. Proof of Theorem 1

In the main paper, Theorem 1 states that, for either regularizer associated with Eq. (2) or Eq. (3),

$$\begin{aligned}\mathcal{H}(T) &\coloneqq \mathbb{E}_{v\sim\mathrm{Unif}(\mathbb{S}^{D-1})}\big[L_{\mathrm{ent}}(T,v)\big] \\ &= \alpha L_{\mathrm{reg}}(T) + \beta + O(D^{-3/2}),\end{aligned} \tag{A.1}$$

with $\alpha > 0$. We now make this statement precise and prove it.

Recall that $T \in \mathbb{R}^{K\times D}$ denotes the text feature matrix, and $t_i \in \mathbb{R}^D$ denotes its i-th row, where $\|t_i\|_2 = 1$. This assumption is imposed by the CLIP model, which normalizes feature embeddings and computes cosine similarity. To facilitate the analysis, we introduce equivalent surrogate losses that share the same optimal states as the original regularizers:

$$L_{\mathrm{disp}}(T) \coloneqq -\|PT\|_F^2, \qquad P \coloneqq I - \frac{1}{K}11^\top, \tag{A.2}$$

for the regularizer in Eq. (2), and

$$L_{\mathrm{orth}}(T) \coloneqq \sum_{1\le i<j\le K} t_i^\top t_j, \tag{A.3}$$

for the regularizer in Eq. (3). We will show that both losses depend only on the same term, and that the expected EM loss is also determined by that term.

Let $\mu(T) = \frac{1}{K}\sum_{i=1}^K t_i$ denote the mean text feature. We then define the following quantity:

$$\begin{aligned}S(T) &\coloneqq \|PT\|_F^2 \\ &= \sum_{i=1}^K \|t_i - \mu(T)\|_2^2.\end{aligned} \tag{A.4}$$

We first relate $L_{\mathrm{disp}}(T)$ and $L_{\mathrm{orth}}(T)$ to $S(T)$. Since $\sum_{i=1}^K (t_i - \mu(T)) = 0$ and $\|t_i\|_2 = 1$, we have

$$S(T) = K - K\|\mu(T)\|_2^2. \tag{A.5}$$

Moreover, since each t_i has unit norm, $\|\mu(T)\|_2^2$ is determined by the pairwise inner products:

$$\begin{aligned}\|\mu(T)\|_2^2 &= \frac{1}{K^2}\Big\|\sum_{i=1}^K t_i\Big\|_2^2 \\ &= \frac{1}{K^2}\left(K + 2\sum_{1\le i<j\le K} t_i^\top t_j\right).\end{aligned} \tag{A.6}$$

Substituting this into the expression for $S(T)$, we obtain

$$S(T) = K - 1 - \frac{2}{K}\sum_{1\le i<j\le K} t_i^\top t_j. \tag{A.7}$$

Therefore,

$$L_{\mathrm{disp}}(T) = -S(T), \qquad L_{\mathrm{orth}}(T) = \frac{K}{2}\big(K - 1 - S(T)\big). \tag{A.8}$$

Thus, both losses depend only on $S(T)$.

We next relate $S(T)$ to the expected EM loss. For an image feature $v \sim \mathrm{Unif}(\mathbb{S}^{D-1})$, let

$$s(T,v) \coloneqq Tv \tag{A.9}$$

denote the logit vector, and define

$$f(s) \coloneqq H\big(\mathrm{softmax}(s)\big), \tag{A.10}$$

where H denotes entropy. Since softmax is invariant under adding the same scalar to all logits, we have

$$f(s + c1) = f(s) = f(Ps) \qquad \text{for all } c \in \mathbb{R}, \tag{A.11}$$

where $Ps = s - \frac{1}{K}(1^\top s)1$. Thus, $f(s)$ is fully determined by the centered logits Ps.

We now expand f around the origin. Since $\|t_i\|_2 = \|v\|_2 = 1$, each logit satisfies $|s_i(T,v)| = |t_i^\top v| \le 1$. Therefore, $s(T,v) \in [-1,1]^K$, and the centered logits $Ps(T,v)$ lie in the compact set $\{Ps : s \in [-1,1]^K\}$. As f is smooth, its third derivatives are uniformly bounded on this set. At $s = 0$, the softmax distribution is uniform, so

$$f(0) = \log K, \qquad \nabla f(0) = 0, \qquad \nabla^2 f(0) = -\frac{1}{K}P. \tag{A.12}$$

Hence, Taylor expansion around the origin gives

$$f(s) = \log K - \frac{1}{2K}\|Ps\|_2^2 + R(s), \tag{A.13}$$

where $R(s)$ is the remainder term satisfying

$$|R(s)| \le M_K\|Ps\|_2^3 \tag{A.14}$$

for some constant $M_K < \infty$ depending only on K.

Substituting $s(T,v) = Tv$ and taking expectation over $v \sim \mathrm{Unif}(\mathbb{S}^{D-1})$, we obtain

$$\begin{aligned}\mathcal{H}(T) &= \mathbb{E}_v\big[L_{\mathrm{ent}}(T,v)\big] \\ &= \log K - \frac{1}{2K}\mathbb{E}_v\|PTv\|_2^2 + \mathbb{E}_v[R(Tv)].\end{aligned} \tag{A.15}$$

Since v is uniform on the unit sphere,

$$\mathbb{E}_v[vv^\top] = \frac{1}{D} I_D. \tag{A.16}$$

Therefore, since P is a projection matrix satisfying $P^\top = P$ and $P^2 = P$, we obtain

$$\begin{aligned} \mathbb{E}_v\|PTv\|_2^2 &= \operatorname{tr}\big(T^\top P T\, \mathbb{E}_v[vv^\top]\big) \\ &= \frac{1}{D}\operatorname{tr}(T^\top P T) \\ &= \frac{1}{D}\|PT\|_F^2 \\ &= \frac{1}{D} S(T). \end{aligned} \tag{A.17}$$

Thus,

$$\mathcal{H}(T) = \log K - \frac{1}{2KD} S(T) + \mathbb{E}_v[R(Tv)]. \tag{A.18}$$

It remains to bound the remainder uniformly in T. Let $A := PT$. By Cauchy–Schwarz,

$$\mathbb{E}_v\|Av\|_2^3 \le \sqrt{\mathbb{E}_v\|Av\|_2^2}\sqrt{\mathbb{E}_v\|Av\|_2^4}. \tag{A.19}$$

For $v \sim \mathrm{Unif}(\mathbb{S}^{D-1})$, the standard fourth-moment identity gives

$$\begin{aligned} \mathbb{E}_v\|Av\|_2^4 &= \mathbb{E}_v\big[(v^\top A^\top A v)^2\big] \\ &\le \frac{3\|A\|_F^4}{D(D+2)} \\ &\le \frac{3\|A\|_F^4}{D^2}, \end{aligned} \tag{A.20}$$

while

$$\mathbb{E}_v\|Av\|_2^2 = \frac{\|A\|_F^2}{D}. \tag{A.21}$$

Combining the two bounds yields

$$\begin{aligned} \mathbb{E}_v\|PTv\|_2^3 &\le \sqrt{3}\,\frac{\|PT\|_F^3}{D^{3/2}} \\ &= \sqrt{3}\,\frac{S(T)^{3/2}}{D^{3/2}}. \end{aligned} \tag{A.22}$$

Moreover, since

$$S(T) = K - K\|\mu(T)\|_2^2, \tag{A.23}$$

$S(T)$ is bounded between 0 and K. That is,

$$0 \le S(T) \le K. \tag{A.24}$$

Therefore,

$$\mathbb{E}_v\|PTv\|_2^3 = O(D^{-3/2}), \tag{A.25}$$

uniformly over all admissible text features T, and hence

$$\mathcal{H}(T) = \log K - \frac{1}{2KD} S(T) + O(D^{-3/2}). \tag{A.26}$$

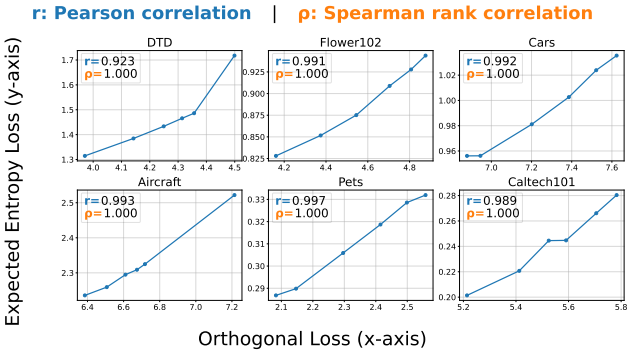

Figure A.1. Correlation between expected EM loss and regularization loss.

Finally, we rewrite this expansion in terms of the corresponding regularizer.

For Eq. (2), since $L_{\mathrm{disp}}(T) = -S(T)$, we have

$$\mathcal{H}(T) = \frac{1}{2KD} L_{\mathrm{disp}}(T) + \log K + O(D^{-3/2}). \tag{A.27}$$

Thus, in this case,

$$\alpha_{\mathrm{disp}} = \frac{1}{2KD}, \qquad \beta_{\mathrm{disp}} = \log K. \tag{A.28}$$

For Eq. (3), since

$$L_{\mathrm{orth}}(T) = \frac{K}{2}\big(K - 1 - S(T)\big),$$

we obtain

$$S(T) = K - 1 - \frac{2}{K} L_{\mathrm{orth}}(T). \tag{A.29}$$

Substituting this into the entropy expansion gives

$$\begin{aligned} \mathcal{H}(T) = &\frac{1}{K^2 D} L_{\mathrm{orth}}(T) \\ &+ \left(\log K - \frac{K-1}{2KD}\right) + O(D^{-3/2}). \end{aligned} \tag{A.30}$$

Thus, in this case,

$$\alpha_{\mathrm{orth}} = \frac{1}{K^2 D}, \qquad \beta_{\mathrm{orth}} = \log K - \frac{K-1}{2KD}. \tag{A.31}$$

Therefore, for either of the regularizers associated with Eq. (2) and Eq. (3), the expected EM loss satisfies

$$\mathcal{H}(T) = \alpha L_{\mathrm{reg}}^{\mathrm{th}}(T) + \beta + O(D^{-3/2}), \qquad \alpha > 0. \tag{A.32}$$

Hence, up to an $O(D^{-3/2})$ remainder, the expected EM loss increases with the corresponding regularization loss. This completes the proof. □

Fig. A.1 further supports this relationship across the settings used in our experiments.

B. Connection Between the Proposed Flatness Loss and Sharpness of the Loss Landscape

Our goal is to clarify how the flatness loss $\mathcal{L}_{\text{flat}}$ in Eq. (9) helps reduce the sharpness of other differentiable losses derived from the model output. Specifically, we aim to show that constraining the deviation of the model output $f_T(C;\theta)$ under small perturbations effectively decreases the sharpness of any loss function defined in terms of the model output, $\ell(f_T(C,\theta))$, with respect to (C,θ).

To formalize this connection, we first analyze how small perturbations in (C,θ) affect the model output. In particular, we approximate $f_T(C+\varepsilon_1;\,\theta+\varepsilon_2)$ using a second-order Taylor expansion:

$$f_T(C+\varepsilon_1;\,\theta+\varepsilon_2) \approx f_T(C;\theta) + J_C\varepsilon_1 + J_\theta\varepsilon_2 + \tfrac{1}{2}\varepsilon_1^\top H_{CC}\varepsilon_1 + \varepsilon_1^\top H_{C\theta}\varepsilon_2 + \tfrac{1}{2}\varepsilon_2^\top H_{\theta\theta}\varepsilon_2, \quad \text{(B.1)}$$

where $J_C = \partial f_T/\partial C$ and $J_\theta = \partial f_T/\partial\theta$ denote the Jacobians of the model output. H_{CC}, $H_{C\theta}$, and $H_{\theta\theta}$ represent the corresponding Hessian blocks.

For a small deviation $\Delta f_T = f_T(C+\varepsilon_1,\,\theta+\varepsilon_2) - f_T(C,\theta)$, the flatness loss $\mathcal{L}_{\text{flat}}$ based on the cosine distance can be locally approximated as

$$\mathcal{L}_{\text{flat}} = \text{dist}_{\cos}(f_T+\Delta f_T,\,f_T) \approx \frac{1}{2\|f_T\|^2}\,\|\Delta f_T\|^2, \quad \text{(B.2)}$$

where this approximation ignores higher-order terms and the dependence on the precise alignment between Δf_T and the normalized direction $u = f_T/\|f_T\|$. This local approximation shows that the flatness loss grows quadratically with the magnitude of the output deviation.

By substituting the Taylor expansion in Eq. (B.1) into Eq. (B.2) and taking expectations over zero-mean isotropic Gaussian perturbations $\varepsilon_1, \varepsilon_2$, we obtain:

$$\mathbb{E}_{\varepsilon_1,\varepsilon_2}[\mathcal{L}_{\text{flat}}] \propto \underbrace{\|J_C\|_F^2}_{\text{input deviation}} + \underbrace{\|J_\theta\|_F^2}_{\text{parameter deviation}} + \underbrace{\|H_{C\theta}\|_F^2}_{\text{cross sharpness}} + \underbrace{\|H_{CC}\|_F^2}_{\text{input sharpness}} + \underbrace{\|H_{\theta\theta}\|_F^2}_{\text{parameter sharpness}}. \quad \text{(B.3)}$$

For simplicity, we ignore constant factors and higher-order terms and view Eq. (B.3) as capturing the dominant dependence of $\mathbb{E}[\mathcal{L}_{\text{flat}}]$ on the derivatives of f_T. This expression shows that the expectation of the flatness loss is proportional to the sum of the squared magnitudes of both first- and second-order derivatives of the model output. Consequently, $\mathcal{L}_{\text{flat}}$ explicitly penalizes output-level sharpness by reducing excessive sensitivity with respect to both inputs C and parameters θ.

Let the scalar loss be defined as $\ell(f_T(C,\theta)) = \phi(f_T(C;\theta),v)$, where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a differentiable mapping and v is arbitrary image features. Denote $g = \nabla_f\phi\big(f_T(C;\theta),v\big)$ as the gradient of ℓ with respect to the model output. By applying the multivariate chain rule, the Hessian of ℓ with respect to C and θ can be expressed as:

$$\begin{aligned}\nabla^2_{CC}\ell &= J_C^\top H_f J_C + g^\top H_{CC},\\ \nabla^2_{\theta\theta}\ell &= J_\theta^\top H_f J_\theta + g^\top H_{\theta\theta},\end{aligned} \quad \text{(B.4)}$$

where H_f denotes the Hessian of the function ϕ with respect to the output f_T. The terms J_C, J_θ, H_{CC} and $H_{\theta\theta}$ correspond to the Jacobians and Hessians of $f_T(C;\theta)$ with respect to C and θ, as defined in Eq. (B.1).

Consequently, minimizing flatness loss $\mathcal{L}_{\text{flat}}$—which reduces $\|J_C\|_F$, $\|J_\theta\|_F$, $\|H_{CC}\|_F$, and $\|H_{\theta\theta}\|_F$—tends to diminish the corresponding components of the Hessian of $\ell(f_T(C,\theta))$. Because the Hessian reflects the sharpness of the loss surface, reducing it directly leads to a smoother landscape [48]. Therefore, suppressing the sharpness of the model output inherently smooths the loss landscape, guiding optimization toward a flatter loss landscape.

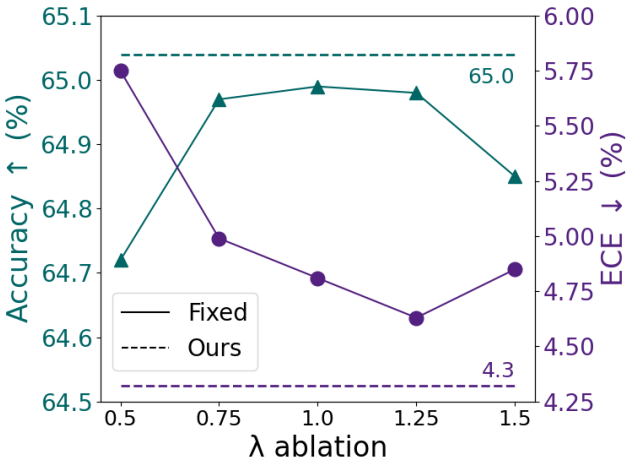

Figure B.1. Comparison between fixed hyperparameter λ and the proposed dynamic λ strategy.

C. Hyperparameter Settings

C.1. Analysis of Lambda

In Eq. (11), λ scales the flatness loss $\mathcal{L}_{\text{flat}}$ during prompt optimization. Rather than using a fixed value of λ across all datasets, we allow it to adapt automatically to dataset-specific characteristics. Concretely, we set $\lambda = \gamma_1 + \frac{\gamma_2}{K}$ with $\gamma_1 = 1.0$ and $\gamma_2 = 0.15$, where K denotes the number of classes. This design reflects the intuition that preserving the original text features becomes more difficult on datasets with larger label spaces; accordingly, diminishing the contribution of the flatness loss helps maintain the underlying text semantics. As shown in Fig. B.1, this dataset-dependent scheduling of λ consistently outperforms a fixed

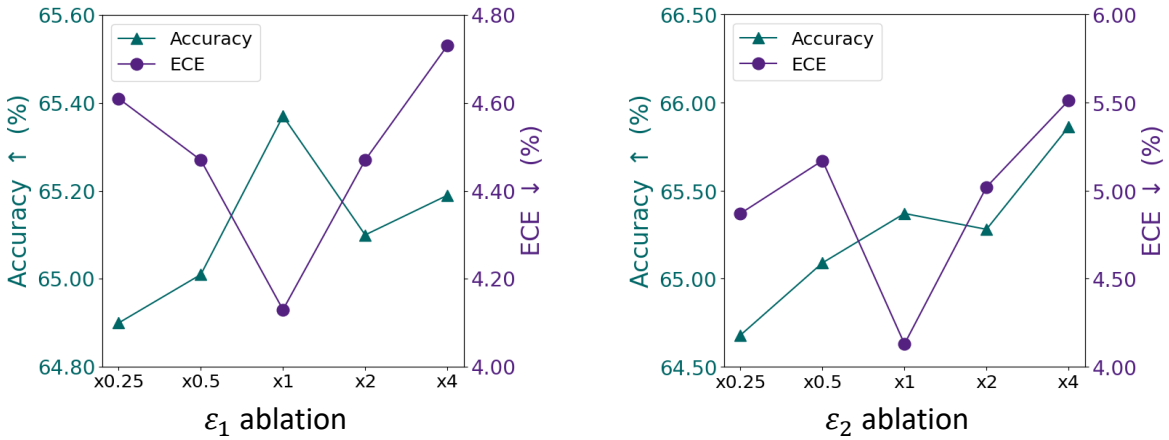

Figure B.2. Ablation on perturbation magnitude. We scale the variances of the Gaussian distributions used to sample perturbations ε_1 and ε_2 and report the resulting performance. The results indicate that tuning the perturbation magnitude within an appropriate range is essential for achieving optimal performance.

Align Loss	Flat Loss	Acc.	ECE
Cosine	L2	64.70	4.87
Cosine	Cosine	64.81	4.91
L2	L2	64.56	4.80
L2	Cosine	**65.37**	**4.13**

Table C.1. Ablation study on distance functions used in the losses of the proposed pretraining stage. The method demonstrates robust performance, achieving stable accuracy and ECE regardless of the function selected.

choice, yielding improvements in both accuracy and ECE across the fine-grained classification datasets.

C.2. Analysis of the Magnitude of Perturbation

In Fig. B.2, we conduct an ablation study on the magnitude of the two perturbations used in the flatness loss. Specifically, starting from our default setting where ε_1 and ε_2 are sampled from Gaussian distributions with variances 0.02 and 0.005, respectively, we scale these variances by constant factors and measure the resulting performance on the fine-grained classification datasets. Here, a scaling factor of $\times 1$ corresponds to our default configuration. Across both experiments, we observe a consistent trend: when the perturbation magnitudes are too small, the model exhibits lower accuracy and higher calibration error, while increasing ϵ_1 and ϵ_2 gradually improves both metrics. However, once the perturbations exceed our default values, the calibration error starts to increase again and the overall performance degrades. This pattern indicates that choosing ε_1 and ε_2 within an appropriate range is crucial, and that our default setting yields a favorable trade-off between accuracy and calibration.

C.3. Analysis of Distance Function

In our pretraining objective, we consider two choices of distance metrics for both the alignment loss in Eq. (10) and the flatness loss in Eq. (9): L2 distance and cosine distance. Tab. C.1 presents an ablation study in which we interchange these distance functions for each loss term while keeping all other components unchanged. Across all configurations, our method consistently delivers higher accuracy than the baseline methods C-TPT and O-TPT, and also achieves lower ECE than C-TPT, demonstrating that the effectiveness of our approach is robust to the specific choice of distance metric.

D. Datasets

Following the baseline papers [33, 41], we adopt the ten fine-grained classification datasets originally introduced in CoOp [45]. These datasets span a broad set of visual domains, including plants and animals (Flower102, OxfordPets), textures (DTD), food imagery (Food101), and scenes (SUN397). They also cover human action recognition (UCF101), satellite imagery analysis (EuroSAT), transportation categories such as cars and aircraft (StanfordCars, Aircraft), and the general-purpose dataset (Caltech101).

In addition, we evaluate on four ImageNet variants that are widely used to assess robustness under natural distribution shifts: ImageNet-V2, ImageNet-A, ImageNet-R, and ImageNet-Sketch. Compared to the standard ImageNet validation set, these datasets respectively contain re-collected samples, naturally adversarial images, artistic renditions (e.g., paintings and cartoons), and sketch drawings, thereby providing a diverse set of challenging test conditions for analyzing both accuracy and calibration.

Method	Metric	Air	Calt	Car	DTD	SAT	FLW	Food	Pets	SUN	UCF	Avg.
O-TPT [33]	AECE	3.96	4.78	1.72	8.21	13.92	4.07	4.81	2.16	8.51	1.96	5.41
	AURC	0.57	0.01	0.14	0.31	0.40	0.10	0.05	0.02	0.18	0.14	0.19
	MCE	13.91	86.10	13.07	17.05	29.25	18.26	11.06	22.57	45.27	8.74	26.53
FPP (Ours)	AECE	7.52	6.24	2.16	7.83	5.36	2.98	1.98	2.46	3.20	3.04	**4.28**
	AURC	0.57	0.01	0.13	0.30	0.31	0.10	0.05	0.02	0.18	0.12	**0.18**
	MCE	27.02	62.33	7.94	20.03	23.52	11.80	6.82	24.78	8.52	10.80	**20.36**

Table D.1. Comparison of calibration metrics (AECE, AURC, MCE) between O-TPT and FPP using the CLIP-ViT/B16 backbone within the TPT framework under a predefined hard prompt ("a photo of a"). The best result for each metric is highlighted in **bold**.

E. Calibration Metrics

Following previous works [33, 41], we evaluate calibration using two metrics: (1) Expected Calibration Error (ECE) [28] and Static Calibration Error (SCE) [30]. Specifically, ECE measures the discrepancy between predicted confidence and accuracy by partitioning predictions into M bins:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|A_b|}{M} \Big|\text{acc}(A_b) - \text{conf}(A_b)\Big|, \tag{E.1}$$

where B is the number of bins used to divide the prediction confidences, A_b denotes the set of samples whose confidence scores fall into the b-th bin, and $|A_b|$ indicates the number of samples in A_b. M is the total number of predictions, $\text{acc}(\cdot)$ represents the prediction accuracy, and $\text{conf}(\cdot)$ is the average confidence of the samples. In addition, the SCE extends this idea to multi-class settings by computing calibration error across both confidence bins and classes:

$$\text{SCE} = \frac{1}{K} \sum_{k=1}^{K} \sum_{b=1}^{B} \frac{|A_{k,b}|}{M} \Big|\text{acc}(A_{k,b}) - \text{conf}(A_{k,b})\Big|. \tag{E.2}$$

Here, $A_{k,b}$ denotes the set of samples belonging to class k whose prediction confidence falls into the b-th bin.

F. Results on Additional Calibration Metrics

We further evaluated our method on a broader set of calibration metrics beyond ECE and SCE, as shown in Tab D.1. Specifically, we compared our method against O-TPT on AECE, AURC, and MCE. AECE [26] provides a more robust estimate of calibration by reducing dependence on a fixed binning scheme. AURC [6] captures the overall risk-coverage trade-off when predictions are ranked by confidence. MCE [28] measures the largest calibration error across bins. Our method consistently achieved better results than O-TPT across all three metrics.

G. Additional Sharpness Analysis

To further support our claim that the proposed method converges to flatter minima, we provide an additional sharpness analysis in Tab. F.1. Specifically, we compare the Fisher-Rao norm of O-TPT and FPP within the TPT framework. The results show that FPP consistently achieves smaller values than O-TPT across all datasets, reducing the average score from 0.681 to 0.137.

H. Standard Deviation Across Different Seeds

In Tab. F.2, following the evaluation protocol of O-TPT, we report the standard deviation of accuracy and ECE across five fine-grained datasets. For three different random seeds, we independently perform both the pretraining and test-time adaptation (TTA) stages, compute the standard deviation for each dataset, and then report the mean standard deviation across datasets. Specifically, we compute the standard deviation over three different random seeds for both the pretraining and TTA stages, and then report the mean standard deviation across datasets. Our method exhibits an accuracy standard deviation comparable to O-TPT, while attaining the lowest ECE deviation among all methods. These results indicate that our approach not only achieves strong average performance but also maintains consistently stable behavior across multiple runs with different random initializations.

Method	Metric	Air	Calt	Car	DTD	SAT	FLW	Food	Pets	UCF	Avg.
O-TPT [33]	Fisher-Rao	0.871	0.452	1.939	0.920	0.255	0.279	1.163	0.035	0.213	0.681
FPP (Ours)	Fisher-Rao	0.251	0.048	0.115	0.228	0.244	0.123	0.049	0.057	0.122	**0.137**

Table F.1. Comparison of Fisher-Rao flatness between O-TPT and FPP using the CLIP-ViT/B16 backbone within the TPT framework under a predefined hard prompt ("a photo of a"). Lower values indicate a flatter loss landscape.

Method	Metric	Calt	Car	DTD	FLW	Food	Avg.
C-TPT	Std. Acc.	0.12	0.16	0.16	0.22	0.20	0.17
	Std. ECE	0.24	0.18	0.24	0.12	0.19	0.194
O-TPT	Std. Acc.	0.14	0.11	0.03	0.10	0.19	**0.11**
	Std. ECE	0.17	0.25	0.14	0.20	0.10	0.177
FPP (Ours)	Std. Acc.	0.05	0.05	0.27	0.10	0.18	0.13
	Std. ECE	0.12	0.27	0.15	0.21	0.04	**0.158**

Table F.2. Standard deviation across three different seed runs.