# Structured Semantics Meet Uncertain Visuals: A Unified Approach to Calibrated Test-Time Prompt Tuning

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Large vision-language models (VLMs) generalize well zero-shot but become overconfident and poorly calibrated under distribution shifts. Existing test-time adaptation (TTA) methods largely apply uniform entropy minimization with fixed geometric regularizers, ignoring instance-wise uncertainty and domain-specific visual cues. We propose **Uncertainty-Calibrated Test-Time Prompt Tuning (UC-TPT)**, a label-free TTA framework targeted at improving reliability rather than solely maximizing accuracy. UC-TPT consists of three theoretically motivated components: (i) lightweight visual-to-text conditioning that injects *shallow* visual statistics—where shift is most pronounced—into prompts, yielding domain-conditioned predictions; (ii) an *uncertainty-tempered* entropy objective that adaptively controls distribution sharpness to curb overconfidence; and (iii) a topology-aware prompt regularizer that approximately preserves the pairwise semantic relations of manual prompts, stabilizing adaptation in the pretrained embedding space. Experiments on CLIP and BiomedCLIP across diverse benchmarks demonstrate that UC-TPT **consistently outperforms existing methods in calibration robustness**, yielding significant reductions in Expected Calibration Error (ECE) across a wide range of distribution shifts while maintaining competitive classification accuracy. [1]

## 1 Introduction

Despite the strong generalization of large-scale Vision–Language Models (VLMs) Radford et al. (2021a); Jia et al. (2021); Yang et al. (2022), their reliability can degrade sharply under real-world distribution shifts (e.g., imaging conditions, sensors, environments), limiting deployment in safety-critical settings such as autonomous driving and medical diagnosis. *Test-time adaptation (TTA)* Liang et al. (2025); Xiao & Snoek (2024) mitigates this by adapting at inference using only unlabeled target data, avoiding retraining and supervision. For VLMs, *Test-Time Prompt Tuning (TPT)* Shu et al. (2022) is especially attractive, as it updates only a small set of prompt tokens while keeping pretrained encoders frozen.

However, existing TPT/TTA pipelines expose a key research gap: they primarily adapt *confidence geometry* without explicitly controlling *when* and *where* updates should occur. Early TPT Shu et al. (2022) relies on entropy minimization, which encourages peaky posteriors and can amplify overconfidence on ambiguous or out-of-distribution (OOD) inputs, harming *calibration* Guo et al. (2017). Later methods add global geometric regularizers—C-TPT Yoon et al. (2024) (isotropic dispersion), O-TPT Sharifdeen et al. (2025) (orthogonality), and A-TPT Ahamed et al. (2026) (angular diversity)—to stabilize predictions, but these constraints remain *static* and *sample-agnostic*: they apply identical update pressure regardless of instance uncertainty and are blind to the evolving target-domain visual distribution. Moreover, the pretrained CLIP space encodes a meaningful *semantic topology*, where distances reflect linguistic relatedness (e.g., "cat"–"tiger"–"leopard"). Rigid constraints such as orthogonality Sharifdeen et al. (2025) or aggressive dispersion Yoon et al. (2024); Ahamed et al. (2026) can distort this structure during adaptation, decoupling semantic similarity from logit

[1] Code will be released upon acceptance.

[2] **ECE**: Expected Calibration Error, **SCE**: Static Calibration Error, **MCE**: Maximum Calibration Error, **ACE**: Adaptive Calibration Error, **Brier score**: Mean squared probability error, **NLL**: Negative Log-Likelihood. See **Appendix E** for further details, All results in Fig. 1 are averaged over 9 image-classification datasets, excluding ImageNet with CLIP-ViT B/16. See Table 1 for details.

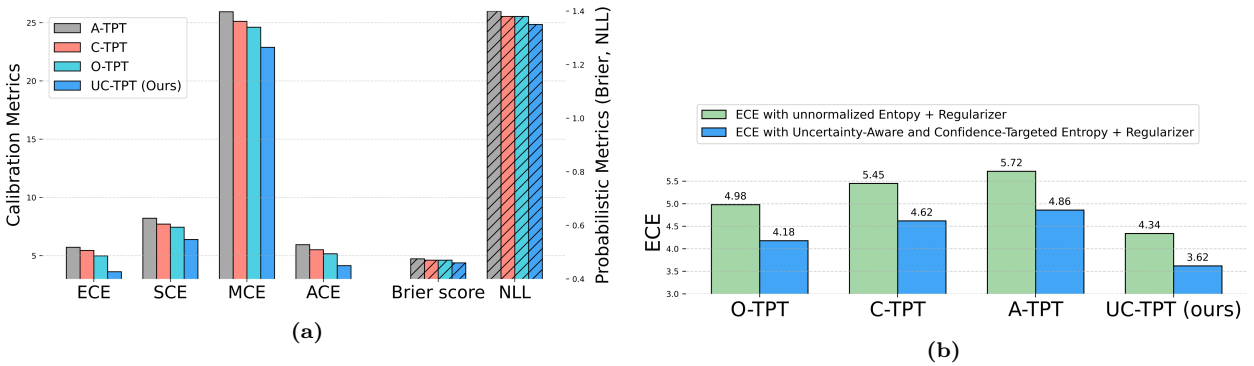



**Figure 1:** (a) **UC-TPT** consistently achieves superior calibration across multiple metrics (↓)[2]surpassing TPT variants C-TPT Yoon et al. (2024), O-TPT Sharifdeen et al. (2025), and A-TPT Ahamed et al. (2026). (b) Replacing the unnormalized entropy in prior methods with our adaptive, uncertainty weighted formulation (Section 3.3.1) along with their respective regularizer markedly reduces miscalibration, showing that ignoring representational uncertainty hinders TTA. Adding our topology-weighted prompt regularizer (Section 3.3.2) further yields the lowest ECE (↓) in our case.

geometry and leading to unstable logit scaling and calibration drift under shift (Fig. 1a–b, Fig. 5). These observations suggest that reliable TPT needs (i) *uncertainty-aware* updates (to avoid sharpening unreliable predictions), (ii) *visual grounding* (to condition prompts on target-domain evidence), and (iii) *topology preservation* (to prevent semantic neighborhood collapse).

**Our Proposal.** We present **UC-TPT**, a unified framework that improves test-time prompt tuning with a *primary goal of reliability and calibration* (rather than maximizing accuracy) along three theoretically justified axes: (i) *visual-to-textual shallow conditioning*, (ii) *adaptive uncertainty-guided entropy optimization*, and (iii) *topology-weighted prompt regularization*.

First, UC-TPT introduces **shallow visual-to-text prompt conditioning**—a minimal, domain-grounded mechanism that injects early-layer visual shift statistics into prompt tokens via a lightweight cross-modal projection, yielding *instance-conditioned* prompts while keeping both encoders frozen (Fig. 7).

Second, UC-TPT proposes an **uncertainty-calibrated entropy objective** that *replaces uniform sharpening* with *risk-aware* test-time optimization: a confidence-shaped target entropy sets the desired sharpness, while a sigmoid uncertainty gate attenuates gradients from ambiguous/OOD samples, preventing overconfident collapse and improving calibration stability (Fig. 1a–b).

Third, UC-TPT introduces a **topology-weighted prompt diversity regularizer** that *preserves CLIP's semantic neighborhood structure* during adaptation by scaling repulsion according to manual-prompt semantic distances. Unlike hard orthogonality Sharifdeen et al. (2025) or uniform angular dispersion Ahamed et al. (2026), this soft cosine-weighted constraint improves separability *without* topology-breaking drift, stabilizing inter-class logit geometry and reducing gradient volatility (Fig. 1b, Fig. 4b, Fig. 5).

Together, these contributions provide a **principled decomposition of test-time prompt tuning for calibration**: visual conditioning supplies *domain evidence*, uncertainty calibration decides *update magnitude per instance*, and topology weighting constrains *update direction* to respect pretrained semantics. This synergy yields stable test-time dynamics and consistently better calibration across datasets and backbones (Fig. 1a–b).

**Our major contributions are:**

- We propose **UC-TPT**, an uncertainty-calibrated test-time prompt tuning framework that targets *reliable confidence* under distribution shifts.
- UC-TPT combines *shallow visual conditioning* for domain-aware grounding, an *uncertainty-guided entropy objective* for selective confidence updates, and a *topology-weighted regularizer* that preserves semantic geometry while encouraging diversity.
- Extensive evaluations show that UC-TPT achieves state-of-the-art calibration while maintaining robust accuracy across architectures and benchmarks, advancing multimodal test-time adaptation.

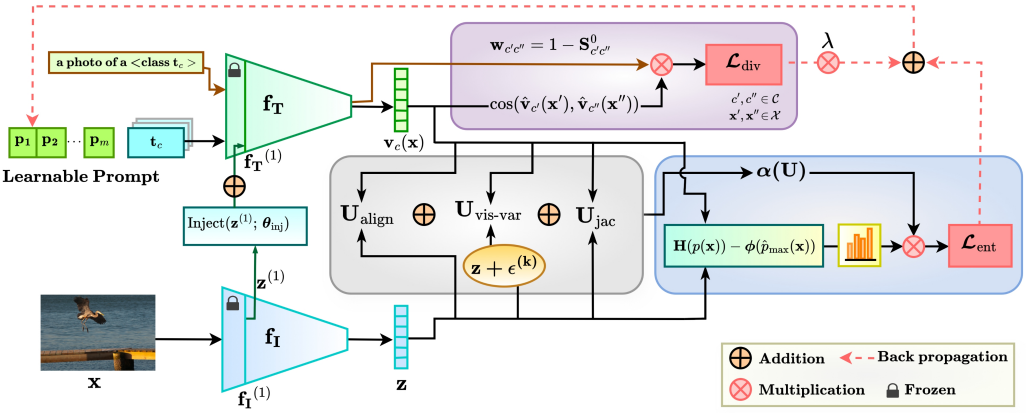



**Figure 2:** Overview of the **UC-TPT** framework. Given an unseen test sample $\mathbf{x}$, UC-TPT performs instance-wise adaptation of textual prompts $\mathbf{s_c(P)}$ through uncertainty-guided updates without requiring labeled data. The frozen CLIP encoders with visual conditioning (Inject($\mathbf{z}$); $\boldsymbol{\theta}_{\text{inj}}$) extract visual $\mathbf{z}$ and textual $\mathbf{v}_c(\mathbf{x})$ embeddings. An uncertainty module estimates alignment-based ($\mathbf{U}_{\text{align}}$), variance-based ($\mathbf{U}_{\text{vis-var}}$), and Jacobian-based ($\mathbf{U}_{\text{jac}}$) uncertainty components, fused via a gating function $\boldsymbol{\alpha}(\cdot)$ to modulate the adaptive confidence-aware entropy loss $\boldsymbol{\mathcal{L}}_{\text{ent}}$. Simultaneously, the semantics preserving diversity loss $\boldsymbol{\mathcal{L}}_{\text{div}}$ enforces discriminative consistency among prompts. Together, these components yield uncertainty-aware, topology-stable prompt adaptation that enhances calibration.

## 2 Related Works

**(a) Test-Time Tuning.** Deep neural networks achieve impressive accuracy yet remain vulnerable under distribution shifts Shimodaira (2000). TTA addresses this by allowing pretrained models to adapt online using unlabeled target data, without retraining. Prior TTA research has largely focused on entropy minimization Wang et al. (2020); Han et al. (2025) and consistency regularization Wang et al. (2022); Liu et al. (2023) or other confidence estimation for standard classifiers Lee et al. (2024). Within this paradigm, TPT Shu et al. (2022) optimizes prompt tokens with entropy minimization and view augmentation, achieving lightweight adaptation for VLMs and inspiring several extensions Xiao et al. (2025); Sheng et al. (2025); Feng et al. (2023).

**(b) Uncertainty Estimation.** Uncertainty quantification in deep models has been approached via Bayesian networks and approximate inference, e.g., Monte Carlo dropout for epistemic uncertainty Gal & Ghahramani (2016), or ensemble-based methods capturing predictive variance Lakshminarayanan et al. (2017). Predictive entropy provides a scalar confidence proxy Depeweg et al. (2017), while perturbation-based methods estimate both aleatoric and epistemic uncertainty Ovadia et al. (2019). In VLMs, alignment confidence between image–text embeddings serves as a natural uncertainty cue Shu et al. (2022), and prompt sensitivity has been shown to correlate with calibration quality Yoon et al. (2024). Bayesian prompt ensembles Tonolini et al. (2024); Daneshfar et al. (2025), stochastic embedding models Pautsch et al. (2023); Erick et al. (2024), and Bayesian adapters Alvarez et al. (2024) further enhance reliability. Yet, most of these approaches remain offline, with limited exploration of uncertainty modeling in online or test-time prompt tuning settings across domains.

**(c) Calibration in VLMs.** CLIP-based models exhibit strong zero-shot generalization but often become miscalibrated under domain shifts, where prediction confidence diverges from correctness. Previous works address this via temperature scaling using text–class distances Wang et al. (2024), unseen-label augmentation Wang et al. (2025), dominant-dimension suppression Han & Hwang (2025), or covariance-regularized tuning Oh et al. (2024). Within TPT frameworks, calibration is commonly enforced through geometric constraints such as dispersion Yoon et al. (2024); Ahamed et al. (2026) or orthogonality Sharifdeen et al. (2025), while cross-regularized tuning Li et al. (2023) and logit-range normalization Murugesan et al. (2024) further stabilize adaptation. However, these methods apply uniform regularization, ignoring instance-specific uncertainty and often producing overconfident predictions under shift. Moreover, rigid geometric constraints

(e.g., angular orthogonality) distort CLIP's pretrained semantic topology, weakening inter-class relationships crucial for calibrated reasoning.

### 2.1 Problem Formulation and Background

We consider a pretrained vision–language model (VLM), such as CLIP Radford et al. (2021c), consisting of a frozen image encoder $\mathbf{f}_I : \mathcal{X} \to \mathbb{R}^d$ and a frozen text encoder $\mathbf{f}_T : \mathcal{T} \to \mathbb{R}^d$, where $\mathcal{X}$ and $\mathcal{T}$ denote the image and text spaces, respectively. Both encoders map inputs to a shared $d$-dimensional embedding space. Let $\mathcal{C} = \{1, \ldots, C\}$ denote the set of semantic classes. Each class $c \in \mathcal{C}$ is associated with a textual template $\mathbf{t}_c \in \mathcal{T}$.

**Prompt Tokens and Class Representations.** Prompt tuning augments each class template with a sequence of $m$ learnable prompt tokens $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_m] \in \mathbb{R}^{m \times d}$. The resulting class-specific prompt is defined as

$$\mathbf{s}_c(\mathbf{P}) = [\mathbf{p}_1, \ldots, \mathbf{p}_m, \mathbf{t}_c], \tag{1}$$

which is encoded by the text encoder to produce a class embedding

$$\mathbf{v}_c = \mathbf{f}_T\big(\mathbf{s}_c(\mathbf{P})\big) \in \mathbb{R}^d. \tag{2}$$

**Prediction Model.** Given an image $\mathbf{x} \in \mathcal{X}$, the image encoder outputs a normalized visual embedding $\mathbf{z} = \mathbf{f}_I(\mathbf{x}) \in \mathbb{R}^d$. The model predicts class probabilities via a temperature-scaled softmax over cosine similarities:

$$p_c(\mathbf{x}; \mathbf{P}) = \frac{\exp\big(\mathbf{z}^\top \mathbf{v}_c / \tau\big)}{\sum_{c'=1}^{C} \exp(\mathbf{z}^\top \mathbf{v}_{c'} / \tau)}, \tag{3}$$

where $\tau > 0$ denotes a temperature parameter.

**Test-Time Adaptation (TTA).** At test time, the model is exposed to unlabeled target samples $\mathcal{D}_t = \{\mathbf{x}_j\}_{j=1}^n$ drawn from an unknown target distribution $\mathcal{P}_t$. Following the test-time prompt tuning (TPT) paradigm Shu et al. (2022); Sharifdeen et al. (2025), adaptation is performed by optimizing only the prompt parameters $\mathbf{P}$, while keeping both encoders frozen. The test-time adaptation objective is formulated as

$$\min_{\mathbf{P}} \; \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_t}\big[\mathcal{L}_{\mathrm{TTA}}(\mathbf{x}; \mathbf{P})\big], \tag{4}$$

where $\mathcal{L}_{\mathrm{TTA}}$ is a self-supervised loss, typically based on entropy minimization or its regularized variants.

**Calibration and Confidence Reliability.** Calibration measures the alignment between predictive confidence and empirical accuracy Guo et al. (2017). In practice, it is quantified using Expected Calibration Error (ECE), which compares accuracy and confidence across confidence bins (see Appendix E for the formal derivation). Poor calibration is particularly harmful in test-time prompt tuning, where entropy minimization can over-sharpen uncertain predictions under domain shift, leading to overconfident errors.

## 3 Taking through the UC-TPT Framework

Our UC-TPT is a unified, calibration-aware prompt tuning framework that performs uncertainty-guided adaptation of vision–language models at test time (Fig. 2). It couples instance-aware visual conditioning (Section 3.1), multi-signal uncertainty estimation (Section 3.2), and adaptive entropy optimization with a prompt regularizer that preserves linguistic geometry (Section 3.3). The subsequent sections describe each component in detail.

### 3.1 Shallow Visual-to-Textual Conditioning

Conventional TPT methods Shu et al. (2022); Sharifdeen et al. (2025) adapt prompts with a *single, sample-agnostic* parameterization, applying identical updates to all test images. Under domain shift, this is limiting: the mismatch largely resides in *visual* features, yet adaptation occurs only in prompt space without injecting target-domain evidence into text representations.

A natural remedy is stronger cross-modal coupling. MaPLe Khattak et al. (2023), for example, couples text and image branches across multiple layers, including text→image pathways. While effective for supervised prompt learning, such deep bi-directional coupling is less suited for *test-time* tuning: it increases update capacity and lets transient test-time signals steer visual features, potentially distorting CLIP's pretrained geometry and destabilizing calibration (higher ECE in Table 15).

UC-TPT instead adopts *minimal conditioning*: make prompts input-dependent via a *one-way, shallow* visual→text injection while keeping both encoders frozen. By injecting early-layer shift statistics—where shift is strongest and least entangled with class semantics—UC-TPT grounds adaptation without deep coupling. Fig. 3 shows shallow injection improves calibration over deeper injection and no-conditioning, yielding a favorable stability–adaptivity trade-off.

**Definition 3.1** (Shallow Visual-to-Textual Injection)**.** Let $\mathbf{f_I}^{(1)}$ and $\mathbf{f_T}^{(1)}$ denote the first layers of the frozen image and text encoders. Given an image $\mathbf{x}$, the first-layer visual embedding is

$$\mathbf{z}^{(1)} = \mathbf{f_I}^{(1)}(\mathbf{x}) \in \mathbb{R}^d. \tag{5}$$

A lightweight MLP projection $\text{Inject}(\cdot; \boldsymbol{\theta}_{\text{inj}})$ with `GELU` (details in **Appendix B**) maps $\mathbf{z}^{(1)}$ to

$$\mathbf{q_x} = \text{Inject}(\mathbf{z}^{(1)}; \boldsymbol{\theta}_{\text{inj}}) \in \mathbb{R}^{1\times m}. \tag{6}$$

Let $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^m$ be base prompt tokens and $\mathbf{P}^{(1)} = \mathbf{f_T}^{(1)}([\mathbf{p}_1, \ldots, \mathbf{p}_m])$ their first-layer embeddings. The conditioned prompt is

$$\tilde{\mathbf{P}}^{(1)}(\mathbf{x}) = [\mathbf{P}_1^{(1)} + \mathbf{q}_{\mathbf{x}1}, \ \ldots, \ \mathbf{P}_m^{(1)} + \mathbf{q}_{\mathbf{x}m}]. \tag{7}$$

For class template $\mathbf{t}_c$, the conditioned class embedding is

$$\mathbf{v}_c(\mathbf{x}) = \mathbf{f_T}^{(>1)}(\mathbf{s}_c(\tilde{\mathbf{P}}^{(1)}(\mathbf{x}))), \tag{8}$$

where $\mathbf{s}_c$ is defined in Eq. 1; prediction follows Eq. 3.

**Remark.** UC-TPT transforms *static* prompts into *instance-conditioned* ones through lightweight visual conditioning, incurring negligible computational cost ($< 0.1\%$) while keeping the visual backbone and deeper text layers frozen. By computing a single gradient step over augmented views, UC-TPT extracts a robust visual bias for the test time update without the instability typically associated with extended optimization. This design ensures a fast, lightweight adaptation that preserves the calibrated nature of the pretrained embedding space.

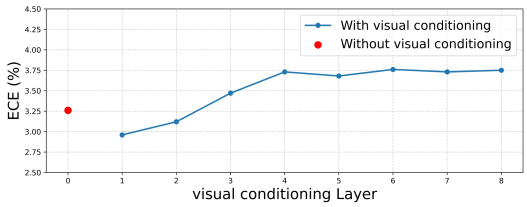



**Figure 3:** Effect of visual conditioning depth on ECE for Caltech101 Fei-Fei et al. (2004) using CLIP–ViT/B/16, comparing shallow conditioning to deeper and unconditioned variants.

### 3.2 Per-Sample Uncertainty Quantification

Prior TPT methods Shu et al. (2022); Ahamed et al. (2026); Sharifdeen et al. (2025) apply *uniform* adaptation, implicitly assuming equal reliability across test samples. Under distribution shift, instances vary widely in ambiguity and local stability; uniform entropy minimization indiscriminately sharpens posteriors, often *amplifying overconfidence* on hard or OOD inputs and degrading calibration.

UC-TPT instead estimates *per-sample uncertainty* to control test-time updates. Rather than using heavy post-hoc Bayesian tools or many stochastic forward passes Brahma & Rai (2023); Tan et al. (2024), we design an *on-the-fly*, label-free estimator that is (i) *multimodal* (image–text agreement and predictive stability), (ii) *optimization-aware* (prompt-update fragility), and (iii) *lightweight* (negligible overhead beyond standard TPT).

**Why uncertainty, and what should it measure?** The key failure mode under shift is *misplaced confidence*: confident errors arise when inputs are off-manifold, image–text alignment is weak, or predictions

are unstable to small perturbations. Accordingly, UC-TPT models three complementary "adaptation risks": (i) *semantic* risk (weak alignment to any class text prototype), (ii) *predictive* risk (output variance under small visual changes), and (iii) *optimization* risk (sensitivity to prompt updates). These cues act at representation, prediction, and update levels, providing non-redundant uncertainty for selective adaptation.

**Unified uncertainty score.** Given a test sample $\mathbf{x} \in \mathcal{X}$ with visual embedding $\mathbf{z}$ and class text embeddings $\mathbf{v}_c(\mathbf{x})$ (Eq. 8), we define the total uncertainty $\mathbf{U}(\mathbf{x})$ as the sum of three efficiently computable components.

**Definition 3.2** (Image–Text Alignment Uncertainty)**.** If the visual embedding $\mathbf{z}$ is poorly aligned with *all* class text embeddings $\{\mathbf{v}_c(\mathbf{x})\}$, the sample is likely off-manifold or semantically ambiguous. We define

$$\mathbf{U}_{\text{align}}(\mathbf{x}) = 1 - \max_{c \in \mathcal{C}} \cos(\mathbf{z}, \mathbf{v}_c(\mathbf{x})), \tag{9}$$

where larger values indicate weaker multimodal agreement (higher semantic risk).

**Definition 3.3** (Visual Perturbation Variance)**.** A reliable predictor should be locally stable: small perturbations in the visual representation should not cause large changes in the output distribution. We inject Gaussian noise into $\mathbf{z}$ to obtain perturbed embeddings

$$\mathbf{z}^{(k)} = \mathbf{z} + \boldsymbol{\epsilon}^{(k)}, \quad \boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{10}$$

and compute

$$p^{(k)} = \text{softmax}\big(\cos(\mathbf{z}^{(k)}, \{\mathbf{v}_c(\mathbf{x})\}_{c=1}^{C})\big). \tag{11}$$

Let $\bar{p}_v = \frac{1}{K}\sum_{k=1}^{K} p^{(k)}$. The visual variance uncertainty is

$$\mathbf{U}_{\text{vis-var}}(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} \|p^{(k)} - \bar{p}_v\|_2^2. \tag{12}$$

This term captures local predictive instability (predictive risk).

**Definition 3.4** (Prompt Sensitivity)**.** Even when predictions are sharp, test-time updates can be brittle: if the loss landscape is steep w.r.t. prompts, small prompt steps can induce large distributional shifts. We quantify this *optimization risk* via the prompt Jacobian:

$$\mathbf{U}_{\text{jac}}(\mathbf{x}) = \sum_{c \in \mathcal{C}} \big(1 - \cos(\mathbf{z},\, \mathbf{v}_c(\mathbf{x}))\big) \left\| \frac{\partial p_c}{\partial \mathbf{P}} \right\|_2, \tag{13}$$

where the alignment term emphasizes gradients along semantically unreliable directions.

**Computational Complexity.** To ensure test-time efficiency, UC-TPT employs several lightweight approximations. $\mathbf{U}_{\text{align}}$ reuses standard CLIP cosine similarities, adding no overhead. $\mathbf{U}_{\text{vis-var}}$ operates directly in the low-dimensional embedding space via $K$ noise samples ($K = 5$). Naïvely, $\mathbf{U}_{\text{jac}}$ would require explicit Jacobian trace computation scaling with the total number of classes $C$; instead, we restrict it to the top-$\tilde{C}$ classes that capture $> 99.9\%$ of the probability mass and approximate the trace with a Hutchinson estimator Avron & Toledo (2011). As a result, the asymptotic complexity is bounded as follows:

**Theorem 3.5** (UC-TPT Complexity)**.** *Let $N$ denote the batch size, $C_B \ll C$ the active classes in a batch, $d$ the embedding dimension, and $K$ the number of perturbations. The total computational complexity of UC-TPT is:*

$$\mathcal{C}_{\text{UC-TPT(H)}} = \mathcal{O}(KN\tilde{C} + C_B^2 d + Nd)$$

By leveraging $\tilde{C} \ll C$ and $C_B \ll C$, this formulation entirely bypasses the strict $\mathcal{O}(C^2 d)$ bottleneck of prior class-interaction regularizers (e.g., A-TPT, O-TPT). A full formal derivation (including intermediate lemmas) and empirical validation demonstrating that UC-TPT maintains peak memory and runtime comparable to baselines while significantly improving calibration are provided in Appendix H.

**Total uncertainty.** The final per-sample uncertainty is

$$\mathbf{U}(\mathbf{x}) = \mathbf{U}_{\text{align}}(\mathbf{x}) + \mathbf{U}_{\text{vis-var}}(\mathbf{x}) + \mathbf{U}_{\text{jac}}(\mathbf{x}). \tag{14}$$

**Theoretical insight and fusion strategy.** The three terms in Eq. 14 upper-bound complementary sources of harmful adaptation: $\mathbf{U}_{\text{align}}$ detects samples with weak multimodal agreement (high risk of confident mismatch), $\mathbf{U}_{\text{vis-var}}$ estimates local output sensitivity (a proxy for the local Lipschitzness of the predictor around $\mathbf{z}$), and $\mathbf{U}_{\text{jac}}$ measures the sensitivity of predictions to prompt parameters (controlling step safety in prompt space). Together, $\mathbf{U}(\mathbf{x})$ serves as a lightweight surrogate for the *expected calibration risk* of applying entropy-sharpening updates on $\mathbf{x}$, enabling UC-TPT to suppress updates when either semantic support, local stability, or optimization stability is lacking.

We intentionally adopt direct additive fusion. Conceptually, because these metrics track independent failure modes, summation acts as a continuous logical "OR" gate: if any single risk factor (e.g., semantic mismatch) spikes, the total uncertainty correctly scales to protect the model. Empirically, this unweighted addition proves highly robust across distribution shifts, avoiding the manual tuning or magnitude loss associated with weighted or rank-based alternatives. A comprehensive ablation validating this aggregation strategy is provided in Appendix F.

### 3.3 Proposed Loss Functions

UC-TPT updates prompts at test time using two complementary objectives: an *Adaptive Uncertainty-Aware Entropy Loss* (Eq. 18) and a *Topology-Weighted Prompt Diversity Regularizer* (Sec. 3.3.2). The first controls *when* and *how strongly* to sharpen predictions (sample-wise reliability), while the second constrains *where* prompts are allowed to move (semantic geometry), yielding stable adaptation by bounding logit variance and stabilizing gradients (Fig. 4(b)).

#### 3.3.1 Adaptive Uncertainty-Aware Entropy Loss

Standard entropy minimization sharpens predictions but can harm calibration by treating all samples equally and propagating gradients from uncertain/OOD inputs. UC-TPT instead formulates test-time entropy optimization as a *heteroscedastic* objective: both the entropy target and the update weight depend on confidence and per-sample uncertainty, preventing indiscriminate over-sharpening.

For $\mathbf{x} \in \mathcal{X}$ with probabilities $p_c(\mathbf{x})$ (Eq. 3), the predictive entropy is

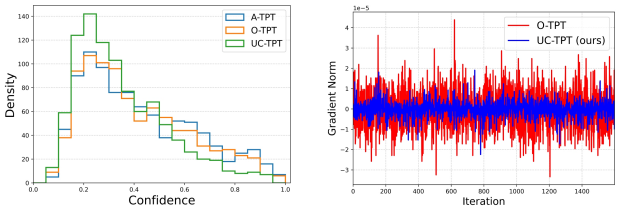
$$\mathbf{H}(p(\mathbf{x})) = -\sum_{c=1}^{C} p_c(\mathbf{x}) \log(p_c(\mathbf{x}) + \nu), \tag{15}$$

where $\nu$ is a stability constant and $\hat{p}_{\max} = \max_c p(y{=}c \mid \mathbf{x}; \mathbf{P})$ denotes confidence. Prior TPT methods Shu et al. (2022); Ahamed et al. (2026); Sharifdeen et al. (2025) directly minimize $\mathbf{H}$, which implicitly drives all samples toward near-delta posteriors. Under shift, this increases the risk of *confident mistakes*. UC-TPT makes two key changes.

**(i) Confidence-shaped entropy target.** We set

$$\phi(\hat{p}_{\max}(\mathbf{x})) = \log C \cdot (1 - \hat{p}_{\max}(\mathbf{x})), \tag{16}$$

which lies in $[0, \log C]$. Intuitively, $\phi$ implements a soft, confidence-aligned target: for confident samples, $\phi$ is small and sharpening is encouraged; for low-confidence samples, $\phi$ remains large, preserving entropy mass and avoiding premature collapse. This can be viewed as imposing a *data-dependent prior* on the posterior sharpness, improving calibration under shift.




**(a)** **(b)**

**Figure 4:** On DTD Cimpoi et al. (2014) with CLIP–ViT-B/16: (a) Max confidence of misclassified samples: UC-TPT shifts errors to lower confidence than Sharifdeen et al. (2025); Ahamed et al. (2026). (b) Gradient-norm variance: $\mathcal{L}_{\text{div}}$ has tighter variance than O-TPT Sharifdeen et al. (2025).

**(ii) Uncertainty-gated update weight.** Let $\mathbf{U}(\mathbf{x})$ be the unified uncertainty (Eq. 14) and $\mathbf{U}_{\text{norm}} = \mathbf{norm}(\mathbf{U}(\mathbf{x}))$ **min-max normalization, (Appendix F).** We define

$$\boldsymbol{\alpha}(\mathbf{U}_{\text{norm}}) = \frac{1}{1 + \exp\big(k_0(\mathbf{U}_{\text{norm}} - u_0)\big)}, \tag{17}$$

where $k_0$ controls the slope and $u_0$ the midpoint. The sigmoid yields smooth, monotone attenuation: reliable samples receive stronger updates and uncertain/OOD samples are down-weighted **(Appendix K).** The smoothness is important for test-time stability: it avoids hard thresholding that can cause abrupt regime switches across nearby samples.

The final objective is

$$\mathcal{L}_{\text{ent}}(\mathbf{x}) = \boldsymbol{\alpha}(\mathbf{U}_{\text{norm}}(\mathbf{x}))\big(\mathbf{H}(p(\mathbf{x})) - \boldsymbol{\phi}(\hat{p}_{\max}(\mathbf{x}))\big)^2. \tag{18}$$

**Theoretical insight (safe sharpening as inverse-variance weighting).** Eq. 18 is a heteroscedastic least-squares form: $\boldsymbol{\alpha}(\mathbf{U})$ plays the role of an *inverse-variance weight*, suppressing gradients from samples with high uncertainty (high estimated noise/instability), while $\boldsymbol{\phi}(\hat{p}_{\max})$ defines a confidence-aligned target entropy. This replaces *uniform* entropy minimization with *risk-aware* sharpening: only samples that are simultaneously confident and stable meaningfully drive adaptation. Consequently, UC-TPT shifts misclassified samples toward lower confidence and yields smoother confidence evolution (Fig. 4(a), Fig. 15(a,b)). Further theoretical insight is in **Appendix I.**

### 3.3.2 Topology-Weighted Prompt Diversity Regularizer

CLIP's manual prompts induce a semantic topology where cosine proximity reflects linguistic relatedness, providing a strong prior for zero-shot generalization. Uniform diversity constraints can violate this prior: strict orthogonality Sharifdeen et al. (2025) forces *equal* separation even for semantically close classes, while uniform angular inflation Ahamed et al. (2026) can over-repel dense semantic neighborhoods, distorting inter-class logit ratios and harming calibration. UC-TPT therefore enforces *topology-aware* diversity: separation strength is proportional to pretrained semantic distance, improving discriminability without breaking local neighborhoods.

**Definition.** Let $\hat{\mathbf{v}}(\mathbf{x}) \in \mathbb{R}^d$ be the $\ell_2$-normalized *conditioned prompt embedding* and $\tilde{c} = \arg\max_c p(c \mid \mathbf{x})$ the pseudo-label. Define

$$\mathbf{w}_{ij} = 1 - \mathbf{S}^0_{\tilde{c}_i \tilde{c}_j}, \qquad \mathbf{S}^0_{\tilde{c}_i \tilde{c}_j} = \cos(\mathbf{v}^0_{\tilde{c}_i}, \mathbf{v}^0_{\tilde{c}_j}), \tag{19}$$

where $\mathbf{v}^0_c$ is the frozen manual-prompt embedding (`"a photo of a <class>"`). For a minibatch $\{\mathbf{x}_i\}_{i=1}^B$,

$$\mathcal{L}_{\text{div}} = \sum_{1 \le i < j \le B} \mathbf{w}_{ij} \cos\big(\hat{\mathbf{v}}(\mathbf{x}_i), \hat{\mathbf{v}}(\mathbf{x}_j)\big). \tag{20}$$

**Theoretical insight (graph-based stabilization).** Eq. 20 can be interpreted as a soft semantic-graph constraint: $\mathbf{S}^0$ defines a target similarity graph from pretrained manual prompts, and the weighting $\mathbf{w}_{ij}$ prevents collapsing edges that should remain close while promoting separation where the prior indicates dissimilarity. This reduces topology-breaking prompt drift and stabilizes the inter-class logit geometry during test-time optimization. Empirically, UC-TPT increases dispersion of pairwise similarities with minimal mean shift (Fig. 5), indicating improved separability without erasing the pretrained semantic scaffold **(Appendix J).**

**Overall Objective.** The final test-time adaptation objective combines uncertainty-aware entropy minimization with topology-weighted prompt diversity, where $\lambda$ balances entropy minimization and prompt diversity:

$$\mathcal{L}_{\text{TTA}} = \mathcal{L}_{\text{ent}} + \lambda\, \mathcal{L}_{\text{div}}. \tag{21}$$

Fig. 1(b) shows consistent ECE reduction when $\mathcal{L}_{\text{ent}}$ is applied across baseline methods (see Table 13), while combining $\mathcal{L}_{\text{ent}}$ & $\mathcal{L}_{\text{div}}$ yields further gains for UC-TPT, highlighting the broader applicability of this principle in TTA.

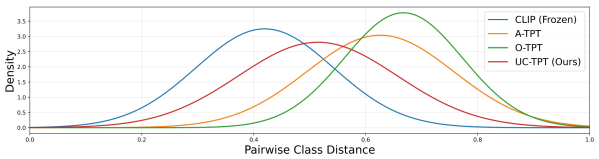



**Figure 5:** On Pets Parkhi et al. (2012) with CLIP–ViT/B-16: fitted Gaussians of pairwise prompt similarities. UC-TPT increases dispersion with minimal mean shift, improving separation while preserving semantics.

## 4 Experimental Evaluations

**Datasets.** We evaluate on a comprehensive suite of datasets encompassing general, fine-grained, robustness, medical, Remote sensing, and cross-domain benchmarks using CLIP and BiomedCLIP. Notably, this includes natural distribution shifts evaluated through ImageNet variants as well as specialized remote sensing and texture tasks. Full dataset descriptions and configuration details are provided in Appendix C.

**Implementation Details, Comparative Methods, and Evaluation Protocol.** We evaluate UC-TPT with CLIP-RN50 and CLIP-ViT-B/16 under a unified test-time prompt tuning (TPT) setup, comparing against calibration-aware and TPT-based baselines under identical settings. All methods use a single AdamW update (lr = $5\times10^{-3}$), hard prompt initialization, and a frozen CLIP backbone. For UC-TPT, following A-TPT, we set $\lambda$ in Eq. 21 to 80, and $k_0$ and $u_0$ in Eq. 17 to 10 and 0.01. Implementation details, hyperparameters, and evaluation protocols are provided in **Appendix D** & **K.** Further analysis of different prompt learning frameworks and different prompt initializations, different combination of regularizers, Batch size analysis, adaptation step analysis, and additional ablations reported in **Appendix**.

**Table 1: ECE** (↓) with CLIP-ViT-B/16 backbone across ten datasets. The overall best-performing result is in **bold**, and the second best is underlined. All baselines are reimplemented under reported settings for fair comparison. Standard deviation information across multiple seeds is provided in the Appendix L.

| Method | Metric | INet | DTD | FLW | Food | Air. | Pets | C101 | UCF | SAT | Car | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-Shot | Acc. | 66.8 | 44.5 | 67.4 | 83.8 | 23.8 | 88.1 | 92.9 | 65.0 | 41.3 | 65.4 | 63.9 |
| | ECE | 2.03 | 8.28 | 2.61 | 2.32 | 5.17 | 4.44 | 5.50 | 3.26 | 7.51 | 4.26 | 4.54 |
| TPT | Acc. | 69.0 | 46.7 | 69.1 | 84.6 | 23.9 | 87.1 | 93.8 | 67.4 | 42.5 | 66.2 | 65.0 |
| | ECE | 10.54 | 21.30 | 13.23 | 3.94 | 16.37 | 5.42 | 4.52 | 12.56 | 21.73 | 5.36 | 11.50 |
| R-TPT | Acc. | 66.7 | 44.5 | 67.2 | 83.8 | 23.9 | 88.1 | 92.9 | 64.9 | 41.4 | 65.3 | 63.9 |
| | ECE | 10.30 | 18.80 | 10.80 | 3.30 | 12.60 | 5.40 | 3.60 | 12.10 | 22.00 | 1.93 | 10.10 |
| C-TPT | Acc. | 68.2 | 46.0 | 69.7 | 83.3 | 23.9 | 88.2 | 93.4 | 65.3 | 43.2 | 65.5 | 64.7 |
| | ECE | 3.19 | 12.50 | 5.13 | 3.72 | 4.33 | 1.83 | 4.34 | 2.40 | 13.30 | 1.56 | 5.22 |
| A-TPT | Acc. | 67.7 | 45.9 | 65.9 | 83.2 | 23.1 | 84.1 | 92.0 | 63.3 | 42.9 | 62.9 | 63.1 |
| | ECE | 2.26 | 6.15 | 4.19 | 7.08 | 6.32 | 5.85 | 10.10 | 3.63 | 4.46 | 5.52 | 5.55 |
| O-TPT | Acc. | 67.3 | 45.6 | 69.2 | 82.8 | 23.3 | 88.0 | 93.1 | 64.2 | 43.3 | 64.9 | 64.2 |
| | ECE | 1.98 | 8.08 | 3.87 | 4.63 | 3.97 | 1.96 | 4.64 | 2.28 | 13.80 | 1.61 | 4.68 |
| **UC-TPT (Ours)** | **Acc.** | 68.2 | 44.3 | 67.8 | 83.3 | 24.0 | 88.4 | 93.4 | 63.7 | 42.4 | 64.4 | 64.0 |
| | **ECE** | **1.53** | **4.74** | 3.97 | 3.23 | **2.65** | 2.65 | **2.96** | 2.70 | 8.36 | **1.35** | **3.41** |

**Table 2: Comparison of average acc. / ECE performance** (↓) with CLIP-RN50 backbone on ten datasets as per Table 1.

| Method | Avg. Acc./ECE |
|---|---|
| Zero-Shot | 55.65 / 5.60 |
| TPT | 57.44 / 11.88 |
| R-TPT | 55.81 / 11.71 |
| C-TPT | 57.25 / 6.66 |
| A-TPT | 55.00 / 7.54 |
| O-TPT | 56.92 / 5.84 |
| **UC-TPT (Ours)** | 56.85 / **4.71** |

**Table 3: Average acc. / ECE performance** (↓) across CLIP-RN50 and ViT-B/16 backbones in natural distribution shift datasets (ImageNet-A, V2, R, S).

| Method | RN50 (Acc./ECE) | ViT-B/16 (Acc./ECE) |
|---|---|---|
| Zero-Shot | 40.60 / 7.18 | 57.16 / 4.90 |
| TPT | 43.53 / 16.75 | 60.20 / 11.70 |
| R-TPT | 40.60 / 16.57 | 57.16 / 11.02 |
| C-TPT | 41.70 / 8.92 | 58.37 / 5.26 |
| A-TPT | 42.93 / 14.22 | 59.34 / 7.95 |
| O-TPT | 41.72 / 8.93 | 58.49 / 5.06 |
| **UC-TPT (Ours)** | 41.88 / **6.88** | 57.81 / **4.63** |

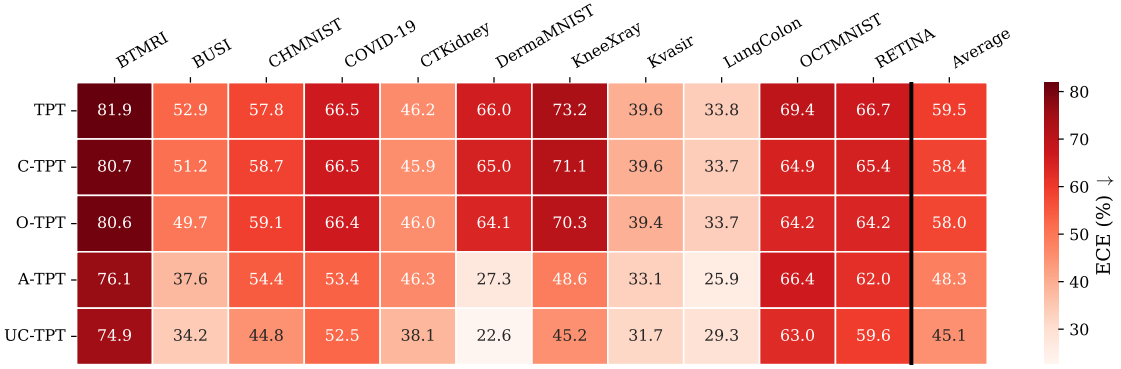



**Figure 6: Comparison of ECE (↓) on** 11 **BioMed datasets** using BioMedCLIP Zhang et al. (2024) as the backbone. UC-TPT achieves the lowest average ECE across all baselines.

### 4.1 Main Results

**Cross-Domain Performance (ViT-B/16).** Table 1 reports calibration across ten diverse cross-domain benchmarks using the CLIP-ViT-B/16 backbone. UC-TPT attains the lowest mean ECE (3.41) while keeping accuracy (64.0%) highly competitive with Zero-Shot (63.9%) and recent test-time adaptation variants. Compared to the strongest baseline, O-TPT (4.68 ECE), UC-TPT achieves an absolute ECE reduction of 1.27 points. This demonstrates that uncertainty-tempered entropy updates, combined with topology-aware regularization, yield significantly more reliable confidence estimates. The improvements are particularly striking on high-variance and texture-heavy datasets such as DTD (ECE reduced from 8.08 to 4.74) and SAT (ECE reduced from 13.80 to 8.36). On these datasets, rigid orthogonality or uniform angular dispersion regularizers (like those in O-TPT and A-TPT) often over-constrain the prompt manifold and exacerbate miscalibration. By down-weighting risky samples via the uncertainty gate, UC-TPT safely preserves CLIP's semantic topology.

**Cross-Domain Performance (RN50).** Table 2 demonstrates that these calibration benefits are architecture-agnostic. When transitioning to a CNN-based CLIP-RN50 backbone across the same ten datasets, UC-TPT consistently achieves the best average ECE (4.71). This outperforms the strongest baseline, O-TPT (5.84), by an absolute margin of 1.13 points. Notably, UC-TPT maintains an average accuracy of 56.85%, fully matching the classification performance of competitive baselines while strictly improving model reliability.

**Natural Distribution Shifts.** Table 3 isolates the evaluation to natural distribution shifts (ImageNet-A, -V2, -R, -S). UC-TPT once again achieves the lowest average ECE across both backbones without any notable loss in classification accuracy when compared to baseline methods. On the RN50 backbone, UC-TPT reduces the average ECE to 6.88, substantially outperforming C-TPT (8.92) and O-TPT (8.93). On the ViT-B/16 backbone, UC-TPT achieves an ECE of 4.63, outperforming the closest baseline, O-TPT (5.06). Full per-dataset tables for these shifts are available in Appendix L, and an extended analysis on additional domain shifts (e.g., PACS, DomainNet) is provided in Appendix M.

**Biomedical Domain Robustness.** Finally, on specialized biomedical benchmarks utilizing BioMedCLIP Zhang et al. (2024) (Fig. 6), UC-TPT achieves the lowest average ECE (45.10) compared to A-TPT (48.30), O-TPT (58.00), C-TPT (58.40), and standard TPT (59.50). This substantial absolute reduction in calibration error confirms the broad applicability and safety of our uncertainty-calibrated framework, even in high-risk, specialized medical domains where overconfidence can be particularly detrimental. Additional analysis using the RemoteCLIP Liu et al. (2024) backbone across 7 remote sensing datasets is presented in the appendix R.

**Accuracy vs. ECE trade-off.** Calibration objectives fundamentally affect accuracy; as established in prior literature Kumar et al. (2018); Karandikar et al. (2021); Yoon et al. (2023), a small decrease in accuracy is an

expected mathematical consequence of improving calibration. Uniform confidence suppression shrinks margins and weakens beneficial sharpening on easy samples. Standard TPT methods aggressively update all samples, forcing marginally higher accuracy. However, because unlabeled adaptation relies on self-training signals (like entropy or pseudo-labels) that become unreliable under severe domain shift, this causes negative prompt drift—resulting in the model being up to 98% confident even when completely wrong. UC-TPT mitigates this via *selective adaptation* and *sharpening*—up-weighting reliable instances while actively down-weighting uncertain or OOD ones—alongside topology-aware regularization that constrains drift and preserves zero-shot structure. Ultimately, trading a fraction of a percent in accuracy to eliminate critically overconfident errors is a necessary trade-off for real-world reliability.

## 5 Ablation Analysis

We analyze the contribution of each component of UC-TPT and its individual impact on optimizing calibration (ECE) and accuracy. All ablation experiments are averaged over nine datasets from Table 1, excluding ImageNet, using the CLIP ViT-B/16 backbone unless otherwise specified.

**(i) Contribution of Each Uncertainty.** Table 4 analyzes the effect of individual uncertainty terms i.e. $\mathbf{U}_{\text{align}}$, $\mathbf{U}_{\text{vis-var}}$, and $\mathbf{U}_{\text{jac}}$. Using only $\mathbf{U}_{\text{align}}$ yields moderate calibration, while adding $\mathbf{U}_{\text{vis-var}}$ improves both metrics by stabilizing predictions under perturbations. Further including $\mathbf{U}_{\text{jac}}$, achieves the best trade-off, confirming that the uncertainties capture complementary cues, semantic consistency, prediction stability, and prompt sensitivity, respectively.

**Table 4: Ablation on uncertainty components.** Green ticks (✓) and red crosses (✗) indicate the inclusion or exclusion of each term.

| $\mathbf{U}_{\text{align}}$ | $\mathbf{U}_{\text{vis-var}}$ | $\mathbf{U}_{\text{jac}}$ | **Avg. Acc. / ECE** |
|---|---|---|---|
| ✓ | ✗ | ✗ | 63.04 / 4.10 |
| ✗ | ✓ | ✗ | 63.25 / 4.35 |
| ✗ | ✗ | ✓ | 63.01 / 4.08 |
| ✓ | ✓ | ✗ | 63.19 / 3.96 |
| ✓ | ✗ | ✓ | 63.08 / 3.86 |
| ✗ | ✓ | ✓ | 63.16 / 3.91 |
| ✓ | ✓ | ✓ | **63.51 / 3.62** |

**(ii) Impact of Model Components.** We ablate the effects of the topology-preserving regularizer $\mathcal{L}_{\text{div}}$, visual conditioning $\mathbf{q}_{\mathbf{x}}$, and confidence-shaped target $\phi(\cdot)$. As shown in Fig. 7, incorporating these modules consistently improves both accuracy and calibration. Visual conditioning contributes the largest accuracy gain, while $\mathcal{L}_{\text{div}}$ most effectively lowers ECE by respecting the semantic topology of the frozen model.

**(iii) Addressing Under- and Over-Confidence.** Fig. 8 compares the reliability of A-TPT, O-TPT, and UC-TPT. The diagonal line represents perfect calibration. A-TPT exhibits underconfidence in the low-confidence region and overconfidence in the mid-confidence region. O-TPT remains largely overconfident, particularly around the mid-confidence levels. In contrast, UC-TPT closely follows the diagonal, demonstrating balanced confidence and improved reliability, with notably reduced overconfidence.

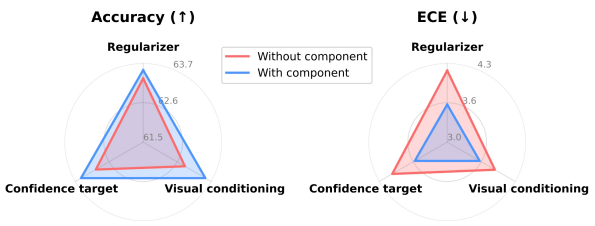



**Figure 7: Component-wise ablations** illustrating their impact.

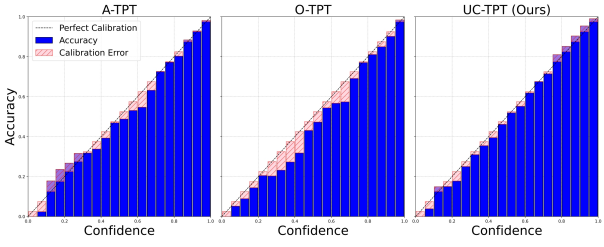



**Figure 8: Reliability plots**, averaged over datasets, showing better calibration with UC-TPT.

**(iv) Isolating Uncertainty Modeling vs. Loss Shape.** To explicitly decouple the effects of the uncertainty-gated weight $\boldsymbol{\alpha}(\mathbf{U}_{\text{norm}})$ and the confidence-shaped target formulation $(\mathbf{H}(p) - \phi)^2$, we provide a controlled ablation study across six distinct datasets (DTD, Pets, Caltech, UCF, Flowers, Aircraft). Modifying both the weight and the loss shape simultaneously could be perceived as lacking a strictly controlled variable setting; this ablation isolates their individual contributions. To maintain a perfectly controlled environment,

our topology-weighted prompt diversity regularizer ($\mathcal{L}_{\text{div}}$) is kept active across all four settings below, ensuring that the *only* variable changing is the formulation of the entropy objective ($\mathcal{L}_{\text{ent}}$).

As shown in Table 5, applying the modified loss shape alone (Shape-Only) causes catastrophic representation collapse, with average accuracy plummeting to 43.76%. Conversely, applying the uncertainty modeling to the traditional entropy loss (Uncertainty-Only) rescues the representations and drives a massive reduction in ECE compared to the base loss. This empirically proves that the dynamic uncertainty gate $\boldsymbol{\alpha}(\mathbf{U}_{\text{norm}})$ acts as the crucial gradient brake and is the true driver of safety and calibration in our framework.

**Table 5:** Ablation decoupling the loss shape from the uncertainty modeling. Averaged across 6 datasets (DTD, Pets, Caltech, UCF, Flowers, Aircraft). The diversity regularizer $\mathcal{L}_{\text{div}}$ is active in all settings to ensure a strictly controlled environment.

| **Setting** | **Entropy Objective ($\mathcal{L}_{\text{ent}}$)** | **Avg. Acc (↑)** | **Avg. ECE (↓)** |
|---|---|---|---|
| Traditional Loss (Base) | $\mathbf{H}(p)$ | 64.88 | 8.30 |
| Shape-Only | $(\mathbf{H}(p) - \boldsymbol{\phi}(\hat{p}_{\max}))^2$ | 43.76 | 6.18 |
| Uncertainty-Only | $\boldsymbol{\alpha}(\mathbf{U}_{\text{norm}}) \cdot \mathbf{H}(p)$ | 63.19 | 3.58 |
| Full UC-TPT | $\boldsymbol{\alpha}(\mathbf{U}_{\text{norm}}) \cdot (\mathbf{H}(p) - \boldsymbol{\phi}(\hat{p}_{\max}))^2$ | **63.60** | **3.28** |

**(v) Preventing Catastrophic Overconfidence.** While aggregate ECE improvements demonstrate overall calibration gains, the true danger of domain shift lies in *confidently wrong* predictions. To evaluate whether our framework successfully prevents logit inflation on these dangerous samples, we isolate top-1 incorrect predictions with high initial confidence ($\hat{p}_{\max} > 0.8$).

As shown in Table 6, traditional baselines like O-TPT act as a destructive cycle, worsening up to 69.6% of these errors by blindly sharpening them. UC-TPT completely breaks this cycle. By correctly identifying their local fragility, the uncertainty gate drops to near-zero on EuroSAT, allowing the model to soften 100% of these severe errors and dropping the subset ECE by a massive 27.16 points compared to O-TPT. This empirically validates that our uncertainty gating successfully acts as a targeted gradient brake exactly when the model is vulnerable.

**Table 6: Diagnostic on confidently wrong samples.** Evaluated on top-1 incorrect predictions with pre-adaptation confidence $\hat{p}_{\max} > 0.8$. Standard entropy minimization baselines (O-TPT, A-TPT) blindly over-sharpen these incorrect predictions, exacerbating miscalibration. In contrast, UC-TPT dynamically detects predictive fragility via the gating coefficient $\boldsymbol{\alpha}(\mathbf{U}_{\text{norm}})$, suppressing harmful updates and drastically improving subset calibration.

| **Dataset** | **Method** | **Pre-Adapt Conf.** | **Post-Adapt Conf. (↓)** | **Mean Gate $\boldsymbol{\alpha}(\mathbf{U}_{\text{norm}})$** | **% Sharpened (Worsened ↓)** | **% Softened (Improved ↑)** | **Subset ECE (↓)** |
|---|---|---|---|---|---|---|---|
| ImageNet-A | O-TPT | 0.89 | 0.90 | 1.000 | 65.30% | 34.70% | 89.70 |
| | A-TPT | 0.89 | 0.89 | 1.000 | 54.80% | 45.20% | 88.90 |
| | **UC-TPT (Ours)** | **0.89** | **0.85** | **0.018** | **25.60%** | **74.40%** | **84.80** |
| EuroSAT | O-TPT | 0.87 | 0.84 | 1.000 | 69.60% | 30.40% | 83.96 |
| | A-TPT | 0.87 | 0.88 | 1.000 | 53.30% | 46.70% | 87.70 |
| | **UC-TPT (Ours)** | **0.87** | **0.57** | **0.013** | **0.00%** | **100.00%** | **56.80** |

## 6 Takeaways

We introduced UC-TPT, a test-time adaptation framework that improves VLM reliability under distribution shifts. UC-TPT fuses visual-to-textual conditioning with a unified uncertainty metric combining alignment confidence, prediction variance, and prompt sensitivity to enable selective, sample-aware updates that sharpen reliable predictions while moderating uncertain ones. A topology-weighted prompt diversity regularizer preserves CLIP's semantic geometry, ensuring stable and interpretable adaptation. Across benchmarks, UC-TPT delivers consistently better calibration and robustness without sacrificing accuracy, positioning it as a strong foundation for reliable continual and open-vocabulary multimodal adaptation.

**Broader Impact Statement**

This paper presents work whose goal is to advance the field of Machine Learning. The proposed methods aim to improve the robustness and calibration of test-time adaptation in vision–language models. While more reliable adaptive inference can support safer deployment of machine learning systems under distribution shift, the techniques introduced in this work do not introduce new ethical risks beyond those commonly associated with pretrained models. We do not foresee immediate negative societal consequences specific to this work.

## References


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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VhlSBZebEw.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863. URL https://doi.org/10.1016/j.dib.2019.104863.

Pablo Morales Alvarez, Stergios Christodoulidis, Maria Vakalopoulou, Pablo Piantanida, and Jose Dolz. Bayesadapter: Enhanced uncertainty estimation in clip few-shot adaptation. *arXiv preprint arXiv:2412.09718*, 2024. doi: 10.48550/arXiv.2412.09718.

Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011.

Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, and Stephen M. Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019. URL https://arxiv.org/abs/1912.12142.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *CVPR*, 2023.

Pingjun Chen. Knee osteoarthritis severity grading dataset. https://data.mendeley.com/datasets/56rmx5bjcr/1, 2018. Accessed: 2025-10-30.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018. doi: 10.1109/ISBI.2018.8363547.

Fatemeh Daneshfar, Abdulhady Abas Abdullah, Moloud Abdar, and Pietro Liò. Ump-net: Uncertainty-aware mixture of prompts network for efficient instruction tuning. *Transactions on Machine Learning Research*, October 2025. URL https://openreview.net/forum?id=EehtvgNXAl. Reviewed on OpenReview.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Stefan Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning (ICML)*, 2017.

Franciskus Xaverius Erick, Mina Rezaei, Johanna Paula Müller, and Bernhard Kainz. Uncertainty-aware vision transformers for medical image analysis. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE 2024)*, volume 15167 of *Lecture Notes in Computer Science*, pp. 171–180, Cham, 2024. Springer. doi: 10.1007/978-3-031-73444-7_17.

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. *Computer Vision and Pattern Recognition Workshop*, 2004.

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*, pp. 2704–2714, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1050–1059. JMLR.org, 2016. URL `http://proceedings.mlr.press/v48/gal16.html`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

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.

Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. *arXiv preprint arXiv:2505.16441*, 2025.

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

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, 2021b.

Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ctradiography. *Scientific Reports*, 12(1):1–14, 2022.

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 *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Ameya Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Rebecca Roelofs. Soft calibration objectives for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29768–29779, 2021.

Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6(1):27988, 2016. doi: 10.1038/srep27988. URL `https://www.nature.com/articles/srep27988`.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. doi: 10.1016/j.cell.2018.02.010. URL https://www.sciencedirect.com/science/article/pii/S0092867418301545.

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*, pp. 19113–19122, 2023.

Thomas Köhler, Attila Budai, Martin Kraus, Jan Odstrcilik, Georg Michelson, and Joachim Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *2013 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 95–100. IEEE, 2013. doi: 10.1109/CBMS.2013.6627771. URL https://doi.org/10.1109/CBMS.2013.6627771.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9w3iw8wDuE.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551, 2017.

Jinlong Li, Zequn Jie, Elisa Ricci, Lin Ma, and Nicu Sebe. Enhancing robustness of vision-language models through orthogonality learning and cross-regularization. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2407.08374.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3390838. URL https://doi.org/10.1109/TGRS.2024.3390838.

Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *The Twelfth International Conference on Learning Representations*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013a.

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

Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *European Conference on Computer Vision*, pp. 147–165. Springer, 2024.

Masoud Nickparvar. Brain tumor mri dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset, 2021. Accessed: 2025-10-30.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. *Advances in Neural Information Processing Systems*, 37:12677–12707, 2024.

Yonatan Ovadia, Elena Fertig, Cynthia Ren, Zachary Nado, D Sculley, Joshua Dillon, Balaji Lakshminarayanan, Jasper Snoek, Dejan Almeida, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Erik Pautsch, John Li, Silvio Rizzi, George K. Thiruvathukal, and Maria Pantoja. Optimized uncertainty estimation for vision transformers: Enhancing adversarial robustness and performance using selective classification. In *Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis (SC-W '23)*, pp. 391–394, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3624062.36241.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Paal Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, New York, NY, USA, 2017. ACM.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169: 337–350, 2020.

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, 2021a. URL `https://arxiv.org/abs/2103.00020`.

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 *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021b.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021c.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19942–19951, 2025.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29958–29967, June 2025.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

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. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

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.

Anas M. Tahir, Muhammad E. H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021. doi: 10.1016/j.compbiomed.2021.105002. URL `https://doi.org/10.1016/j.compbiomed.2021.105002`.

Mingkui Tan, Guohao Chen, Jiaxiang Wu, et al. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2024.

Francesco Tonolini, Nikolaos Aletras, Jordan Massiah, and Gabriella Kazai. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In *Findings of the Association for Computational Linguistics (ACL 2024)*, pp. 12229–12272. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.728. URL `https://aclanthology.org/2024.findings-acl.728/`.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, pp. 180161, 2018. doi: 10.1038/sdata.2018.161. URL `https://pubmed.ncbi.nlm.nih.gov/30106392/`.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *International Conference on Learning Representations*, 2020.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned clip. In *International Conference on Machine Learning (ICML)*, 2024.

Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Understanding and mitigating miscalibration in prompt tuning for vision-language models. In *International Conference on Machine Learning (ICML)*, 2025.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Zehao Xiao and Cees GM Snoek. Beyond model adaptation at test time: A survey. *arXiv preprint arXiv:2411.03687*, 2024.

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=jzzEHTBFOT`.

Sangkyu Yoon, Eunho Kim, Juho Koo, Taeyeong Lee, and Sang-Wook Kim. ESD: Expected squared difference as a tuning-free trainable calibration measure. In *The Eleventh International Conference on Learning Representations*, 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. doi: 10.1056/AIoa2400640. URL `https://ai.nejm.org/doi/full/10.1056/AIoa2400640`.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018.

# A Appendix

In this Appendix, we provide the following components to support and extend the main paper:

1. Details of Visual-to-Textual Injection Module and design choice (Sec. B).
2. Full details of the datasets used for evaluation of UC-TPT and comparative methods (Sec. C).
3. Full implementation and experimental protocols of UC-TPT, Details of baseline methods and their implementation details for reproducibility (Sec. D).
4. Formal definitions, derivations, and complete results for all calibration metrics (Sec. E).
5. Details on the uncertainty normalization strategy (Sec. F).
6. Details on the use of Hutchinson's estimator for Jacobian-based uncertainty (Sec. G).
7. Analysis and breakdown of the computational overhead introduced by UC-TPT (Sec. H).
8. Theoretical insight of safe sharpening as inverse-variance weighting (Sec. I).
9. Interpretive analysis on Topology-Aware Prompt Geometry regularizer (Sec. J).
10. Analysis on selection of key hyperparameters used in our study (Sec. K).
11. Additional and full detailed results across datasets (Sec. L).
12. Extended evaluation on natural distribution shift datasets (Sec. M).
13. Test-time prompt tuning results using different prompt-learning backbones (Sec. N).
14. Dataset-level reliability and robustness analysis and comparison with baseline methods on challenging benchmarks (Sec. O).
15. Evaluation of different combinations of regularizers of baselines along with our framework (Sec. P).
16. Comparison of different prompt initialization strategies with existing methods (Sec. Q).
17. Analysis with RemoteCLIP on remote sensing datasets (Sec. R).
18. Batch size and adaptation step ablations (Sec. S).

# B Details of Visual-to-Textual Injection Module

This appendix provides implementation details for the visual-to-textual conditioning module introduced in Section 3.1.

**Definition B.1** (Visual Injection Mapper)**.** The visual-to-textual conditioning function $\text{Inject}(\cdot;\boldsymbol{\theta}_{inj})$ is implemented as a lightweight two-layer multilayer perceptron (MLP) that maps first-layer visual features to prompt-level modulation coefficients. Specifically,

$$\text{Inject}(\mathbf{z}^{(1)}) = \mathbf{W}_2\,\sigma(\mathbf{W}_1\mathbf{z}^{(1)}), \tag{22}$$

where $\mathbf{z}^{(1)} \in \mathbb{R}^d$ is the first-layer visual embedding, $\mathbf{W}_1 \in \mathbb{R}^{w\times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{m\times w}$ are learnable parameters, $\sigma(\cdot)$ denotes the GELU nonlinearity, $w$ is the text transformer width, and $m$ is the number of prompt tokens.

The output of $\text{Inject}(\cdot)$ is interpreted as a token-wise modulation vector $\mathbf{q}_\mathbf{x} \in \mathbb{R}^m$, which is broadcast across the corresponding prompt embeddings and injected into the first layer of the text encoder as defined in Eq. 7. This design introduces no additional depth, attention, or parameter updates beyond the prompt tokens and preserves the frozen CLIP backbone.

*Remark* B.2 (Shallow Injection Design Choice). We restrict visual-to-text conditioning to the first layer of the text encoder. This design is motivated by stability and calibration considerations rather than expressive capacity. Empirically, deeper-layer injection or richer coupling mechanisms (e.g., cross-attention) were observed to introduce larger confidence distortions and increased expected calibration error (ECE), without yielding consistent improvements in accuracy. Early-layer injection allows image-conditioned prompt modulation while largely preserving the pretrained semantic structure of CLIP, aligning with the goal of label-free and calibration-aware test-time adaptation.

## C Datasets

We conduct a comprehensive empirical evaluation of UC-TPT across a broad and diverse collection of benchmarks, encompassing generic image classification, fine-grained recognition, robustness to distribution shifts, remote sensing, and medical image understanding. This evaluation protocol is designed to rigorously assess the generalization capabilities, robustness, and adaptability of the model across heterogeneous visual domains, semantic granularities, and imaging modalities.

**Generic image classification benchmarks.** We evaluate standard object recognition performance on ImageNet Deng et al. (2009), a large-scale dataset comprising over one million images across 1,000 object categories, and Caltech101 Fei-Fei et al. (2004), which consists of 101 object categories characterized by substantial intra-class variability. These datasets serve as canonical benchmarks for assessing general-purpose visual recognition performance.

**Fine-grained & Cross-domain recognition benchmarks.** Fine-grained and cross-domain generalization is evaluated on a diverse set of datasets designed to capture subtle inter-class variations across multiple visual attributes. DTD Cimpoi et al. (2014) evaluates texture recognition under varying material properties and illumination conditions. FLW Nilsback & Zisserman (2008) focuses on fine-grained flower classification, where class distinctions are often defined by subtle visual cues. Food101 Bossard et al. (2014) assesses recognition of food categories exhibiting significant appearance diversity. UCF101 Soomro et al. (2012) is employed to evaluate action recognition from still frames extracted from video sequences. We further include StanfordCars Maji et al. (2013a) and FGVC-Aircraft Maji et al. (2013b) to assess fine-grained vehicle recognition. OxfordPets Parkhi et al. (2012) evaluates fine-grained animal recognition across cat and dog breeds. Finally, EuroSAT Helber et al. (2019) is included in this suite to evaluate cross-domain adaptation to multi-spectral satellite imagery using standard vision-language backbones.

**Robustness under distribution shifts.** To evaluate robustness to distributional shifts and naturally occurring perturbations, we conduct experiments on several ImageNet variants. ImageNet-A Hendrycks et al. (2021b) contains naturally adversarial examples that pose significant challenges to standard image classifiers. ImageNet-V2 and ImageNet-R Hendrycks et al. (2021a) introduce distribution shifts through changes in image sources and artistic renditions, respectively. ImageNet-S Wang et al. (2019) evaluates robustness with an emphasis on shape-based recognition through segmentation-derived subsets.

**Remote sensing benchmarks.** For a deeper analysis into specialized aerial and Earth observation imagery, we evaluate our method using the domain-specific RemoteCLIP backbone Liu et al. (2024) across seven remote sensing datasets. This includes high-resolution aerial scene classification benchmarks such as MLRSNet Qi et al. (2020), PatternNet Zhou et al. (2018), and RESISC45 Cheng et al. (2017). We also include AID Xia et al. (2017) and the UC Merced Land Use Dataset (UCM) Yang & Newsam (2010) for complex land-use classification, alongside RSICD Lu et al. (2017) to capture rich spatial and semantic earth-observation structures. Finally, we re-evaluate EuroSAT Helber et al. (2019) within this dedicated remote sensing suite. These datasets present unique challenges, such as arbitrary visual orientations, varying spatial resolutions, and extreme domain shifts.

**Medical image classification benchmarks.** For medical image understanding, we adopt Biomed-CLIP Zhang et al. (2024) and evaluate performance across 11 datasets spanning 10 anatomical regions and 9 imaging modalities, thereby covering a wide range of clinically relevant imaging scenarios. Specifically, computed tomography (CT) imaging is evaluated using CTKidney Islam et al. (2022). Dermatoscopic skin lesion classification is assessed using DermaMNIST Codella et al. (2018); Tschandl et al. (2018). Endoscopic

image classification is evaluated on the Kvasir dataset Pogorelov et al. (2017), while retinal disease classification is assessed using fundus images from RETINA Köhler et al. (2013); Porwal et al. (2018). Histopathological image classification is evaluated using LC25000 Borkowski et al. (2019) and CHMNIST Kather et al. (2016), which capture cellular- and tissue-level variations. Magnetic resonance imaging (MRI)–based brain tumor classification is assessed using BTMRI Nickparvar (2021). Optical coherence tomography (OCT)–based retinal imaging is evaluated using OCTMNIST Kermany et al. (2018). Ultrasound-based breast lesion classification is assessed using BUSI Al-Dhabyani et al. (2020). Finally, X-ray imaging is evaluated using COVID-QU-Ex Tahir et al. (2021) and KneeXray Chen (2018), covering thoracic and musculoskeletal imaging tasks, respectively.

## D Implementation details and comparative methods

This section describes the implementation details, optimization settings, and comparative evaluation protocols used in our experiments, with the goal of ensuring clarity, fairness, and reproducibility.

We adopt CLIP models with both ResNet-50 (CLIP-RN50) and Vision Transformer (CLIP ViT-B/16) backbones as the foundational architectures. Test-time prompt tuning (TPT) Shu et al. (2022) serves as the primary baseline framework, upon which all comparative methods are implemented. Additional experiments involving prompt-learning–based initializations, such as CoOp Zhou et al. (2022) and MaPLe Khattak et al. (2023).

All baseline calibration methods—including C-TPT Yoon et al. (2024), O-TPT Sharifdeen et al. (2025), and A-TPT Ahamed et al. (2026)—are implemented on top of the TPT framework, following the configurations and design choices reported in their respective publications. For prompt optimization, all methods employ a single-step update using the AdamW optimizer Loshchilov & Hutter (2019) with a learning rate of $5 \times 10^{-3}$ and a batch size of 64, comprising the original image and 63 augmented images. Following prior TPT-based methods, a batch size of 64 is used solely for stable entropy minimization during optimization; predictions remain label-free and instance-wise. Prompt embeddings are initialized as hard prompts (e.g., `"a photo of a <class>"`) for all methods, consistent with the setup in C-TPT and related works.

During optimization in UC-TPT, the adapted context embeddings are treated as lightweight, learnable parameters within the uncertainty estimation process, enabling stable gradient flow while keeping the CLIP backbone fully frozen. This design choice ensures computational efficiency while preserving the pretrained representation capacity of CLIP.

Regularization hyperparameters are set in accordance with the original implementations of the respective methods. Specifically, C-TPT fixes the regularization weight to $\lambda = 50$ across all experiments. O-TPT sets $\lambda = 18$ for standard experimental settings and reduces it to $\lambda = 2$ when evaluating natural domain shift scenarios. A-TPT employs $\lambda = 80$ for standard experiments and $\lambda = 10$ for natural domain shift evaluations. For our method, we follow A-TPT and set $\lambda$ in Eq. 21 to 80 in all standard experiments. The $K$ in Eq. 10 is set to 5, The $k_0$ and $u_0$ in Eq. 17 is set to 10 and 0.01 respectively. For computing the visual-variance uncertainty $\mathbf{U}_{\text{vis-var}}(\mathbf{x})$ (Eq. 12), we inject isotropic Gaussian noise with fixed standard deviation $\sigma = 0.01$ into $\ell_2$-normalized visual embeddings. This choice induces small local perturbations on the feature manifold and is kept constant across all datasets and methods. The stability constant $\nu$ in Eq.15 is set to 0.1. The parameter $\tilde{C}$ in section 3.2, which determines the subset of classes contributing to the softmax density mass, is chosen in a way that more than 99.9% cumulative probability mass for each dataset, as this proportion consistently captures the majority of the probability mass.

For comparative evaluation, we benchmark our approach against CLIP Radford et al. (2021b), TPT Shu et al. (2022), C-TPT Yoon et al. (2024), O-TPT Sharifdeen et al. (2025), A-TPT Ahamed et al. (2026), and R-TPT Sheng et al. (2025). To ensure fairness and reproducibility, all methods are re-implemented and evaluated under reported experimental settings.

Across all experiments, we report classification accuracy and Expected Calibration Error (ECE) as the primary evaluation metrics. All experiments are conducted on a single NVIDIA RTX Pro 6000 GPU.

## E Calibration Metrics

**Notation.** Let $n$ be the total number of samples, $C$ the number of classes, and $B$ the number of confidence bins. $\mathbf{S}_b$ denotes the set of samples falling into bin $b$, and $\mathbf{S}_b^{(c)}$ the samples of class $c$ in bin $b$. $n_c = |\{i : y_i = c\}|$ is the number of samples of class $c$. $\mathrm{acc}(\cdot)$ and $\mathrm{conf}(\cdot)$ denote empirical accuracy and mean confidence in a set, respectively. $p_{i,c}$ is the predicted probability for class $c$ on sample $i$, and $p_{i,y_i}$ the probability assigned to the true class. $\mathbb{1}[\cdot]$ is the indicator function.

**Expected Calibration Error (ECE).** ECE measures the discrepancy between predicted confidence and empirical accuracy over fixed bins:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathbf{S}_b|}{n} \left|\mathrm{acc}(\mathbf{S}_b) - \mathrm{conf}(\mathbf{S}_b)\right|. \tag{23}$$

**Static Calibration Error (SCE).** We use a simplified, class-agnostic SCE computed over $B$ fixed (equal-width) confidence bins. Let $\mathbf{S}_b$ be the set of samples falling in bin $b$, $\mathrm{acc}(\mathbf{S}_b)$ the empirical accuracy in that bin, and $\mathrm{conf}(\mathbf{S}_b)$ the average confidence in the bin. Then

$$\mathrm{SCE} \;=\; \frac{1}{B} \sum_{b=1}^{B} \left|\mathrm{acc}(\mathbf{S}_b) - \mathrm{conf}(\mathbf{S}_b)\right|. \tag{24}$$

**Maximum Calibration Error (MCE).** MCE reports the worst-case calibration gap across bins:

$$\mathrm{MCE} = \max_{b \in \{1,\ldots,B\}} \left|\mathrm{acc}(\mathbf{S}_b) - \mathrm{conf}(\mathbf{S}_b)\right|. \tag{25}$$

**Adaptive Calibration Error (ACE).** ACE uses adaptive bins with equal sample counts ($|\mathbf{S}_b| = n/B$):

$$\mathrm{ACE} = \sum_{b=1}^{B} \frac{|\mathbf{S}_b|}{n} \left|\mathrm{acc}(\mathbf{S}_b) - \mathrm{conf}(\mathbf{S}_b)\right|. \tag{26}$$

**Brier Score.** A mean squared error between predicted probabilities and one-hot labels:

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C} \big(p_{i,c} - \mathbb{1}[y_i = c]\big)^2. \tag{27}$$

**Negative Log-Likelihood (NLL).** A likelihood-based calibration measure penalizing confident wrong predictions:

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log p_{i,y_i}. \tag{28}$$

We presented the results of all the calibration metrics across datasets in Table 7. Metrics such as ECE, SCE, MCE, and ACE are reported in percentage form, whereas Brier Score and NLL are presented in probabilistic values. We can clearly see that UC-TPT (Ours) consistently performs better across all metrics compared to Sharifdeen et al. (2025); Ahamed et al. (2026); Yoon et al. (2024).

## F Composite Uncertainty Aggregation and Normalization

Recall that UC-TPT combines three complementary per-sample uncertainty signals—image–text alignment, perturbation variance, and Jacobian-based prompt sensitivity—into a unified scalar score:

$$\mathbf{U}(\mathbf{x}) = \mathbf{U}_{\text{align}}(\mathbf{x}) + \mathbf{U}_{\text{vis-var}}(\mathbf{x}) + \mathbf{U}_{\text{jac}}(\mathbf{x}), \tag{29}$$

**Table 7: Calibration metrics across datasets** for OTPT, CTPT, ATPT, and our method on CLIP ViT-B/16. Lower is better (↓).

| Method | Metric | DTD | FLW | Food | Aircraft | Pets | Caltech | UCF | EuroSAT | Cars | INet | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O-TPT | **ECE** | 8.08 | 3.87 | 4.63 | 3.97 | 1.96 | 4.64 | 2.28 | 13.80 | 1.61 | 1.98 | 4.682 |
| | **SCE** | 8.26 | 5.45 | 5.88 | 5.58 | 5.22 | 18.95 | 2.68 | 12.34 | 2.56 | 2.98 | 6.990 |
| | **MCE** | 19.48 | 13.43 | 11.13 | 13.23 | 25.13 | 85.92 | 11.01 | 29.17 | 13.00 | 9.18 | 23.068 |
| | **ACE** | 7.86 | 4.48 | 4.85 | 4.46 | 2.00 | 4.73 | 2.32 | 13.80 | 1.90 | 2.38 | 4.878 |
| | **Brier** | 0.703 | 0.413 | 0.260 | 0.862 | 0.174 | 0.114 | 0.480 | 0.760 | 0.460 | 0.450 | 0.468 |
| | **NLL** | 2.20 | 1.65 | 0.69 | 3.00 | 0.40 | 0.25 | 1.33 | 1.85 | 1.03 | 1.23 | 1.363 |
| C-TPT | **ECE** | 12.45 | 5.13 | 3.72 | 4.33 | 1.83 | 4.34 | 2.40 | 13.25 | 1.56 | 3.19 | 5.220 |
| | **SCE** | 11.67 | 7.13 | 4.62 | 5.41 | 5.19 | 18.55 | 3.38 | 10.75 | 2.73 | 3.51 | 7.294 |
| | **MCE** | 25.45 | 17.26 | 8.86 | 12.17 | 24.72 | 82.18 | 13.16 | 28.79 | 13.49 | 6.36 | 23.244 |
| | **ACE** | 12.40 | 5.33 | 3.90 | 4.66 | 1.73 | 4.48 | 2.12 | 13.26 | 1.71 | 3.16 | 5.275 |
| | **Brier** | 0.720 | 0.440 | 0.250 | 0.860 | 0.173 | 0.110 | 0.460 | 0.750 | 0.450 | 0.440 | 0.465 |
| | **NLL** | 2.26 | 1.68 | 0.68 | 3.01 | 0.41 | 0.26 | 1.28 | 1.83 | 1.01 | 1.23 | 1.365 |
| A-TPT | **ECE** | 6.15 | 4.19 | 7.08 | 6.32 | 5.85 | 10.1 | 3.63 | 4.46 | 5.52 | 2.26 | 5.550 |
| | **SCE** | 9.13 | 6.53 | 4.65 | 7.46 | 5.65 | 22.59 | 2.75 | 12.92 | 2.25 | 3.44 | 7.737 |
| | **MCE** | 32.31 | 16.18 | 9.38 | 18.11 | 24.70 | 86.93 | 7.44 | 30.61 | 7.80 | 8.17 | 24.163 |
| | **ACE** | 9.04 | 4.51 | 3.67 | 7.54 | 1.89 | 5.55 | 2.00 | 17.51 | 1.76 | 2.46 | 5.593 |
| | **Brier** | 0.710 | 0.410 | 0.250 | 0.870 | 0.178 | 0.120 | 0.480 | 0.800 | 0.460 | 0.450 | 0.473 |
| | **NLL** | 2.23 | 1.67 | 0.69 | 3.05 | 0.43 | 0.26 | 1.34 | 1.91 | 1.03 | 1.24 | 1.385 |
| **UC-TPT (Ours)** | **ECE** | 4.74 | 3.97 | 3.23 | 2.65 | 2.65 | 2.96 | 2.70 | 8.36 | 1.35 | 1.53 | 3.414 |
| | **SCE** | 4.65 | 5.60 | 4.88 | 6.31 | 5.52 | 17.49 | 3.36 | 7.32 | 2.40 | 2.25 | 5.978 |
| | **MCE** | 13.81 | 18.35 | 8.94 | 13.92 | 18.71 | 85.62 | 11.17 | 23.16 | 12.38 | 8.15 | 21.421 |
| | **ACE** | 5.18 | 4.74 | 3.61 | 4.24 | 2.66 | 3.30 | 2.55 | 9.25 | 1.72 | 1.78 | 3.903 |
| | **Brier** | 0.690 | 0.420 | 0.240 | 0.860 | 0.170 | 0.110 | 0.470 | 0.750 | 0.450 | 0.450 | 0.461 |
| | **NLL** | 2.19 | 1.61 | 0.67 | 3.00 | 0.38 | 0.23 | 1.32 | 1.76 | 1.03 | 1.24 | 1.343 |

where $\mathbf{U}_{\text{align}}(\mathbf{x})$, $\mathbf{U}_{\text{vis-var}}(\mathbf{x})$, and $\mathbf{U}_{\text{jac}}(\mathbf{x})$ are defined in Eqs. (9)–(13) of the main paper. These terms have heterogeneous ranges and units: $\mathbf{U}_{\text{align}}$ is derived from cosine similarities, $\mathbf{U}_{\text{vis-var}}$ from prediction variance under perturbations, and $\mathbf{U}_{\text{jac}}$ from Jacobian norms. Directly feeding $\mathbf{U}(\mathbf{x})$ into the gating function $\alpha(\cdot)$ in Eq. 17 would make the update strength sensitive to arbitrary scale differences between these components and to dataset- or batch-specific magnitudes.

**Justification for additive fusion.** We intentionally adopt a direct, unweighted additive fusion for Eq. 29 rather than multiplicative or weighted alternatives. Conceptually, because $\mathbf{U}_{\text{align}}$, $\mathbf{U}_{\text{vis-var}}$, and $\mathbf{U}_{\text{jac}}$ track independent failure modes, direct summation acts as a continuous logical "OR" gate. For instance, if a sample exhibits a massive semantic gap but has perfectly stable visual features, the total uncertainty must still spike to protect the model from a harmful update. A multiplicative approach would dangerously suppress this warning if any single term approached zero.

To empirically validate this design, we compared our unweighted additive formulation against two intuitive alternatives:

- **Weighted Fusion:** Prioritizes specific risks by assigning uneven weights (e.g., emphasizing semantic risk via $\mathbf{U}_{\text{w}} = 0.6\mathbf{U}_{\text{align}} + 0.2\mathbf{U}_{\text{vis-var}} + 0.2\mathbf{U}_{\text{jac}}$).
- **Rank-Based Aggregation:** Sums intra-batch rankings to eliminate sensitivity to extreme outliers: $\mathbf{U}_{\text{rank}} = \text{Rank}(\mathbf{U}_{\text{align}}) + \text{Rank}(\mathbf{U}_{\text{vis-var}}) + \text{Rank}(\mathbf{U}_{\text{jac}})$.

As shown in Table 8, simple unweighted addition remains the most robust validation-free approach. Rank aggregation loses critical absolute magnitude information (which dictates the actual severity of the risk), while weighted fusion requires dataset-specific tuning and generally yields inferior average calibration across varied domains.

To obtain a robust and comparable scale across samples after additive fusion, we normalize the composite uncertainty within each mini-batch $B$ using a batch-wise min–max transformation:

$$\mathbf{U}_{\text{norm}}(\mathbf{x}) = \frac{\mathbf{U}(\mathbf{x}) - \min_{\mathbf{x}' \in B} \mathbf{U}(\mathbf{x}')}{\max_{\mathbf{x}' \in B} \mathbf{U}(\mathbf{x}') - \min_{\mathbf{x}' \in B} \mathbf{U}(\mathbf{x}') + \epsilon}, \quad \mathbf{x} \in B, \tag{30}$$

where $\epsilon > 0$ is a small constant for numerical stability. This ensures that $\mathbf{U}_{\text{norm}}(\mathbf{x}) \in [0, 1]$ for all samples in the batch.

**Why batch-wise normalization?** Our entropy objective uses $\mathbf{U}_{\text{norm}}(\mathbf{x})$ only through the bounded gating function $\alpha(\mathbf{U}_{\text{norm}}(\mathbf{x}))$ in Eq. 17,

$$\alpha\big(\mathbf{U}_{\text{norm}}(\mathbf{x})\big) = \frac{1}{1 + \exp\big(k_0(\mathbf{U}_{\text{norm}}(\mathbf{x}) - u_0)\big)}, \tag{31}$$

and the adaptive loss

$$L_{\text{ent}}(\mathbf{x}) = \alpha\big(\mathbf{U}_{\text{norm}}(\mathbf{x})\big)\big(\mathbf{H}(p(\mathbf{x})) - \phi(\hat{p}_{\max}(\mathbf{x}))\big)^2. \tag{32}$$

In this formulation, only the *relative* ordering of uncertainties within a batch matters: samples with larger $\mathbf{U}(\mathbf{x})$ should receive smaller updates (smaller $\alpha$), while more reliable samples with lower $\mathbf{U}(\mathbf{x})$ should be adapted more aggressively (larger $\alpha$). The min–max normalization in Eq. (30) enforces a consistent dynamic range $[0, 1]$ for $\mathbf{U}_{\text{norm}}(\mathbf{x})$ across batches, making the slope parameter $k_0$ and midpoint $u_0$ interpretable and stable across datasets and backbones. The typical behaviour of the gating function $\alpha(u)$ over the normalized range $u \in [0, 1]$ is illustrated in Fig. 9, showing a smooth monotonic decay that naturally suppresses updates for high-uncertainty samples.

**Table 8: Ablation on uncertainty aggregation strategies.** Average accuracy and ECE across benchmarks using the ViT-B/16 backbone.

| Aggregation Strategy | Avg. Acc (↑) | Avg. ECE (↓) |
|---|---|---|
| Weighted (0.6 : 0.2 : 0.2) | 63.42 | 3.68 |
| Weighted (0.2 : 0.6 : 0.2) | **63.55** | 3.71 |
| Rank-Based Aggregation | 63.52 | 3.64 |
| **Additive (Ours: 1:1:1)** | 63.51 | **3.62** |

Empirically, we observe that although $\mathbf{U}_{\text{align}}(\mathbf{x})$, $\mathbf{U}_{\text{vis-var}}(\mathbf{x})$, and $\mathbf{U}_{\text{jac}}(\mathbf{x})$ can differ in absolute scale, their *relative* ordering within a batch is informative and fairly consistent across domains. Batch-wise min–max normalization therefore acts as a monotone reparameterization that preserves this ordering while removing arbitrary scale and offset effects.

In fact, on the EuroSAT dataset, we find that the raw combined uncertainty $\mathbf{U}(\mathbf{x})$ has a very large absolute magnitude (e.g., batch-level mean 242.109 and standard deviation 3.28), making its scale highly dataset-dependent and unsuitable for direct gating. After applying batch-wise min–max normalization, the corresponding $\mathbf{U}_{\text{norm}}(\mathbf{x})$ values lie in a stable and interpretable range with mean 0.38 and standard deviation 0.202. Since $\alpha(\cdot)$ is monotonic, this normalization preserves the adaptation ordering while ensuring numerical stability and allowing $(k_0, u_0)$ to behave consistently across datasets.

**Comparison to alternative normalizations.** Beyond alternative aggregation strategies, we also considered other normalization strategies, such as: (i) z-score normalization of $\mathbf{U}(\mathbf{x})$ over the batch and (ii) independent normalization of each component $\mathbf{U}_{\text{align}}$, $\mathbf{U}_{\text{vis-var}}$, $\mathbf{U}_{\text{jac}}$ before aggregation. In practice, these alternatives either produced unbounded values (requiring additional clipping) or made the gating overly sensitive to outliers in one component. By contrast, the min–max normalization in Eq. (30) guarantees $\mathbf{U}_{\text{norm}}(\mathbf{x}) \in [0, 1]$, keeps the entropy updates bounded, and yields consistently lower ECE across datasets in our experiments.

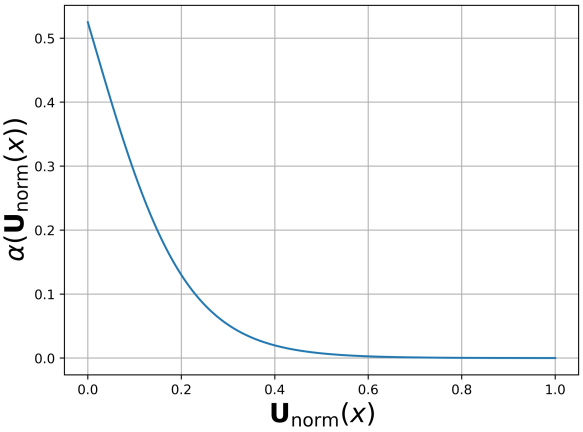



**Figure 9:** Typical behavior of the sigmoid gate $\alpha(\mathbf{U})$ used in our adaptive entropy objective, shown for $k_0 = 10$ and $u_0 = 0.01$ over the normalized uncertainty range $\mathbf{U} \in [0, 1]$.

Finally, we emphasize that the heteroscedastic interpretation of $\mathcal{L}_{\text{ent}}$ (Sec. 3.3.1) is primarily conceptual: $\mathbf{U}_{\text{norm}}(\mathbf{x})$ serves as a surrogate "difficulty" indicator that modulates the effective update strength, analogous to an inverse-variance weight in a het-

eroscedastic regression objective. Our empirical ablations support that this uncertainty-aware reweighting leads to significantly improved calibration compared to uniform entropy minimization.

## G Hutchinson Approximation for Jacobian Sensitivity

In UC-TPT, the Jacobian-based uncertainty $\mathbf{U}_{\text{jac}}$ (Eq. 13) captures the sensitivity of the predicted class probabilities $p_c(\mathbf{x})$ to perturbations of the prompt-token matrix $\mathbf{P} \in \mathbb{R}^{m \times d}$. Computing the full Jacobian norm $\left\| \frac{\partial p_c(\mathbf{x})}{\partial \mathbf{P}} \right\|_2^2$ for all classes requires $O(C)$ backpropagations and becomes computationally prohibitive at test time. To avoid this, we adopt a *Hutchinson trace estimator*, providing an unbiased stochastic estimate of the Jacobian norm using a single vector–Jacobian product.

**Full Jacobian Form.** Let the flattened prompt parameter matrix be $\tilde{\mathbf{p}} = \text{vec}(\mathbf{P}) \in \mathbb{R}^{md}$. For any class $c$, its sensitivity is

$$S_c(\mathbf{x}) = \left\| \frac{\partial p_c(\mathbf{x})}{\partial \tilde{\mathbf{p}}} \right\|_2^2 = \text{Tr}\big(J_c(\mathbf{x})^\top J_c(\mathbf{x})\big)\,, \tag{33}$$

where

$$J_c(\mathbf{x}) = \frac{\partial p_c(\mathbf{x})}{\partial \tilde{\mathbf{p}}}$$

is the class-wise Jacobian with respect to the prompt matrix $\mathbf{P}$.

**Hutchinson Trace Estimator.** For any PSD matrix $A$, the Hutchinson identity is:

$$\text{Tr}(A) = \mathbb{E}_{\mathbf{r}}\big[\mathbf{r}^\top A \mathbf{r}\big], \qquad \mathbf{r} \sim \mathcal{N}(\mathbf{0}, I) \text{ or Rademacher.} \tag{34}$$

Applying Eq. 34 to $A = J_c^\top J_c$ yields the stochastic estimate:

$$S_c(\mathbf{x}) \approx \mathbf{r}^\top \big(J_c(\mathbf{x})^\top J_c(\mathbf{x})\big)\mathbf{r} = \|J_c(\mathbf{x})\,\mathbf{r}\|_2^2\,. \tag{35}$$

Here, $J_c(\mathbf{x})\,\mathbf{r}$ is evaluated via a single *vector–Jacobian product*, computed with one backward pass in modern autodiff frameworks.

**Top–$\widetilde{C}$ Class Restriction.** Following Sec. 3.2, the Jacobian term is computed only for the $\widetilde{C}$ most probable classes:

$$\mathbf{U}_{\text{jac}}(\mathbf{x}) = \sum_{c \in \text{Top-}\widetilde{C}} \big(1 - \cos(z, v_c(\mathbf{x}))\big)\, \|J_c(\mathbf{x})\,\mathbf{r}\|_2^2. \tag{36}$$

The cosine-weighted factor amplifies gradients along semantically misaligned directions, ensuring that $\mathbf{U}_{\text{jac}}(\mathbf{x})$ reflects both representational and semantic fragility.

**Why Top-$\tilde{C}$ Selection Does Not Affect ECE.** In UC-TPT, the Jacobian-based uncertainty $\mathbf{U}_{\text{jac}}(\mathbf{x})$ serves solely as a normalization factor inside the entropy-based adaptation loss and therefore does not alter the softmax probabilities used for prediction. Since ECE depends exclusively on the predicted label $\hat{y}$ and its corresponding confidence $\text{conf}(\mathbf{x})$ (Eq. 23), restricting the Jacobian computation to the top-$\tilde{C}$ classes cannot influence bin assignments or confidence calibration.

In practice, we select $\tilde{C}$ such that more than 99.9% of the predictive class mass is retained. This choice is motivated by empirical evidence across all datasets showing that the cumulative softmax probability of a small subset of high-confidence classes consistently exceeds 0.99. As a result, the remaining low-probability classes contribute negligibly to gradient-based sensitivity, while restricting computation to the top-$\tilde{C}$ classes reduces the Jacobian cost by over 5× without affecting predictions or ECE.

We further empirically validate this design on the EuroSAT dataset. When computing $\mathbf{U}_{\text{jac}}(\mathbf{x})$ using all classes (full Jacobian), the batch-wise min–max normalized uncertainty has a mean of 0.38 and a standard deviation of 0.202. Applying the Hutchinson approximation together with the top-$\tilde{C}$ class restriction yields a closely

**Table 9: Computational complexity, time, and memory usage of TPT variants** on ImageNet-V2 (ViT-B/16). $C$: total classes, $\tilde{C} \ll C$: selected classes capturing more than 99% predictive mass, $C_B$: active classes in batch, $N$: batch size, $d$: embedding dim, $K$: Hutchinson perturbations. Memory denotes peak GPU memory per batch.

| Method | Complexity | Time (s)/batch | Memory (MiB) | ECE (↓) |
|---|---|---|---|---|
| A-TPT Ahamed et al. (2026) | $\mathcal{O}(C^2 d)$ | 0.82 | 21840 | 8.11 |
| O-TPT Sharifdeen et al. (2025) | $\mathcal{O}(C^2 d)$ | 0.90 | 23740 | 4.01 |
| UC-TPT (Full Jacobian) | $\mathcal{O}(Nd) + \mathcal{O}(KNC) + \mathcal{O}(NdC) + \mathcal{O}(C_B^2 d)$ | 1.55 | 25410 | **3.06** |
| UC-TPT (Hutchinson) | $\mathcal{O}(Nd) + \mathcal{O}(KNC) + \mathcal{O}(C_B^2 d)$ | **1.19** | **22650** | **3.06** |

matching distribution (mean 0.39, std. 0.211). This close agreement indicates that the Hutchinson–top-$\tilde{C}$ estimate preserves the relative ordering of uncertain samples within a batch—precisely the quantity utilized by the monotonic gating function $\alpha(\cdot)$—while avoiding the prohibitive cost of full Jacobian computation. Consequently, restricting to the selected top-$\tilde{C}$ classes maintains calibration behavior and leaves ECE unaffected.

## H Complexity Analysis

We analyze the computational complexity of the proposed Uncertainty-Calibrated Test-Time Prompt Tuning (UC-TPT) framework. Let $N$ denote the batch size, $C$ the total number of classes, $C_B \ll C$ the number of classes active within a batch, $d$ the embedding dimension, and $K$ the number of perturbations or Hutchinson samples.

**Lemma H.1** (Visual Encoding Complexity)**.** *Computing CLIP image embeddings for a batch of $N$ samples incurs a computational cost of $\mathcal{O}(Nd)$.*

**Lemma H.2** (Perturbation-Based Uncertainty Complexity)**.** *Estimating uncertainty via $K$ perturbation-consistent logits over $N$ samples and $C$ classes incurs a cost of $\mathcal{O}(KNC)$.*

**Lemma H.3** (Full Jacobian Trace Complexity)**.** *Explicit computation of the Jacobian trace of a $C$-dimensional logit vector with respect to $d$ prompt parameters over $N$ samples incurs a cost of $\mathcal{O}(NdC)$.*

**Lemma H.4** (Hutchinson Trace Approximation)**.** *Using a Hutchinson estimator with $K$ samples approximates the Jacobian trace with computational cost $\mathcal{O}(KNC)$, eliminating the $\mathcal{O}(NdC)$ term.*

**Lemma H.5** (Batch-Restricted Class Interaction)**.** *Restricting class–class regularization to batch-active classes reduces the complexity from $\mathcal{O}(C^2 d)$ to $\mathcal{O}(C_B^2 d)$.*

**Theorem H.6** (Overall UC-TPT Complexity)**.** *The total computational complexity of UC-TPT is given by:*

- ***UC-TPT (Full Jacobian):***

$$\mathcal{C}_{\text{UC-TPT(FJ)}} = \Theta(NdC + KNC + C_B^2 d + Nd),$$

  *which is dominated by the $\Theta(NdC)$ term when $C$ is large.*

- ***UC-TPT (Hutchinson):***

$$\mathcal{C}_{\text{UC-TPT(H)}} = \Theta(KNC + C_B^2 d + Nd),$$

  *which is dominated by $\Theta(KNC)$ under the practical regime $C_B \ll C$ and small $K$.*

*Proof.* The result follows by summing the costs established in Lemmas H.1–H.5. The Hutchinson estimator removes the explicit Jacobian term while preserving an unbiased trace estimate, yielding the stated reduction in asymptotic complexity. □

**Empirical Validation.** Table 9 reports theoretical complexity alongside empirical runtime and memory usage measured on ImageNet-V2 with a ViT-B/16 backbone. Consistent with Theorem H.6, A-TPT and O-TPT incur a global $\mathcal{O}(C^2d)$ overhead due to full class–class interactions, leading to increased latency and memory consumption. In contrast, UC-TPT restricts regularization to batch-active classes and replaces the Jacobian trace with a Hutchinson approximation, substantially reducing both runtime and memory overhead. As a result, UC-TPT introduces only a modest latency increase while achieving significantly improved calibration, reducing ECE from 4.01/8.11 to 3.06.

## I Theoretical insight (safe sharpening as inverse-variance weighting)

We formalize Eq. 18 as a heteroscedastic regression problem and show that (i) the optimal weight is inverse variance, and (ii) the resulting gradient update is provably attenuated for high-uncertainty samples, yielding safer test-time sharpening.

**Setup.** For each test sample $\mathbf{x}$, define the *sharpening residual*

$$r(\mathbf{x};\mathbf{P}) \;=\; \mathbf{H}(p(\mathbf{x};\mathbf{P})) \;-\; \boldsymbol{\phi}(\hat{p}_{\max}(\mathbf{x};\mathbf{P})), \tag{37}$$

where $\mathbf{H}(\cdot)$ is predictive entropy (Eq. 15) and $\boldsymbol{\phi}(\cdot)$ is the confidence-shaped target (Eq. 16). We interpret $r(\mathbf{x};\mathbf{P})$ as a noisy observation of a latent "desired residual" 0 with sample-dependent noise:

$$r(\mathbf{x};\mathbf{P}) \;=\; \epsilon(\mathbf{x}), \qquad \mathbb{E}[\epsilon(\mathbf{x})] = 0, \qquad \mathrm{Var}[\epsilon(\mathbf{x})] = \sigma^2(\mathbf{x}). \tag{38}$$

Under distribution shift, $\sigma^2(\mathbf{x})$ is larger for ambiguous/OOD samples; UC-TPT uses uncertainty $\mathbf{U}(\mathbf{x})$ as a monotone proxy for $\sigma^2(\mathbf{x})$.

**Proposition 1 (ML derivation of inverse-variance weighting).** Assume $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2(\mathbf{x}))$ independent across samples. The negative log-likelihood over a batch $\mathcal{B}$ is

$$\mathcal{L}_{\mathrm{NLL}} = \sum_{\mathbf{x}\in\mathcal{B}} \Big( \frac{r(\mathbf{x};\mathbf{P})^2}{2\sigma^2(\mathbf{x})} + \frac{1}{2}\log\sigma^2(\mathbf{x}) \Big) \;+\; \text{const.} \tag{39}$$

If $\sigma^2(\mathbf{x})$ is treated as fixed w.r.t. $\mathbf{P}$ during prompt updates (or the log-term is absorbed into a constant), minimizing $\mathcal{L}_{\mathrm{NLL}}$ is equivalent to minimizing the weighted least-squares objective

$$\sum_{\mathbf{x}\in\mathcal{B}} \alpha(\mathbf{x})\, r(\mathbf{x};\mathbf{P})^2, \qquad \text{with} \quad \alpha(\mathbf{x}) \propto \frac{1}{\sigma^2(\mathbf{x})}. \tag{40}$$

Thus the statistically optimal weight is inverse variance. In UC-TPT we instantiate $\alpha(\mathbf{x}) = \boldsymbol{\alpha}(\mathbf{U}_{\mathrm{norm}}(\mathbf{x}))$ (Eq. 17), enforcing a monotone mapping from uncertainty to (approx.) inverse variance.

**Proposition 2 (Gauss–Markov / minimum-variance aggregation).** Consider estimating a shared parameter update direction from noisy per-sample residuals with heteroscedastic noise as in Eq. 38. Among all unbiased linear combinations of residual-based gradients, inverse-variance weighting minimizes the estimator variance (classical weighted least squares / Gauss–Markov). Concretely, for scalars $g_i$ with $\mathbb{E}[g_i] = g$ and $\mathrm{Var}(g_i) = \sigma_i^2$, the minimum-variance unbiased estimator of $g$ is

$$\hat{g} = \frac{\sum_i \sigma_i^{-2} g_i}{\sum_i \sigma_i^{-2}}. \tag{41}$$

Hence down-weighting high-uncertainty samples is not heuristic: it is the variance-optimal way to aggregate noisy test-time signals.

**Proposition 3 (safe sharpening via gradient attenuation).** Let the UC-TPT entropy loss for one sample be

$$\mathcal{L}_{\text{ent}}(\mathbf{x}) = \alpha(\mathbf{x})\, r(\mathbf{x};\mathbf{P})^2. \tag{42}$$

Its gradient w.r.t. prompt parameters $\mathbf{P}$ is

$$\nabla_{\mathbf{P}}\mathcal{L}_{\text{ent}}(\mathbf{x}) = 2\,\alpha(\mathbf{x})\, r(\mathbf{x};\mathbf{P})\, \nabla_{\mathbf{P}} r(\mathbf{x};\mathbf{P}) \;+\; r(\mathbf{x};\mathbf{P})^2 \nabla_{\mathbf{P}}\alpha(\mathbf{x}), \tag{43}$$

where the second term is typically small in practice when $\alpha$ depends on $\mathbf{U}$ through weakly varying proxies or is stop-gradient (either choice is valid). Ignoring $\nabla_{\mathbf{P}}\alpha$ (or treating it as bounded), we obtain the key attenuation property:

$$\big\|\nabla_{\mathbf{P}}\mathcal{L}_{\text{ent}}(\mathbf{x})\big\| \;\le\; 2\,\alpha(\mathbf{x})\,|r(\mathbf{x};\mathbf{P})|\,\big\|\nabla_{\mathbf{P}} r(\mathbf{x};\mathbf{P})\big\|. \tag{44}$$

Since $\alpha(\mathbf{x})$ is monotone decreasing in uncertainty, high-uncertainty samples provably contribute smaller gradient magnitudes, preventing aggressive sharpening driven by unreliable/OOD inputs.

**Corollary (risk-aware sharpening).** Under the heteroscedastic model, $\alpha(\mathbf{x}) \propto 1/\sigma^2(\mathbf{x})$ is the maximum-likelihood (and minimum-variance) choice, and Eq. 44 shows it yields conservative updates for samples with high uncertainty (large $\sigma^2$). Therefore, compared to uniform entropy minimization ($\alpha \equiv 1$), UC-TPT performs *risk-aware* sharpening: only samples that are simultaneously (i) aligned with the target entropy (small residual) and (ii) stable (large $\alpha$) meaningfully drive adaptation, which empirically reduces confident errors and stabilizes confidence dynamics (Fig. 4(a), Fig. 15(a,b)).

## J Topology-Aware Prompt Geometry: Interpretive Analysis

This section analyzes the geometric structure induced by the proposed topology-aware diversity regularizer in UC-TPT, and its relation to the frozen semantic space of CLIP.

**Frozen CLIP Semantic Geometry.** Let $\mathbf{v}_c^0 \in \mathbb{R}^d$ denote the frozen CLIP text embedding corresponding to class $c$, obtained from the manual prompt `"a photo of a <class>"`. The pretrained class–class cosine similarity is defined as

$$\mathbf{S}^0_{c'c''} = \cos(\mathbf{v}^0_{c'}, \mathbf{v}^0_{c''}), \qquad -1 \le \mathbf{S}^0_{c'c''} \le 1, \tag{45}$$

which characterizes the semantic relations encoded by the frozen CLIP text encoder.

**Instance-Conditioned Prompt Geometry.** For a test sample $\mathbf{x}$, UC-TPT produces instance-conditioned prompt embeddings $\hat{\mathbf{v}}_c(\mathbf{x})$ for each class $c$. The adapted similarity between two classes for a given sample is

$$\hat{\mathbf{S}}_{c'c''}(\mathbf{x}) = \cos\big(\hat{\mathbf{v}}_{c'}(\mathbf{x}), \hat{\mathbf{v}}_{c''}(\mathbf{x})\big). \tag{46}$$

The diversity regularizer assigns topology-aware weights

$$\mathbf{w}_{c'c''} = 1 - \mathbf{S}^0_{c'c''}, \tag{47}$$

so that class pairs that are semantically close under frozen CLIP receive weaker repulsion, while distant pairs receive stronger repulsion.

The resulting per-sample diversity objective is

$$\mathcal{L}_{\text{div}}(\mathbf{x}) = \sum_{c'<c''} \mathbf{w}_{c'c''}\, \hat{\mathbf{S}}_{c'c''}(\mathbf{x}), \tag{48}$$

which biases the adapted prompt geometry toward respecting relative semantic relations encoded in $\mathbf{S}^0$.

**Class-Averaged Prompt Structure.** For analysis, we consider the class-averaged adapted prompt

$$\hat{\mathbf{v}}_c = \frac{1}{|\mathcal{X}_c|} \sum_{\mathbf{x} \in \mathcal{X}_c} \hat{\mathbf{v}}_c(\mathbf{x}), \qquad \|\hat{\mathbf{v}}_c\|_2 = 1, \tag{49}$$

and stack them as

$$\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_C] \in \mathbb{R}^{d \times C}. \tag{50}$$

The corresponding Gram matrix

$$\hat{\mathbf{S}} = \hat{\mathbf{V}}^\top \hat{\mathbf{V}} \tag{51}$$

summarizes the average inter-class geometry induced by adaptation. The deviation $\|\hat{\mathbf{S}} - \mathbf{S}^0\|_F$ serves as a quantitative measure of how much the adapted prompts depart from the frozen CLIP semantic structure.

**Comparison with Other TPT Regularizers.** O-TPT enforces orthogonality among adapted class prompts, while A-TPT promotes uniform angular separation. Both objectives impose geometry that is independent of the pretrained semantic relations encoded in $\mathbf{S}^0$. In contrast, UC-TPT modulates inter-class repulsion according to frozen CLIP similarity, leading to an adapted geometry that more closely reflects the original semantic topology.

**Logit Sensitivity Perspective.** Given a visual feature $\mathbf{z} \in \mathbb{R}^d$, the TPT logit map is

$$g(\mathbf{z}) = \tau\, \mathbf{z}^\top \hat{\mathbf{V}}, \tag{52}$$

where $\tau > 0$ is the temperature. The sensitivity of logits to perturbations in $\mathbf{z}$ is influenced by the spectral norm $\|\hat{\mathbf{V}}\|_2$, which in turn depends on the geometry of the adapted prompts. By discouraging excessive distortion of inter-class relations, the proposed regularizer induces smoother variations in the logit space, aligning with the observed improvements in calibration reported in the empirical results.

## K Analysis on hyper parameters

Hyperparameters were selected using a held-out validation split of the analyzed datasets, In Fig. 10, we present the variation in ECE with respect to different model parameters: $k_0$ and $u_0$ (Eq. 17), $\lambda$ (Eq. 21), and $K$(Eq. 10), shown in Fig. 10(a), Fig. 10(b), Fig. 10(c), and Fig. 10(d), respectively. The plots report the average performance over five datasets—DTD, Flowers, Pets, EuroSAT, and Cars—using the CLIP ViT-B/16 backbone. While ablating one parameter, we kept all the remaining parameters fixed at their optimal values to ensure a fair comparison. From these figures, we observe that the best calibration performance (ECE ↓) is achieved at $k_0 = 10$, $u_0 = 0.01$, $\lambda = 80$, and $K = 5$. We adopt these parameter settings for all our analysis.

We performed an ablation study by replacing the sigmoid-based gating function $\alpha(\cdot)$ (Eq. 17) in our uncertainty-aware update rule with several alternative gating mechanisms. The objective of this analysis was to evaluate the effectiveness and robustness of different gates across six datasets—Pets, DTD, Flowers, Aircraft, Caltech, and UCF—using CLIP ViT-B/16 as the backbone. Specifically, we evaluated the following gating functions:

- **Tanh-based gate** — a smoother and symmetric alternative to sigmoid:

$$\alpha(\mathbf{U}) = 0.5 \left(1 - \tanh\left(k_0\big(\mathbf{U}_{\text{norm}} - u_0\big)\right)\right). \tag{53}$$

- **Inverse-square gate** — a polynomial attenuation mechanism:

$$\alpha(\mathbf{U}) = \frac{1}{1 + \left(k_0(\mathbf{U}_{\text{norm}} - u_0)\right)^2}. \tag{54}$$

- **Exponential decay gate** — a monotonic exponential suppression:

$$\alpha(\mathbf{U}) = e^{-k_0(\mathbf{U}_{\text{norm}} - u_0)}. \tag{55}$$

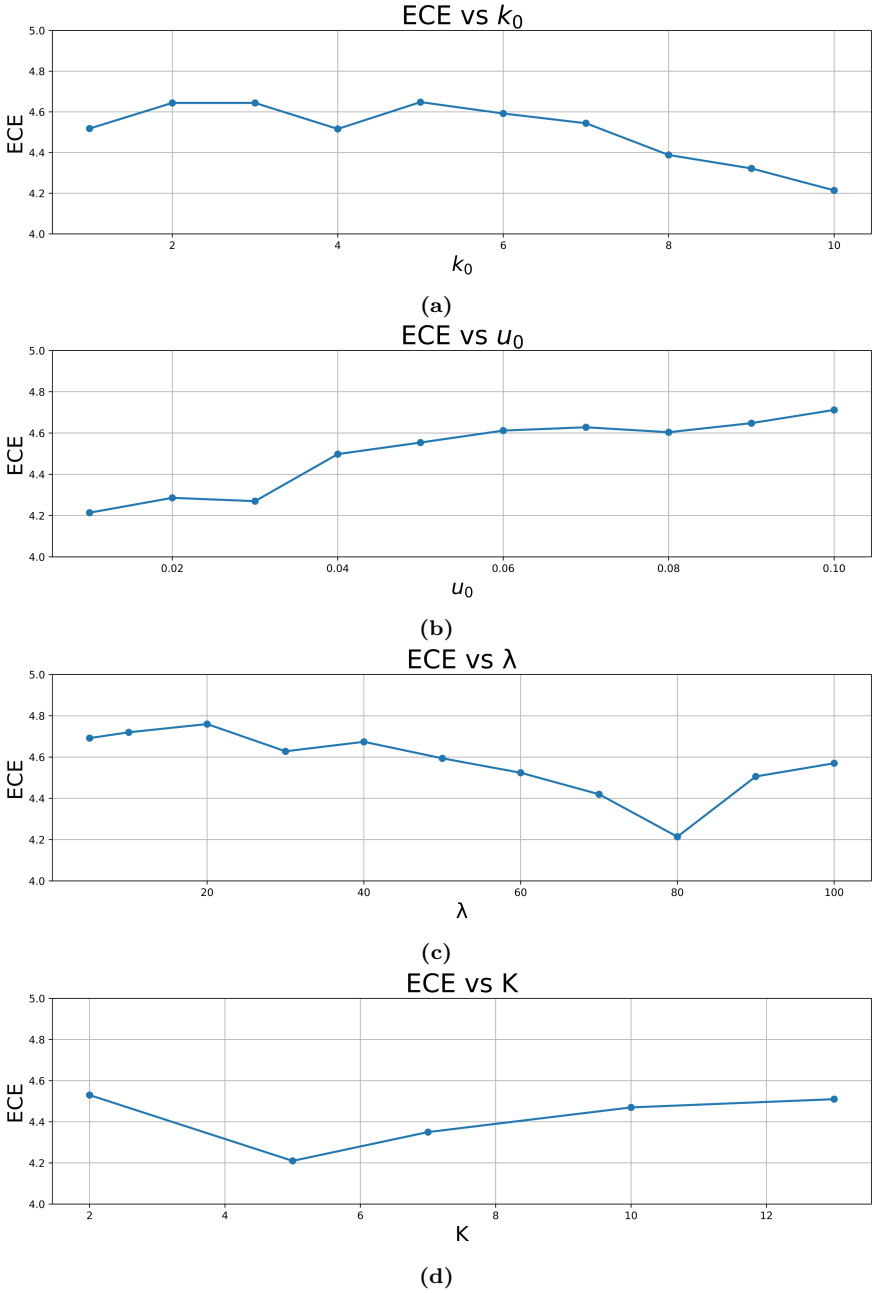



**Figure 10:** Ablations on model parameters with CLIP-ViT B/16 backbone averaged over DTD, Flower, Pets, Eurosat, Stanfordcar datasets (a) ECE vs $k_0$. (b) ECE vs $u_0$. (c) ECE vs $\lambda$. (d) ECE vs $K$.

As illustrated in Fig. 11, the sigmoid-based gate achieves the best calibration performance (ECE↓) on Aircraft, Caltech, and DTD. While it produces slightly higher ECE on Pets, Flowers, and UCF compared to certain alternatives, the **overall trend consistently favors sigmoid gating**. When averaged across all six datasets, the sigmoid gate achieves the **lowest mean ECE of 3.28**, outperforming the Tanh-based gate (3.49), Inverse-square gate (3.52), and Exponential decay gate (3.50).

Given its superior average calibration performance and stable behavior across diverse datasets, we adopt the **sigmoid gating function** as the default choice in our method.

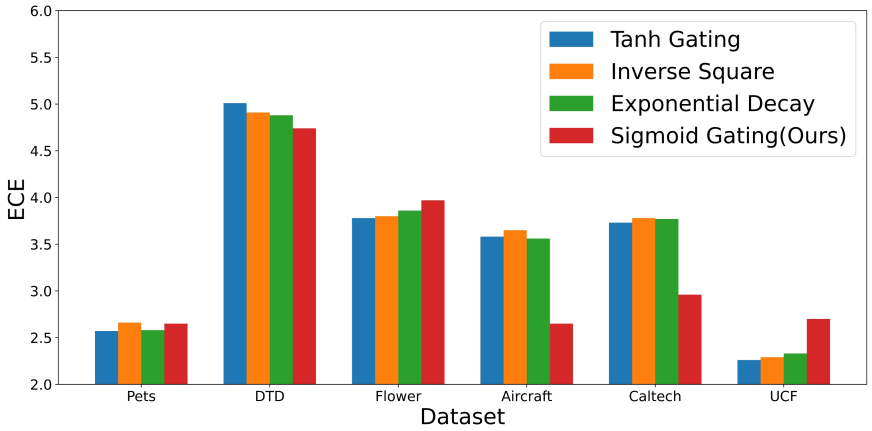



**Figure 11: Comparison of ECE (↓) across datasets** using CLIP-ViT B/16 as the backbone with different gating functions.

## L Detailed experimental analysis

In this section, we provide detailed tables for better clarity. Table 10 presents a comparison of evaluation metrics across all baseline methods on all datasets using the CLIP RN50 backbone. We observe that UC-TPT achieves the best (lowest) average ECE (↓) while maintaining competitive accuracy compared to the baselines.

**Table 10: Comparison of calibration performance** with CLIP-RN50 backbone. Best average result is in **bold**.

| Method | Metric | ImageNet | DTD | Flowers | Food | Aircraft | Pets | Caltech | UCF | EuroSAT | Car | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-Shot | **Acc.** | 58.20 | 39.95 | 60.94 | 73.80 | 15.66 | 83.73 | 85.88 | 58.42 | 24.21 | 55.68 | 55.65 |
| | **ECE** | 2.08 | 9.66 | 3.12 | 2.48 | 6.31 | 6.01 | 4.22 | 3.03 | 14.42 | 4.62 | 5.60 |
| TPT | **Acc.** | 60.66 | 41.49 | 62.40 | 75.00 | 16.98 | 84.79 | 86.65 | 59.53 | 28.53 | 58.38 | 57.44 |
| | **ECE** | 11.41 | 25.72 | 13.59 | 5.24 | 16.15 | 3.55 | 4.80 | 12.35 | 22.46 | 3.60 | 11.88 |
| R-TPT | **Acc.** | 58.20 | 39.83 | 61.02 | 75.76 | 15.63 | 83.65 | 85.80 | 58.37 | 24.20 | 55.68 | 55.81 |
| | **ECE** | 11.48 | 24.83 | 12.36 | 5.20 | 12.37 | 4.04 | 4.94 | 9.64 | 30.93 | 1.33 | 11.71 |
| C-TPT | **Acc.** | 60.00 | 42.20 | 65.08 | 74.67 | 16.86 | 83.67 | 86.73 | 59.56 | 27.57 | 56.14 | 57.25 |
| | **ECE** | 3.02 | 20.75 | 4.04 | 1.86 | 10.99 | 2.67 | 2.14 | 4.05 | 15.19 | 1.95 | 6.66 |
| A-TPT | **Acc.** | 59.82 | 41.84 | 64.51 | 74.19 | 15.81 | 82.99 | 86.20 | 58.57 | 29.87 | 55.85 | 56.96 |
| | **ECE** | 2.15 | 16.16 | 3.25 | 1.78 | 9.18 | 2.7 | 2.99 | 4.53 | 5.22 | 1.96 | 4.99 |
| O-TPT | **Acc.** | 58.76 | 41.55 | 65.65 | 74.54 | 16.77 | 83.21 | 86.65 | 59.00 | 27.80 | 55.29 | 56.92 |
| | **ECE** | 3.19 | 16.70 | 2.64 | 1.13 | 8.42 | 3.15 | 3.35 | 2.21 | 15.72 | 1.88 | 5.84 |
| **UC-TPT (Ours)** | **Acc.** | 58.19 | 41.62 | 65.88 | 74.31 | 16.95 | 82.31 | 87.00 | 58.39 | 28.30 | 55.15 | 56.81 |
| | **ECE** | 2.07 | 11.62 | 3.05 | 1.22 | 7.23 | 3.74 | 2.76 | 2.71 | 11.33 | 1.85 | **4.76** |

The detailed results for natural domain shift on the ImageNet variants are presented in Table 11 for the CLIP-RN50 backbone and Table 12 for the CLIP-ViT B/16 backbone. From both tables, we can observe that UC-TPT, our proposed method, outperforms all other baseline approaches and achieves the best average ECE.

In Table 13, we present the baseline methods—O-TPT Sharifdeen et al. (2025), C-TPT Yoon et al. (2024), R-TPT Sheng et al. (2025), and A-TPT Ahamed et al. (2026)—with and without our uncertainty modeling. The results clearly show that incorporating uncertainty estimation into test-time prompt tuning is effective not only for our method but also for all baseline approaches. A significant reduction in ECE is consistently observed when uncertainty estimation is applied, demonstrating that it is a powerful tool for achieving better-calibrated and more trustworthy models in test-time prompt tuning.

In Table 14, we report the standard deviation across three random seeds for A-TPT Ahamed et al. (2026), O-TPT Sharifdeen et al. (2025), and UC-TPT (Ours) using the CLIP-ViT B/16 backbone on the DTD, Flower, Food, Caltech, and Cars datasets. We observe that all methods exhibit similar standard deviation

**Table 11: Calibration performance** on the ImageNet suite (CLIP-RN50) for natural distribution shifts. Best average result is in **bold**.

| Method | Metric | I-A | I-V2 | I-R | I-S | Avg |
|---|---|---|---|---|---|---|
| CLIP-RN50 | **Acc.** | 21.69 | 51.44 | 55.94 | 33.33 | 40.60 |
| | **ECE** | 21.30 | 3.35 | 1.95 | 3.14 | 7.43 |
| TPT | **Acc.** | 25.17 | 54.58 | 59.10 | 35.26 | 43.53 |
| | **ECE** | 31.04 | 13.18 | 9.12 | 13.67 | 16.75 |
| R-TPT | **Acc.** | 21.64 | 51.51 | 55.95 | 33.30 | 40.60 |
| | **ECE** | 29.98 | 13.54 | 9.79 | 12.99 | 16.57 |
| C-TPT | **Acc.** | 22.20 | 53.37 | 56.87 | 34.34 | 41.69 |
| | **ECE** | 22.78 | 5.09 | 1.34 | 6.46 | 8.92 |
| A-TPT | **Acc.** | 23.74 | 54.50 | 58.42 | 35.09 | 42.94 |
| | **ECE** | 27.87 | 11.08 | 6.20 | 11.74 | 14.22 |
| O-TPT | **Acc.** | 22.57 | 53.15 | 57.11 | 34.07 | 41.72 |
| | **ECE** | 24.25 | 3.82 | 2.58 | 5.06 | 8.93 |
| **UC-TPT (Ours)** | **Acc.** | 24.80 | 52.81 | 55.97 | 33.92 | 41.87 |
| | **ECE** | 18.25 | 3.19 | 1.85 | 4.23 | **6.88** |

**Table 12: Calibration performance** on the ImageNet suite (CLIP-ViT B/16) for natural distribution shifts. Best average result is in **bold**.

| Method | Metric | I-A | I-V2 | I-R | I-S | Avg |
|---|---|---|---|---|---|---|
| CLIP-ViT B/16 | **Acc.** | 47.73 | 60.79 | 74.01 | 46.12 | 57.16 |
| | **ECE** | 8.40 | 2.79 | 3.59 | 4.84 | 4.90 |
| TPT | **Acc.** | 52.85 | 63.12 | 76.91 | 47.93 | 60.20 |
| | **ECE** | 16.37 | 11.14 | 4.36 | 14.94 | 11.70 |
| R-TPT | **Acc.** | 47.75 | 60.75 | 73.98 | 46.18 | 57.16 |
| | **ECE** | 13.95 | 11.52 | 4.58 | 14.02 | 11.02 |
| C-TPT | **Acc.** | 49.37 | 61.99 | 74.82 | 47.29 | 58.37 |
| | **ECE** | 6.45 | 4.51 | 2.87 | 7.23 | 5.26 |
| A-TPT | **Acc.** | 51.02 | 62.46 | 76.20 | 47.68 | 59.34 |
| | **ECE** | 10.23 | 8.11 | 1.91 | 11.58 | 7.96 |
| O-TPT | **Acc.** | 49.94 | 61.69 | 75.26 | 47.08 | 58.49 |
| | **ECE** | 7.14 | 4.01 | 2.15 | 6.93 | 5.06 |
| **UC-TPT (Ours)** | **Acc.** | 47.82 | 60.89 | 73.48 | 46.46 | 57.16 |
| | **ECE** | 7.11 | 3.06 | 2.08 | 6.28 | **4.63** |

**Table 13:** ECE comparison across different Test-Time Prompt Tuning methods with and without uncertainty modeling.

| Method | Metric | DTD | FLW | Food | Air. | Pets | C101 | UCF | SAT | Cars | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| O-TPT | ECE | 8.08 | 3.87 | 4.63 | 3.97 | 1.96 | 4.64 | 2.28 | 13.80 | 1.61 | 4.98 |
| O-TPT+Uncertainty | ECE | 4.74 | 3.66 | 4.79 | 3.45 | 2.41 | 4.22 | 2.36 | 10.55 | 1.48 | **4.18** |
| C-TPT | ECE | 12.45 | 5.13 | 3.72 | 4.33 | 1.83 | 4.34 | 2.40 | 13.25 | 1.56 | 5.45 |
| C-TPT+Uncertainty | ECE | 5.39 | 4.01 | 5.24 | 3.74 | 2.56 | 3.70 | 2.85 | 12.49 | 1.64 | **4.62** |
| R-TPT | ECE | 18.79 | 10.84 | 3.30 | 12.65 | 5.40 | 3.60 | 12.07 | 22.02 | 1.93 | 10.07 |
| R-TPT+Uncertainty | ECE | 3.72 | 5.34 | 4.50 | 2.55 | 5.51 | 7.19 | 3.70 | 9.57 | 10.11 | **5.80** |
| A-TPT | ECE | 8.92 | 3.96 | 3.58 | 6.77 | 2.34 | 5.31 | 1.77 | 17.56 | 1.23 | 5.72 |
| A-TPT+Uncertainty | ECE | 5.08 | 4.34 | 3.57 | 4.77 | 2.48 | 4.59 | 2.51 | 15.03 | 1.43 | **4.87** |

**Table 14: Standard deviation across 3 random seeds** for A-TPT, O-TPT, and UC-TPT (Ours) with the CLIP-ViT B/16 backbone. Lower is better (↓).

| Method | Metric | DTD | Flower | Food | Caltech | Car | Avg |
|---|---|---|---|---|---|---|---|
| A-TPT | **ACC** | 0.3014 | 0.0732 | 0.0516 | 0.1189 | 0.1020 | 0.12942 |
| | **ECE** | 0.2459 | 0.1103 | 0.0941 | 0.1293 | 0.1944 | 0.15480 |
| O-TPT | **ACC** | 0.1514 | 0.1529 | 0.0416 | 0.1189 | 0.1205 | 0.11706 |
| | **ECE** | 0.0983 | 0.1980 | 0.0163 | 0.2170 | 0.1702 | 0.13996 |
| **UC-TPT (Ours)** | **ACC** | 0.1316 | 0.1541 | 0.1225 | 0.1537 | 0.0612 | 0.12462 |
| | **ECE** | 0.0902 | 0.1287 | 0.0300 | 0.2614 | 0.1225 | **0.12656** |

values. Compared to the others, UC-TPT shows a slightly lower average ECE standard deviation, indicating more stable and reproducible results.

## M Domain shift analysis

For further analysis in domain shifts, we choose PACS Li et al. (2017) and DomainNet Peng et al. (2019) datasets.

The PACS Li et al. (2017) dataset is a widely used benchmark for domain-shift and domain-generalization studies. It contains four distinct domains—**Photo, Art Painting, Cartoon, and Sketch**—that exhibit large variations in texture, style, and abstraction. Despite sharing the same set of object categories, the

visual appearance differs drastically across domains, making PACS an effective testbed for evaluating a model's robustness to distribution shifts. Its diverse domain composition helps assess how well a method can generalize from one visual style to another, especially in real-world scenarios where models often encounter unseen or shifted data distributions.

The DomainNet Peng et al. (2019) dataset is one of the largest and most challenging benchmarks for domain-shift and domain-adaptation research. It spans six highly diverse domains—**Clipart, Infograph, Painting, Quickdraw, Real, and Sketch**—covering over 300 object categories. The dataset exhibits significant variations in drawing style, abstraction level, texture, and visual complexity, making cross-domain generalization particularly difficult. Due to its large scale and strong inter-domain discrepancies, DomainNet provides a rigorous testbed for evaluating the robustness, scalability, and adaptability of models under substantial distribution shifts.

The analysis on the PACS dataset using the CLIP-ViT B/16 backbone, comparing O-TPT Sharifdeen et al. (2025) with UC-TPT (ours), is presented in Fig. 12. We observe that UC-TPT consistently outperforms O-TPT across all domain-shift variants of PACS. Similarly, the results on the DomainNet dataset are shown in Fig. 13. Here as well, UC-TPT achieves better performance than O-TPT across all six domain-shift variants, demonstrating its robustness under diverse and challenging distribution shifts.

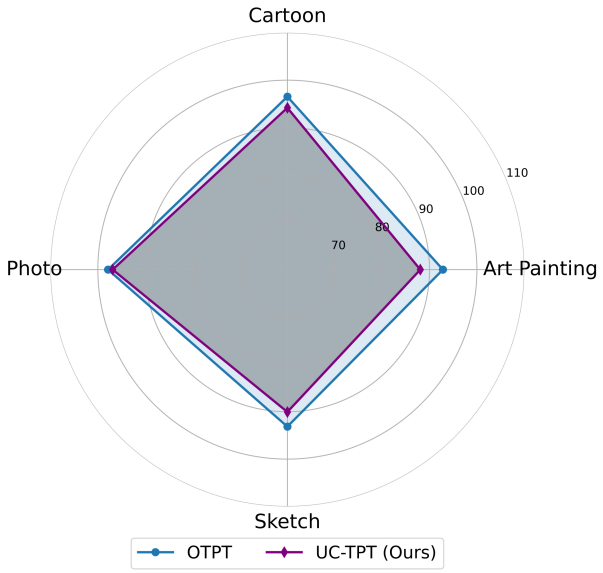



**Figure 12: Comparison of ECE (↓) on 4 Domain shift datasets of PACS** using CLIP-ViT B/16 as the backbone.

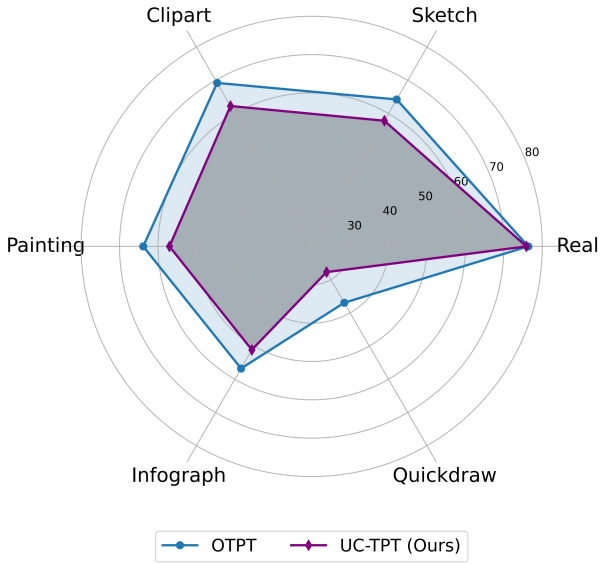



**Figure 13: Comparison of ECE (↓) on 6 Domain shift datasets of DomainNet** using CLIP-ViT B/16 as the backbone.

## N TPT with prompt learning backbone

The analysis of test-time prompt tuning methods—O-TPT Sharifdeen et al. (2025), A-TPT Ahamed et al. (2026), and UC-TPT (ours)—combined with prompt-learning backbones such as CoOp Zhou et al. (2022) and MaPLe Khattak et al. (2023) is presented in Table 15. This evaluation is conducted on DTD, Flower, Food, Aircraft, Pets, Caltech101, UCF, EuroSAT, and Cars datasets using the CLIP-ViT B/16. From the table, we observe that UC-TPT achieves the best performance within the CoOp framework, as reflected by its lower average ECE.

For the MaPLe backbone, UC-TPT attains average ECE performance comparable to O-TPT and clearly outperforms A-TPT, while maintaining competitive accuracy across datasets. This behavior highlights an important insight: methods like MaPLe, which integrate multi-layer text–visual fusion, tend to distort the natural CLIP embedding space, negatively affecting calibration. We observed a similar effect when replacing

the shallow visual conditioning in our method with deeper visual fusion—the ECE values increased, indicating that excessive fusion alters the inherent CLIP semantic structure and consequently degrades calibration performance.

**Table 15:** Comparison of MaPLe and CoOp based prompt learning backbones combined with different TPT variants across datasets.

| Method | Metric | DTD | FLW | Food | Air. | Pets | C101 | UCF | SAT | Cars | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **MaPLe + A-TPT** | Acc | 42.38 | 66.63 | 83.33 | 22.65 | 86.07 | 93.67 | 65.29 | 48.22 | 63.28 | 63.50 |
| | ECE | 11.39 | 3.88 | 1.84 | 6.88 | 3.48 | 3.56 | 3.99 | 2.73 | 2.87 | 4.51 |
| **MaPLe + O-TPT** | Acc | 42.44 | 67.07 | 83.38 | 22.74 | 86.24 | 93.51 | 64.98 | 48.63 | 62.87 | 63.54 |
| | ECE | 11.77 | 3.23 | 2.06 | 6.10 | 3.58 | 3.50 | 3.26 | 2.45 | 3.07 | 4.34 |
| **MaPLe + UC-TPT** | Acc | 42.43 | 67.03 | 83.36 | 22.71 | 86.26 | 93.55 | 64.68 | 49.47 | 62.75 | 63.58 |
| | ECE | 10.31 | 3.40 | 2.11 | 6.04 | 3.59 | 3.67 | 3.17 | 3.81 | 3.05 | 4.35 |
| **CoOp + A-TPT** | Acc | 45.69 | 68.33 | 83.56 | 18.90 | 89.04 | 93.23 | 65.93 | 40.51 | 62.78 | 63.11 |
| | ECE | 16.10 | 9.03 | 3.94 | 20.12 | 1.57 | 1.15 | 10.84 | 12.83 | 2.55 | 8.68 |
| **CoOp + O-TPT** | Acc | 45.10 | 68.37 | 83.55 | 18.66 | 89.02 | 93.83 | 65.66 | 40.41 | 62.53 | 63.01 |
| | ECE | 16.41 | 6.96 | 3.56 | 16.87 | 2.10 | 0.97 | 9.11 | 13.71 | 2.78 | 8.05 |
| **CoOp + UC-TPT** | Acc | 44.86 | 68.58 | 83.47 | 17.70 | 88.96 | 93.63 | 64.63 | 40.56 | 62.18 | 62.73 |
| | ECE | 10.96 | 5.60 | 3.16 | 13.53 | 2.22 | 1.23 | 8.26 | 10.95 | 2.71 | 6.51 |

## O Reliability analysis

Further reliability analysis on selected datasets—Pets, Aircraft, and DTD—using A-TPT Ahamed et al. (2026), O-TPT Sharifdeen et al. (2025), and UC-TPT (ours) with the CLIP-ViT B/16 backbone is presented in Fig. 14. In Fig. 14(a), we show the results on the Pets dataset. Both A-TPT and O-TPT exhibit strong overconfidence in the low-confidence regions, whereas UC-TPT demonstrates noticeably reduced overconfidence.

Fig. 14(b) illustrates the analysis on the Aircraft dataset. A-TPT remains largely overconfident across the confidence spectrum, while O-TPT performs slightly better but becomes underconfident at higher confidence levels. In contrast, UC-TPT shows significantly reduced underconfidence and overconfidence compared to both baselines.

Finally, Fig. 14(c) presents the results for the DTD dataset. Here, both A-TPT and O-TPT display strong overconfidence, whereas UC-TPT offers a more balanced reliability curve with substantially lower miscalibration.

For the analysis of incorrect confidences, we compared UC-TPT (ours) against A-TPT Ahamed et al. (2026) and O-TPT Sharifdeen et al. (2025) using the CLIP-ViT B/16 backbone on the Aircraft and Caltech datasets. These results are presented in Fig. 15.

In Fig. 15(a), we show the results for the Aircraft dataset. As illustrated, UC-TPT exhibits a higher density of incorrect samples in the low-confidence region and a much lower density in the high-confidence region. This behavior is desirable, as it indicates that the model assigns low confidence to its mistakes. In contrast, both A-TPT and O-TPT display the opposite trend: they produce fewer incorrect predictions in the low-confidence region but noticeably more incorrect predictions in the high-confidence region. Such patterns reflect overconfident errors, making those methods less reliable than UC-TPT.

Fig. 15(b) presents the analysis on the Caltech dataset. Even under this setting, UC-TPT shows a more favorable trend, with incorrect sample density peaking in the mid-confidence region and remaining lower in the high-confidence region compared to A-TPT and O-TPT. This again demonstrates that UC-TPT is better calibrated and more trustworthy, as it avoids producing overly confident incorrect predictions.

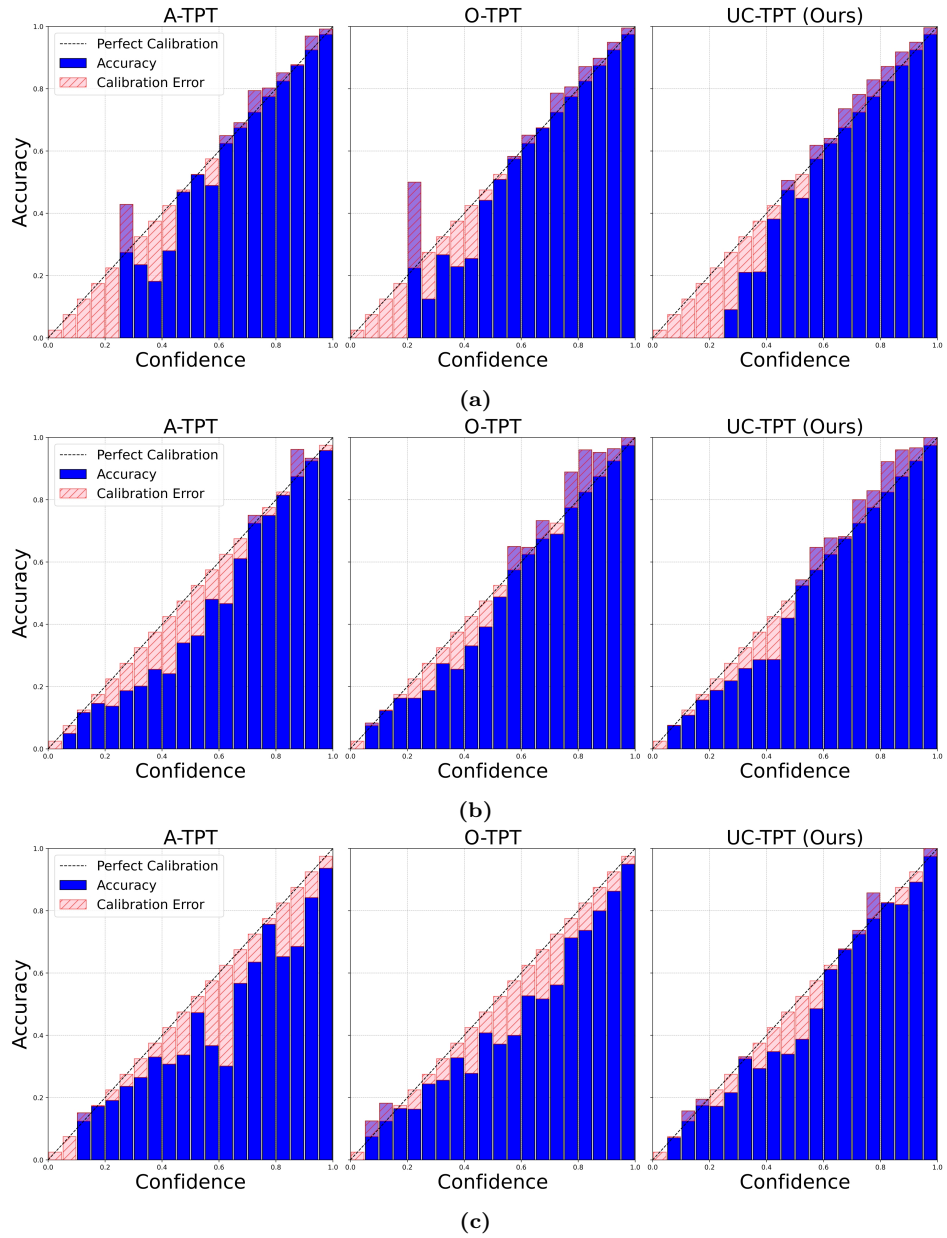
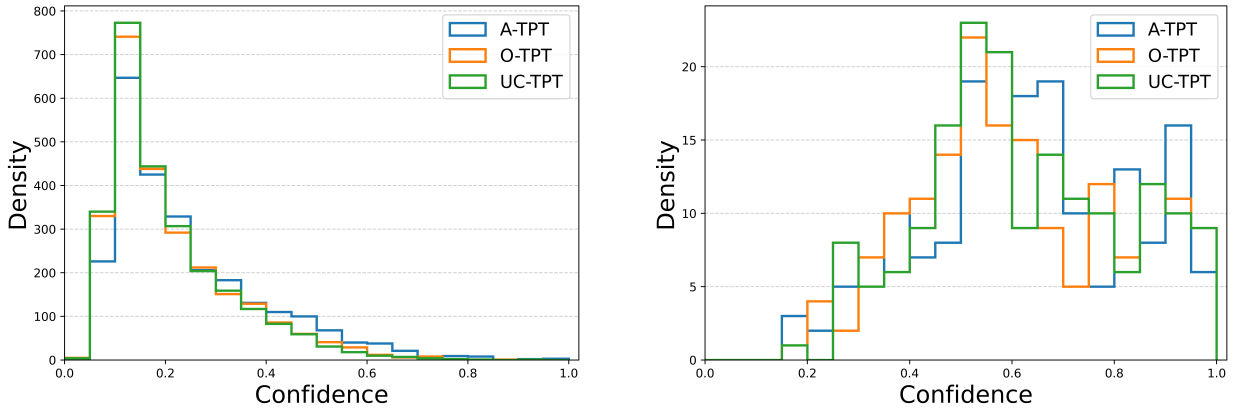



**Figure 14:** Reliability plots with CLIP-ViT B/16 backbone. (a) Pets. (b) Aircraft. (c) DTD.

## P Analysis on combination of regularizers

In this section, we investigate how our UC-TPT behaves when combined with the regularizers used in other test-time prompt tuning methods, namely A-TPT Ahamed et al. (2026), O-TPT Sharifdeen et al. (2025), and C-TPT Yoon et al. (2024). This analysis is conducted using the CLIP-ViT B/16 backbone across six datasets: Pets, DTD, Flower, Aircraft, Caltech, and UCF. We report the average performance across these datasets to understand the overall effect of combining different regularizers. The complete results are presented in Table 16.

From the table, we observe that UC-TPT alone consistently achieves the lowest average ECE, while still maintaining competitive accuracy compared to all other combinations. This highlights the importance




**(a)** **(b)**

**Figure 15:** Incorrect confidence plots with CLIP-ViT B/16 backbone. (a) Aircraft. (b) Caltech.

**Table 16:** Performance comparison when regularizers from other methods are combined with UC-TPT (ours) across multiple datasets using CLIP-ViT-B/16 as the backbone. The overall best results are highlighted in **bold**.

| Method | Metric | Pets | DTD | FLW | Air. | C101 | UCF | Avg |
|---|---|---|---|---|---|---|---|---|
| Ours + O-TPT | **Acc.** | 87.94 | 43.79 | 68.55 | 23.10 | 92.49 | 63.97 | 63.31 |
| | **ECE** | 2.34 | 4.60 | 3.37 | 3.51 | 4.76 | 2.40 | 3.50 |
| Ours + A-TPT | **Acc.** | 87.62 | 43.61 | 67.21 | 23.64 | 92.25 | 63.20 | 62.92 |
| | **ECE** | 2.29 | 5.59 | 4.78 | 5.23 | 5.51 | 2.33 | 4.29 |
| Ours + C-TPT | **Acc.** | 87.84 | 44.08 | 67.83 | 23.25 | 92.53 | 63.50 | 63.17 |
| | **ECE** | 2.46 | 5.01 | 3.77 | 3.67 | 4.87 | 2.19 | 3.66 |
| **UC-TPT (Ours)** | **Acc.** | 88.40 | 44.30 | 67.80 | 24.00 | 93.40 | 63.70 | 63.60 |
| | **ECE** | 2.65 | 4.74 | 3.97 | 2.65 | 2.96 | 2.70 | **3.28** |

of preserving the semantic structure of textual embeddings in the frozen CLIP space—a central property maintained by our topology-preserving diversity regularizer.

The next best-performing combination is UC-TPT + O-TPT, followed by combinations with C-TPT and A-TPT. Although some of these combinations yield improvements on specific datasets (especially Pets and UCF), they fail to generalize across the remaining datasets. This suggests that applying aggressive transformations such as orthogonal projections, angular dispersion, or excessive separation of embeddings may distort the natural semantic relationships between closely related classes. Such distortions often lead to degraded calibration performance.

In contrast, UC-TPT alone maintains the intrinsic CLIP semantics, ensuring that class relationships remain meaningful while still enabling instance-wise uncertainty-aware prompt adaptation. This balanced behavior results in superior overall calibration performance compared to all the combined regularization approaches.

## Q Analysis on different prompt initializations

In this section, we present a detailed analysis of different prompt initialization strategies. The experiments are conducted on the Pets, DTD, Flowers, Aircraft, Caltech, and UCF datasets using both CLIP-ViT-B/16 and RN50 backbones. Table 17 reports the results obtained with random prompt initialization. For both CLIP-ViT-B/16 and RN50, the average ECE achieved by our UC-TPT method is consistently lower compared to O-TPT Sharifdeen et al. (2025) and A-TPT Ahamed et al. (2026). This demonstrates that the uncertainty estimates used in our approach make the model more robust across different prompt initializations, allowing

it to generalize better than the competing methods. In contrast, A-TPT is highly sensitive to prompt initialization, showing a significant increase in ECE, followed by O-TPT in both backbones.

Table 18 summarizes the analysis using the prompt template "the photo of the cool $< class >$". A similar trend is observed here: our method continues to perform reliably across initialization schemes, maintaining low ECE values, while A-TPT again exhibits strong sensitivity to prompt choices, followed by O-TPT. The consistently lower ECE of UC-TPT across both initialization scenarios highlights its robustness and stability in comparison to existing approaches.

**Table 17: Comparison of accuracy and calibration performance** across datasets for the random prompt initialization. Best average performance is in **bold**.

| Method | Metric | Pets | DTD | FLW | Air. | C101 | UCF | Avg |
|---|---|---|---|---|---|---|---|---|
| ViT-B/16 O-TPT | **Acc.** | 80.32 | 34.87 | 59.72 | 22.11 | 86.97 | 55.62 | 56.60 |
| | **ECE** | 4.53 | 23.38 | 6.64 | 8.58 | 4.72 | 10.81 | 9.78 |
| ViT-B/16 A-TPT | **Acc.** | 77.10 | 34.51 | 53.59 | 21.18 | 86.53 | 50.17 | 53.85 |
| | **ECE** | 9.19 | 22.92 | 11.22 | 10.33 | 3.55 | 16.80 | 12.34 |
| ViT-B/16 UC-TPT (Ours) | **Acc.** | 82.72 | 32.68 | 61.59 | 21.54 | 85.43 | 53.89 | 56.31 |
| | **ECE** | 2.29 | 16.40 | 5.65 | 8.29 | 4.63 | 9.73 | **7.83** |
| RN 50 O-TPT | **Acc.** | 68.36 | 26.12 | 52.37 | 10.74 | 80.73 | 49.91 | 48.04 |
| | **ECE** | 10.19 | 10.67 | 7.89 | 16.34 | 6.60 | 6.21 | 9.65 |
| RN 50 A-TPT | **Acc.** | 61.29 | 26.54 | 49.74 | 10.68 | 79.11 | 49.70 | 46.18 |
| | **ECE** | 14.96 | 17.85 | 12.27 | 24.56 | 3.93 | 9.72 | 13.88 |
| RN 50 UC-TPT (Ours) | **Acc.** | 71.35 | 21.51 | 51.81 | 10.44 | 80.32 | 49.38 | 47.47 |
| | **ECE** | 5.53 | 6.87 | 6.91 | 15.34 | 7.67 | 5.53 | **7.98** |

**Table 18: Comparison of accuracy and calibration performance** across datasets with the prompt "the photo of the cool $< class >$". Best average performance is in **bold**.

| Method | Metric | Pets | DTD | FLW | Air. | C101 | UCF | Avg |
|---|---|---|---|---|---|---|---|---|
| ViT-B/16 O-TPT | **Acc.** | 88.55 | 47.17 | 70.36 | 23.88 | 91.44 | 65.71 | 64.51 |
| | **ECE** | 2.31 | 4.05 | 6.22 | 8.17 | 1.71 | 5.66 | 6.34 |
| ViT-B/16 A-TPT | **Acc.** | 85.85 | 46.57 | 63.52 | 23.55 | 91.12 | 66.08 | 62.78 |
| | **ECE** | 2.71 | 14.75 | 10.25 | 14.52 | 2.28 | 8.03 | 8.74 |
| ViT-B/16 UC-TPT (ours) | **Acc.** | 88.12 | 45.98 | 68.62 | 22.95 | 92.25 | 64.63 | 63.76 |
| | **ECE** | 2.63 | 8.71 | 4.88 | 8.12 | 2.21 | 4.45 | **5.15** |
| RN 50 O-TPT | **Acc.** | 82.23 | 39.18 | 67.15 | 16.74 | 86.69 | 59.55 | 58.58 |
| | **ECE** | 2.29 | 14.73 | 3.61 | 10.84 | 2.21 | 4.91 | 6.43 |
| RN 50 A-TPT | **Acc.** | 83.56 | 39.83 | 62.08 | 16.41 | 85.06 | 59.42 | 57.82 |
| | **ECE** | 2.10 | 10.48 | 6.33 | 11.68 | 3.06 | 6.32 | 6.65 |
| RN 50 UC-TPT (ours) | **Acc.** | 82.04 | 36.94 | 66.83 | 16.58 | 87.50 | 59.25 | 58.08 |
| | **ECE** | 2.09 | 13.15 | 3.66 | 9.32 | 2.62 | 2.95 | **5.62** |

## R Calibration on Remote Sensing Datasets (RemoteCLIP)

In addition to the natural, fine-grained, and biomedical domains discussed in the main text, we further evaluate the calibration performance of our method on specialized aerial and remote sensing imagery. For this analysis, we utilize the domain-specific **RemoteCLIP** Liu et al. (2024) backbone across 7 diverse remote sensing datasets: MLRSNet, PatternNet, RESISC45, AID, UCM, EuroSAT, and RSICD.

As illustrated in Figure 16, we report the Expected Calibration Error (ECE) for various Test-Time Prompt Tuning methods.

Our proposed **UC-TPT** achieves the lowest average ECE (**3.32%**) across the 7 datasets. It successfully suppresses the severe overconfidence degradation introduced by standard TPT (15.53%) and A-TPT (7.88%), while also slightly improving upon the highly conservative Zero-Shot baseline (3.36%). This demonstrates a crucial safety benefit of our approach: even in high-difficulty scenarios, the uncertainty-aware objective acts as a strong safeguard. It ensures the tuned model does not collapse into unwarranted overconfidence, keeping its predictive probabilities faithfully aligned with its actual capabilities.

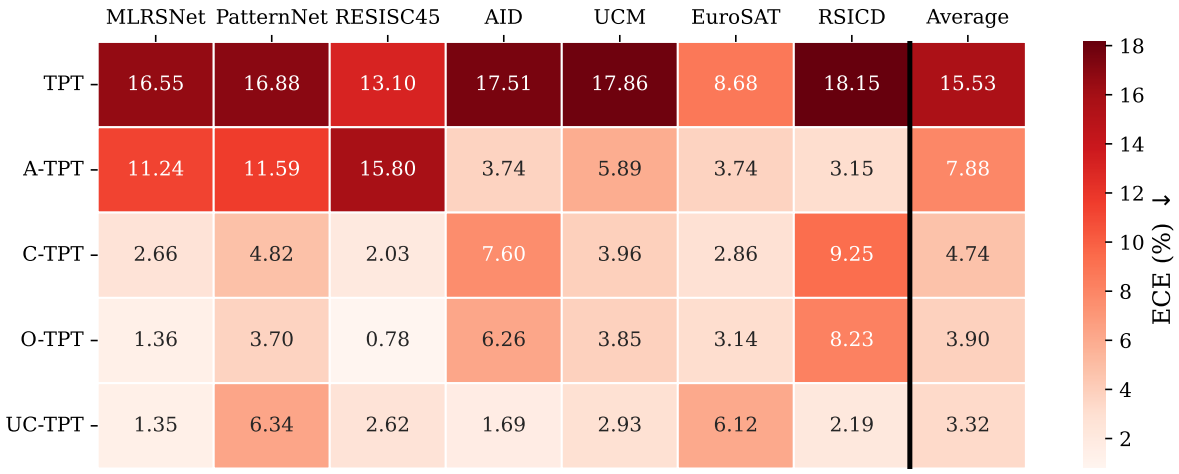



**Figure 16: Calibration performance (ECE % ↓) on Remote Sensing Datasets.** Using the RemoteCLIP backbone, UC-TPT achieves the lowest average calibration error compared to all other test-time tuning methods, effectively mitigating the severe overconfidence typically induced by standard TPT.

## S Batch-Wise Normalization, Sample Efficiency, and Adaptation Steps

**The Adaptation Batch and Normalization.** Following the standard TPT protocol Shu et al. (2022), UC-TPT processes a single test image at a time by generating a batch of $N = 64$ augmented views. In this context, our batch-wise min-max normalization (described in Appendix F) is strictly well-defined: it computes the *relative reliability* across these 64 variations of the same underlying instance. This allows the gating function to dynamically guide selective sharpening, suppressing updates for highly distorted or uninformative views while aggressively adapting on reliable ones.

**Sample Efficiency and Batch-Size Generalizability.** To evaluate the sample efficiency of UC-TPT, we ablate the number of augmented views per test instance. Table 19 reports the average performance across six representative datasets (DTD, Pets, Caltech, UCF, Flowers, Aircraft) using the ViT-B/16 backbone.

A known limitation of the general TPT paradigm is that decreasing the number of augmented views starves the model of the diverse distribution necessary to compute stable entropy gradients. To confirm that our normalization strategy does not introduce unique fragility, we compared UC-TPT against O-TPT under extreme low-view conditions. When reduced to 4 views, O-TPT degrades from 4.13 to 5.05 ECE. While UC-TPT similarly converges toward this noisy gradient floor at 4 views (4.91 ECE), it remains strictly superior

**Table 19: Ablation on the number of augmented views** ($N$). Average Accuracy and ECE across six datasets (DTD, Pets, Caltech, UCF, Flowers, Aircraft) with the ViT-B/16 backbone.

| Number of Views ($N$) | Avg. Acc (↑) | Avg. ECE (↓) |
|---|---|---|
| 4 | 62.95 | 4.91 |
| 16 | 63.06 | 3.79 |
| 32 | 63.18 | 3.58 |
| **64 (Default)** | **63.60** | **3.28** |

to the baseline. This confirms that the degradation stems from the inherently noisy low-view estimates of test-time prompt tuning, not from our uncertainty-calibrated framework.

**Strict Batch Size 1.** In a strict batch-size 1 regime where augmentations are entirely removed, batch-wise normalization becomes mathematically undefined. To adapt our method to this setting, we substitute the batch-wise min-max scaling with running statistics normalization (tracking a moving average of the minimum and maximum uncertainty scores observed during test time). Under this augmentation-free protocol, UC-TPT achieves 62.10% Accuracy and 5.12 ECE, demonstrating that the method can still function safely, though it sacrifices the performance gains unlocked by standard TPT augmentations. Furthermore, we emphasize that we operate under a strictly validation-free protocol, keeping all hyperparameters fixed across all datasets and batch-size variations.

**Test-Time Adaptation (TTA) Steps.** We also analyze the sensitivity of UC-TPT to the number of gradient steps taken during adaptation. Table 20 details the performance on ImageNet-V2 using the ResNet-50 (RN50) backbone as the number of adaptation steps is varied.

**Table 20: Ablation on the number of Test-Time Adaptation steps.** Evaluated on ImageNet-V2 with the RN50 backbone.

| TTA Steps | Accuracy (↑) | ECE (↓) |
|---|---|---|
| **1 (Default)** | 52.81 | **3.19** |
| 2 | **52.93** | 3.54 |
| 3 | 52.76 | 3.98 |
| 4 | 52.56 | 4.54 |
| 5 | 52.49 | 6.28 |

As shown in Table 20, increasing the number of adaptation steps beyond a single update strictly degrades calibration. While a second step yields a negligible accuracy bump (52.81% to 52.93%), the Expected Calibration Error (ECE) steadily worsens from 3.19 to 6.28 by the fifth step. This empirical trend confirms that extended optimization causes the prompt to overfit to the noisy test-time entropy objective, destabilizing the well-calibrated geometry of the pretrained embedding space.

Furthermore, extending adaptation to multiple steps is highly impractical for real-world deployment. Test-time prompt tuning is specifically designed for immediate, on-the-fly adaptation; multi-step TTA linearly multiplies both computational complexity and inference latency per sample. By restricting UC-TPT to a single, uncertainty-tempered step, we ensure that the framework remains lightweight and deployment-friendly without sacrificing model reliability.