RA-TPT: STOPPING LOGIT INFLATION AT TEST TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) enable strong zero-shot transfer, yet their predictive probabilities can become miscalibrated under distribution shift. Test-Time Prompt Tuning (TPT) adapts continuous prompts using unlabeled target inputs and can improve target-domain accuracy; however, its entropy-minimization objective may increase predictive concentration for incorrect or ambiguous samples. We propose **Reliability-Aware Test-Time Prompt Tuning (RA-TPT)**, a logit-space approach for improving the calibration-accuracy trade-off during test-time adaptation. RA-TPT replaces entropy minimization with an uncertainty-aware bounded-margin objective and a pseudo-label anchor that provide a self-training signal while limiting margin-expansion pressure on uncertain inputs. It further applies uncertainty-routed centered-logit dispersion regularization to discourage excessive relative logit variation when predictions are uncertain, with a complementary mean-logit regularizer. A soft conflict-aware gradient projection attenuates the regularization-gradient component that opposes the separability gradient. Across 31 benchmarks spanning standard classification, ImageNet distribution shifts, biomedical imaging, and remote sensing, RA-TPT improves the accuracy-calibration trade-off in CLIP-style adaptation, substantially reducing Expected Calibration Error (ECE) versus entropy-based TPT and matching calibration-aware baselines at competitive accuracy. Because the objective acts purely on logits, it further transfers to non-VLM test-time adaptation, matching strong corruption-robustness baselines on CIFAR-10/100-C and ImageNet-C.[1]

1 INTRODUCTION

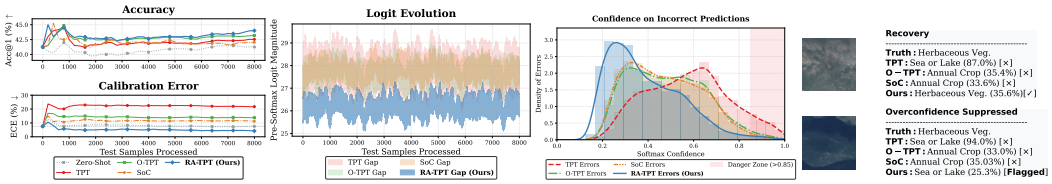
Vision–Language Models (VLMs) (Radford et al., 2021; Jia et al., 2021) provide strong zero-shot transfer through aligned visual and textual representations. However, their predictive probabilities can become miscalibrated under distribution shift, even when top-1 predictions remain useful. This is important when confidence informs downstream decisions in medical imaging, remote sensing, and autonomous systems. Test-Time Prompt Tuning (TPT) (Shu et al., 2022) adapts continuous prompt tokens using unlabeled target inputs while freezing the visual and textual encoders, providing lightweight test-time adaptation.

The challenge: adaptation can increase predictive concentration without improving reliability. TPT minimizes predictive entropy over augmented views. Although this can improve target-domain discrimination, it can also sharpen predictions for incorrect or intrinsically ambiguous inputs, worsening calibration (Figure 1(a)). Let $\mathbf{z} \in \mathbb{R}^K$ denote the pre-softmax logits for K classes and $\Delta = z_{(1)} - z_{(2)}$ the top-two logit gap. Entropy minimization promotes low-entropy softmax distributions by increasing relative logit separation. Without target labels, it has no direct mechanism to distinguish evidence for a correct prediction from confidence assigned to an incorrect one. Consequently, it can increase predictive concentration on erroneous examples under shift. We term this tendency toward increasing relative logit separation and predictive concentration *logit inflation*. It concerns relative logits, not a common additive offset, because $\mathrm{softmax}(\mathbf{z} + c\mathbf{1}) = \mathrm{softmax}(\mathbf{z})$ for any $c \in \mathbb{R}$.

Limitations of existing calibration-aware TPT. Recent methods including (Yoon et al., 2024; Ahamed et al., 2026; Sharifdeen et al., 2025; Fillioux et al., 2026) regularize textual features or their inter-class geometry during adaptation. These constraints provide useful structural inductive biases, but do not explicitly target an individual sample's predictive distribution or necessarily adapt to

[1] Code will be publicly released upon acceptance.

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



(a) (b) (c) (d)

Figure 1: **Calibration behavior during test-time prompt adaptation** on EuroSAT with CLIP ViT-B/16. **(a)** Accuracy and Expected Calibration Error (ECE) across Online steps; all methods start from the same zero-shot prompt state at step 0. **(b)** Post-adaptation top-two pre-softmax logit magnitude across the test stream. **(c)** Distribution of maximum confidence among incorrect predictions. **(d)** A representative target-domain prediction.

estimated uncertainty. We instead study a complementary approach that directly controls sample-level relative logit separation and dispersion. Figure 1 illustrates the motivation, and we compare this logit-space formulation with embedding-geometry methods across standard and distribution-shifted benchmarks.

Our approach. We propose **Reliability-Aware Test-Time Prompt Tuning (RA-TPT)**, a logit-space formulation for improving the calibration–accuracy trade-off during prompt adaptation. RA-TPT replaces entropy minimization with two complementary objectives. First, an **uncertainty-aware bounded margin** encourages class separation up to an instance-dependent target and is combined with a pseudo-label anchor to provide a self-training direction. Conditional on the estimated uncertainty and augmentation-coherence weight, the hinge term has zero subgradient with respect to the top-two gap after reaching its target. It therefore no longer directly expands the gap beyond the sample-specific cap, although the pseudo-label anchor remains active.

Second, RA-TPT introduces uncertainty-routed logit regularization. The **centered-logit dispersion term** penalizes excessive relative logit variation more strongly for uncertain predictions. Centering is essential because predictive probabilities are invariant to a common-mode logit shift. A complementary **mean-logit regularizer** controls the common-mode logit component and, at CLIP's logit scale, empirically contributes to calibration in addition to stabilizing optimization (Section 3.3). For fixed centered-logit dispersion, we further derive an upper bound on maximum softmax confidence; this characterizes predictive concentration rather than guaranteeing ECE.

The separability and regularization losses can yield locally opposing prompt gradients. RA-TPT therefore applies a **soft conflict-aware gradient projection** that attenuates the regularization-gradient component locally opposing the separability gradient while retaining the complete separability update. This provides a first-order reduction of gradient interference at an adaptation step, without implying global convergence or an accuracy guarantee.

Our contributions are:

- We identify a critical limitation of entropy-minimization TPT under distribution shift: it aggressively increases relative logit separation and predictive concentration, driving unwarranted overconfidence on incorrect predictions.
- We propose **RA-TPT**, combining uncertainty-aware bounded margin separation, pseudo-label anchoring, uncertainty-routed centered-logit dispersion regularization, and a complementary mean-logit regularizer.
- We introduce soft conflict-aware gradient projection, which reduces the locally conflicting regularization-gradient component while retaining the separability gradient.
- Across four VLM backbones (CLIP ViT-B/16 and RN50, BiomedCLIP, RemoteCLIP) and 31 benchmarks, RA-TPT improves the accuracy–calibration trade-off, reducing ECE relative to entropy-based TPT baselines at competitive accuracy; being logit-only, it further transfers to non-VLM adaptation, matching strong baselines on CIFAR-10/100-C and ImageNet-C.

2 Related Works

Vision–Language Prompt Tuning. VLMs, such as CLIP (Radford et al., 2021), enable zero-shot transfer through aligned image–text representations. Source-supervised methods such as CoOp (Zhou

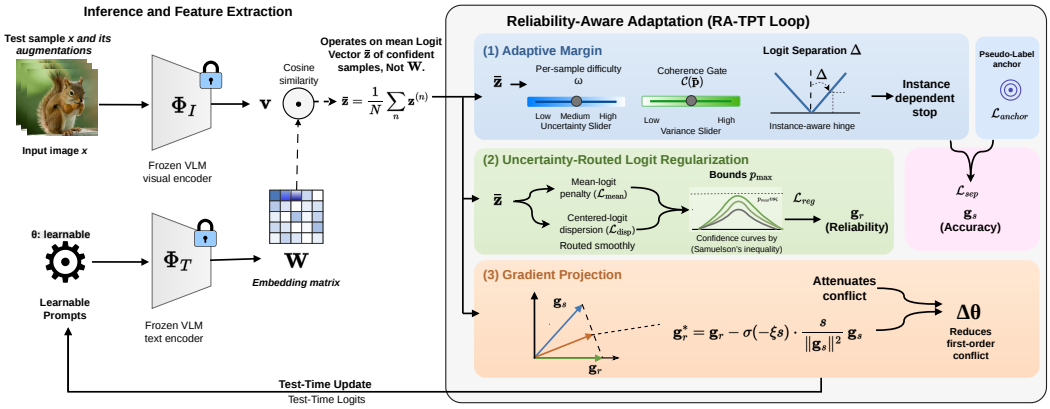


Figure 2: **Overview of the RA-TPT Framework.** *Left:* frozen encoders map test sample x to the mean logit vector $\bar{\mathbf{z}}$. *Right:* (1) an *Adaptive Margin* whose hinge stops once an instance-dependent cap is reached (sec: 3.2); (2) *uncertainty-routed logit regularization* bounding $p_{\max}$ via centered-logit dispersion, with a complementary mean-logit term (sec: 3.3); and (3) a soft *Gradient Projection* attenuating the regularization gradient where it conflicts with the separability update (sec: 3.4).

et al., 2022b), CoCoOp (Zhou et al., 2022a), and MaPLe (Khattak et al., 2023) adapt prompts with labeled source data, motivating unlabeled adaptation under unseen target shifts.

Test-Time Prompt Tuning. Tent (Wang et al., 2021) introduced entropy-minimization test-time adaptation, which TPT (Shu et al., 2022) extended to VLMs using augmented views. EATA (Niu et al., 2022), SAR (Niu et al., 2023), and R-TPT (Sheng et al., 2025) improve robustness through sample selection or gradient stabilization. However, entropy minimization cannot distinguish correct from incorrect target predictions and may increase predictive concentration on ambiguous samples.

Calibration-Aware TPT. FPP (Jang et al., 2026) uses flatness-aware pretraining, whereas C-TPT (Yoon et al., 2024), O-TPT (Sharifdeen et al., 2025), SoC (Fillioux et al., 2026), and A-TPT (Ahamed et al., 2026) regularize embedding-space geometry. These approaches apply uniform pressure across all inputs, failing to relax regularization for uncertain predictions. RA-TPT addresses this by acting directly on the logit vector, routing separation and centered dispersion through per-sample uncertainty. Complementary training-free methods avoid backpropagation via aggregation (Farina et al., 2024; Zanella & Ben Ayed, 2024) or caching (Karmanov et al., 2024), alongside post-hoc calibration (Wang et al., 2024); RA-TPT is orthogonal, utilizing logit-space gradients.

Margin and Energy-Based Regularization. Margin objectives (Cortes & Vapnik, 1995; Deng et al., 2019) and energy-based penalties (Liu et al., 2020; Wei et al., 2022) are widely used in supervised learning and OOD detection. RA-TPT adapts these ideas to unlabeled TTA through an entropy-routed bounded margin, augmentation-coherence weighting, and centered-logit dispersion regularization.

3 Methodology

3.1 Problem Formulation and Background

We consider a pretrained VLM with visual encoder $\Phi_I : \mathcal{X} \to \mathbb{R}^d$ and textual encoder $\Phi_T : \mathcal{Y} \to \mathbb{R}^d$. For K classes, learnable context tokens $\Theta \in \mathbb{R}^{M\times d}$ are prepended to class descriptors y_k, yielding $t_k(\Theta) = [\Theta; y_k]$. The corresponding text features are $\mathbf{w}_k = \Phi_T(t_k(\Theta))$, stacked as $\mathbf{W} \in \mathbb{R}^{d\times K}$. All VLM parameters are frozen; only Θ is adapted at test time.

For a test image $\mathbf{x}$, let $\mathbf{v} = \Phi_I(\mathbf{x}) \in \mathbb{R}^d$ be its ℓ_2-normalized visual feature. Classification is performed using temperature-scaled image–text similarity:

$$z_k = \frac{1}{\tau}\langle \mathbf{v}, \mathbf{w}_k \rangle, \qquad p_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}, \tag{1}$$

where $\tau > 0$ is the CLIP temperature. We write $\mathbf{z} \in \mathbb{R}^K$, $\mathbf{p} \in \Delta^{K-1}$, and $z_{(1)} \geq \cdots \geq z_{(K)}$.

Following TPT (Shu et al., 2022), we generate augmented views and retain the N views with the lowest predictive entropy. Let $\mathbf{z}^{(n)}$ and $p_{n,k} = \mathrm{softmax}(\mathbf{z}^{(n)})_k$ denote the logits and class probabilities of the retained views. We aggregate their logits and compute the corresponding predictive

distribution as

$$\bar{\mathbf{z}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{z}^{(n)}, \qquad \bar{\mathbf{p}} = \mathrm{softmax}(\bar{\mathbf{z}}). \tag{2}$$

Let

$$\Delta = \bar{z}_{(1)} - \bar{z}_{(2)}, \qquad \bar{z} = \frac{1}{K}\sum_{k=1}^{K}\bar{z}_k, \qquad \tilde{\mathbf{z}} = \bar{\mathbf{z}} - \bar{z}\mathbf{1}.$$

The centered-logit dispersion is

$$\varsigma^2 = \frac{1}{K}\|\tilde{\mathbf{z}}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}(\bar{z}_k - \bar{z})^2. \tag{3}$$

The pseudo-label is $k^* = \arg\max_k \bar{p}_k$. Prompts are adapted for T steps and reset to their initial value after processing each test input.

Entropy minimization encourages concentrated predictive distributions but, without target labels, cannot distinguish confidence associated with correct predictions from confidence associated with incorrect predictions. It can therefore increase relative logit separation and predictive concentration on ambiguous or shifted inputs; we refer to this tendency as *logit inflation*. This notion concerns relative logit differences and dispersion, rather than a common additive logit offset:

$$\mathrm{softmax}(\mathbf{z} + c\mathbf{1}) = \mathrm{softmax}(\mathbf{z}), \qquad \forall c \in \mathbb{R}. \tag{4}$$

RA-TPT consequently distinguishes shift-invariant logit quantities that can affect predictive probabilities from common-mode components that do not directly alter softmax probabilities.

Existing calibration-aware prompt-tuning methods regularize textual features or their pairwise geometry. Such regularizers can provide useful inductive biases, but they do not explicitly target the predictive distribution of an individual sample and need not be modulated by its estimated uncertainty. RA-TPT provides a complementary logit-space formulation that controls relative logit separation and centered-logit dispersion, while separately regularizing the common-mode logit component for optimization stability.

As shown in Figure 2, RA-TPT combines an uncertainty-aware bounded margin, uncertainty-routed logit regularization, and soft conflict-aware gradient projection. The margin encourages separation for coherent inputs up to an instance-dependent cap; the regularization objective combines centered-logit dispersion with a complementary mean-logit penalty; and the projection attenuates local conflict between their prompt gradients.

3.2 Adaptive Margin: Bounding Relative Logit Separation

We estimate uncertainty using the normalized entropy of the averaged prediction:

$$\omega = -\frac{1}{\log K}\sum_{k=1}^{K}\bar{p}_k \log \bar{p}_k \in [0, 1], \tag{5}$$

where $\omega \approx 0$ indicates a concentrated predictive distribution and $\omega \approx 1$ indicates a diffuse predictive distribution. As ω is obtained from the model prediction, it serves as an uncertainty proxy rather than a direct measure of correctness. To measure agreement among augmented views, we define

$$\mathcal{C}(\bar{\mathbf{p}}) = \exp(-\gamma\mathcal{V}(\bar{\mathbf{p}})) \in (0, 1], \tag{6}$$

where $\gamma > 0$ and

$$\mathcal{V}(\bar{\mathbf{p}}) = \sum_{k=1}^{K}\frac{1}{N}\sum_{n=1}^{N}(p_{n,k} - \bar{p}_k)^2 \tag{7}$$

quantifies disagreement among retained views. The factor $\mathcal{C}(\bar{\mathbf{p}})$ downweights the margin loss when retained views disagree. Conditional on the observed views, it rescales the corresponding gradient; it is not an unbiased gradient estimator.

We specify a smaller target gap for uncertain predictions:

$$m_{\mathrm{cap}}(\omega) = \max(0, \alpha - \lambda\omega), \tag{8}$$

where $\alpha > 0$ denotes the maximum target margin and $\lambda > 0$ controls the reduction of the target gap with uncertainty. For $\omega \geq \alpha/\lambda$, the target margin is zero and the hinge term does not request additional separation. The margin loss is

$$\mathcal{L}_{\mathrm{margin}} = \mathcal{C}(\bar{\mathbf{p}})(1 - \omega)[m_{\mathrm{cap}}(\omega) - \Delta]_+, \tag{9}$$

Table 1: Behavior of the bounded-margin objective across the uncertainty range; the zero-cap case is a limiting regime (requires $\lambda \geq \alpha$) not entered under our settings (Appendix Q).

Sample type	ω	$m_{\text{cap}}(\omega)$	Margin pressure
Concentrated prediction	≈ 0	$\approx \alpha$	Strong when views are coherent
Moderately diffuse prediction	≈ 0.5	$\alpha - 0.5\lambda$	Reduced
Highly diffuse prediction	$\geq \alpha/\lambda$	0	None

where $[a]_+ = \max(a, 0)$. Thus, the margin term provides stronger direct separation pressure for coherent, concentrated predictions; this pressure decreases as ω increases and becomes inactive once the instance-dependent cap is attained. Table 1 summarizes these operating regimes.

Claim 3.1 (Bounded hinge contribution)**.** For fixed ω and $\mathcal{C}(\bar{\mathbf{p}})$,

$$\frac{\partial \mathcal{L}_{\text{margin}}}{\partial \Delta} \in \begin{cases} \{-\mathcal{C}(1-\omega)\}, & \Delta < m_{\text{cap}}(\omega), \\ [-\mathcal{C}(1-\omega), 0], & \Delta = m_{\text{cap}}(\omega), \\ \{0\}, & \Delta > m_{\text{cap}}(\omega). \end{cases} \tag{10}$$

Thus, conditional on ω and $\mathcal{C}(\bar{\mathbf{p}})$, the hinge contribution has zero subgradient with respect to Δ after the sample-specific cap is reached. *(Proof in Appendix B.2.)*

The claim is the standard bounded-margin property of a hinge loss. It applies to $\mathcal{L}_{\text{margin}}$ alone and does not imply that the complete adaptation objective has a stationary point once the cap is reached. To provide a class-specific self-training direction, we use the current pseudo-label k^*:

$$\mathcal{L}_{\text{sep}} = \mathcal{L}_{\text{margin}} + \eta \mathcal{L}_{\text{anchor}}, \qquad \mathcal{L}_{\text{anchor}} = -\log \bar{p}_{k^*}, \tag{11}$$

where $\eta = 0.5$. The anchor provides a self-training signal, but it can also increase predictive concentration. The uncertainty-routed regularizer introduced next counterbalances this pressure for uncertain predictions. Its ablation is reported in Table 6.

3.3 Uncertainty-Routed Logit Regularization

The top-two gap does not fully characterize the predictive distribution. We therefore decompose the uncentered logit energy into centered and mean components:

$$\mathcal{L}_{\text{disp}} = \frac{1}{K}\|\tilde{\mathbf{z}}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}(\bar{z}_k - \bar{z})^2 = \varsigma^2, \qquad \mathcal{L}_{\text{mean}} = \bar{z}^2. \tag{12}$$

These terms satisfy

$$\frac{1}{K}\|\bar{\mathbf{z}}\|_2^2 = \mathcal{L}_{\text{disp}} + \mathcal{L}_{\text{mean}}. \tag{13}$$

Only $\mathcal{L}_{\text{disp}}$ is invariant under $\bar{\mathbf{z}} \mapsto \bar{\mathbf{z}} + c\mathbf{1}$ and is directly related to softmax concentration; it is the primary mechanism we analyze. The mean term is not shift-invariant and empirically contributes to calibration (Table 5), not merely to optimization stability.

Large centered-logit dispersion can be useful for clear predictions, whereas excessive dispersion can yield overly concentrated probabilities on ambiguous inputs. We therefore define

$$\mathcal{L}_{\text{reg}} = \underbrace{\omega^2 \mathcal{L}_{\text{disp}}}_{\text{dispersion regularization}} + \underbrace{\beta(1-\omega^2)\mathcal{L}_{\text{mean}}}_{\text{mean-logit regularization}}, \qquad \beta \geq 0. \tag{14}$$

The centered-dispersion penalty increases with the estimated uncertainty, whereas the complementary mean-logit penalty receives greater weight for more concentrated predictions. We use quadratic routing because it satisfies $r(0) = 0$, $r(1) = 1$, and $r(\omega) < \omega$ for $\omega \in (0, 1)$. This is a smooth routing choice rather than a uniquely determined one. Its empirical effect is evaluated in Table 5 and comparison with other simple regularizers is shown in Appendix E.

Theorem 3.1 (Maximum-confidence bound from centered-logit dispersion)**.** *Let* $\varsigma^2 = \mathcal{L}_{\text{disp}}$ *and* $K \geq 2$*. By Samuelson's inequality (Samuelson, 1968),*

$$\bar{z}_{(1)} - \bar{z}_{(K)} \leq 2\varsigma\sqrt{K-1}. \tag{15}$$

Consequently,

$$p_{\max} \leq \frac{1}{1 + (K-1)\exp\left(-2\varsigma\sqrt{K-1}\right)}, \tag{16}$$

where $p_{\max} = \max_k \bar{p}_k$. *The right-hand side is increasing in* ς*; hence, smaller centered-logit dispersion yields a smaller upper bound on* $p_{\max}$. (Proof in Appendix B.4.)

Theorem 3.1 concerns only $\mathcal{L}_{\text{disp}}$. It establishes a pointwise relation between centered-logit dispersion and maximum softmax confidence; it does not provide a bound on ECE or general calibration, which also depend on empirical correctness. The mean term $\mathcal{L}_{\text{mean}}$ does not enter this result.

3.4 Soft Conflict-Aware Gradient Projection

The separability and regularization objectives can induce locally competing prompt updates. Let

$$\mathbf{g}_s = \nabla_\Theta \mathcal{L}_{\text{sep}}, \qquad \mathbf{g}_r = \nabla_\Theta \mathcal{L}_{\text{reg}}, \qquad s = \langle \mathbf{g}_r, \mathbf{g}_s \rangle. \tag{17}$$

When $s < 0$, the regularization gradient has a component that locally opposes the separability gradient. Inspired by PCGrad (Yu et al., 2020), we use the sigmoid-smoothed projection

$$\mathbf{g}_r^* = \mathbf{g}_r - \underbrace{\sigma(-\xi s)}_{\phi \in (0,1)} \frac{s}{\|\mathbf{g}_s\|_2^2 + \epsilon} \mathbf{g}_s, \tag{18}$$

where $\xi > 0$ controls projection sharpness and $\epsilon > 0$ is a numerical stability constant. Prompts are updated as

$$\Theta \leftarrow \Theta - \rho(\mathbf{g}_s + \mathbf{g}_r^*), \qquad \rho > 0. \tag{19}$$

For the idealized case $\epsilon = 0$,

$$\langle \mathbf{g}_r^*, \mathbf{g}_s \rangle = s\,(1 - \sigma(-\xi s))\,. \tag{20}$$

For $s < 0$, the magnitude of the residual inner product is smaller than $|s|$ and approaches zero as $\xi \to \infty$. The implementation uses $\epsilon > 0$, which slightly modifies this equality while retaining the intended attenuation. Hence, the projection reduces local first-order interference while retaining $\mathbf{g}_s$ in the update. It does not, by itself, guarantee unchanged top-1 accuracy, improved calibration, or convergence under repeated adaptation.

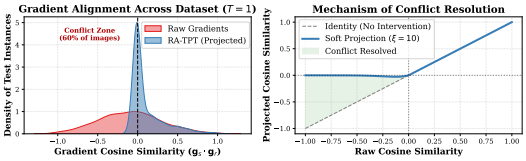


Figure 3: **Gradient conflict at one adaptation step.** Evaluated on Caltech101. Negative values of $s = \langle \mathbf{g}_r, \mathbf{g}_s \rangle$ indicate local conflict. Soft projection reduces the magnitude of negative residual inner products in this setting.

We project $\mathbf{g}_r$ because $\mathcal{L}_{\text{sep}}$ supplies the primary self-training direction, whereas $\mathcal{L}_{\text{reg}}$ supplies a calibration-oriented and optimization-regularization signal. This choice retains the complete separability gradient and attenuates only the locally conflicting component of the regularization gradient. A comparison with reverse projection is reported in Table 7.

The final objectives are

$$\begin{aligned} \mathcal{L}_{\text{sep}} &= \mathcal{C}(\bar{\mathbf{p}})(1-\omega)[m_{\text{cap}}(\omega) - \Delta]_+ + \eta(-\log \bar{p}_{k^*}), \\ \mathcal{L}_{\text{reg}} &= \omega^2 \frac{1}{K} \|\bar{\mathbf{z}} - \bar{z}\mathbf{1}\|_2^2 + \beta(1-\omega^2)\bar{z}^2. \end{aligned} \tag{21}$$

Rather than using the raw gradient of the scalar sum $\mathcal{L}_{\text{sep}} + \mathcal{L}_{\text{reg}}$, RA-TPT computes separate gradients and applies the update in Equations (18)–(19). The resulting update retains the separability gradient while attenuating the component of the regularization gradient that locally conflicts with it.

4 Experiments

Datasets. We evaluate RA-TPT on 31 benchmarks spanning general, fine-grained, and specialized domains: **10 standard datasets**, **4 ImageNet variants** for natural distribution shifts, **7 remote-sensing**, and **11 biomedical** datasets. EuroSAT is shared between the CLIP and RemoteCLIP suites. We use CLIP (Radford et al., 2021) (ViT-B/16 and RN50), RemoteCLIP (Liu et al., 2024), and BiomedCLIP (Zhang et al., 2025) backbones. Full dataset details are provided in Appendix G.

Implementation and protocol. We evaluate RA-TPT on CLIP (ViT-B/16, RN50), BiomedCLIP, and RemoteCLIP against recent test-time tuning baselines, using hard-prompt initialization, a frozen

backbone, and one AdamW step (lr 1×10^{-3}); results are averaged over three runs. Following TPT, augmented views are ranked by predictive entropy and the most-confident subset is retained: the entropy baselines keep their published top-10%, whereas RA-TPT keeps the top 50%, as its bounded regularizer tolerates noisier views. We disclose both settings and, as a control, evaluate the strongest baselines at both retention levels (Appendix H): RA-TPT attains the lowest ECE at both, so its advantage reflects the objective, not the view budget.

Hyperparameters. RA-TPT has five tuned hyperparameters $\{\alpha, \lambda, \gamma, \xi, \beta\}$ (β=0.3 weights a complementary stabilizer and is insensitive across benchmarks), plus two fixed choices: the anchor weight η=0.5 and the top-50% view-retention ratio. All five are selected once on a held-out ImageNet subset following (Shu et al., 2022) and fixed across all 31 benchmarks, including the unseen biomedical and remote-sensing domains. Appendix K shows robustness of hyperparameters in a band around each value. On non-VLM backbones the fixed value of α does not transfer, as their logit scale differs from CLIP's; the vision extension instead sets α from the source top-two logit gap with a single validation-selected scale constant (Appendix F).

4.1 Main Results

Table 2: **Calibration (ECE↓) and Accuracy (Acc.↑)** with **CLIP-ViT-B/16** across ten standard datasets and four ImageNet variants. **Bold** = best average values compared to second best; underline = second best; shaded = our method. **Note**: All baselines are rerun from the official codebases under paper-reported hyperparameter settings. Standard deviation is reported in **Appendix D**.

Method	Metric	Ten Standard Datasets										Avg.	ImageNet Variants				Avg.
		INet	DTD	FLW	Food	Air.	Pets	C101	UCF	SAT	Car		I-A	I-V2	I-R	I-S	
Zero-Shot (ICML 2021)	Acc.	66.7	44.4	67.3	83.8	23.7	88.1	92.9	65.0	41.2	65.3	63.9	47.7	60.8	74.0	46.1	57.2
	ECE	2.07	8.30	2.63	2.32	5.19	4.45	5.51	3.31	7.55	4.29	4.56	8.40	2.79	3.59	4.84	4.90
TPT (NeurIPS 2022)	Acc.	69.0	46.8	69.3	84.8	23.3	87.1	93.9	67.6	42.8	66.2	65.1	52.9	63.1	76.9	47.9	60.2
	ECE	10.57	21.04	13.03	3.97	16.75	5.69	4.52	12.65	21.42	5.33	11.50	16.37	11.14	4.36	14.94	11.70
R-TPT (CVPR 2025)	Acc.	66.5	44.5	67.2	83.6	23.7	88.1	92.7	64.6	41.3	65.2	63.7	47.8	60.8	74.0	46.2	57.2
	ECE	10.25	18.81	10.69	3.42	12.66	5.40	3.64	12.12	22.30	2.01	10.13	13.95	11.52	4.58	14.02	11.02
C-TPT (ICLR 2024)	Acc.	68.0	45.4	69.4	83.1	23.3	88.0	93.2	65.2	43.0	65.3	64.4	49.4	62.0	74.8	47.3	58.4
	ECE	3.18	12.59	5.15	3.69	4.43	1.89	4.43	2.45	13.24	1.65	5.27	6.45	4.51	2.87	7.23	5.26
A-TPT (ICLR 2026)	Acc.	67.6	45.5	65.9	83.2	23.3	84.1	92.0	63.3	42.8	63.0	63.1	51.0	62.5	76.2	47.7	59.3
	ECE	2.28	5.72	4.22	7.06	6.26	5.79	10.04	3.19	4.71	5.53	5.48	8.92	6.47	4.38	10.80	7.64
O-TPT (CVPR 2025)	Acc.	66.7	45.6	69.2	82.8	23.3	87.7	93.2	64.1	43.0	64.9	64.0	49.9	61.6	75.2	47.0	58.4
	ECE	2.12	8.13	3.75	4.61	3.97	2.02	4.65	2.29	13.72	1.65	4.69	7.14	4.01	2.15	6.93	5.06
SoC (CVPR 2026)	Acc.	67.2	45.2	68.8	82.7	23.4	88.1	93.0	63.6	42.0	65.3	63.9	48.3	61.3	74.0	46.6	57.6
	ECE	1.98	7.40	3.49	4.47	3.69	2.19	3.93	2.93	11.17	1.96	4.32	6.67	3.50	3.93	5.86	4.99
RA-TPT (Ours)	**Acc.**	67.6	45.5	68.3	84.2	23.7	88.8	93.2	65.6	44.5	65.2	**64.7**	51.1	61.2	74.7	46.8	**58.5**
	ECE	2.45	7.87	3.45	2.36	5.33	3.19	4.58	2.24	4.28	4.21	**3.99**	7.06	3.23	4.09	3.85	**4.56**

We evaluate accuracy and calibration against zero-shot CLIP and prior test-time prompt-tuning baselines using Expected Calibration Error (ECE). Table 2 and Table 9 summarize CLIP-ViT-B/16 and CLIP-RN50 results across ten standard datasets and four distribution shifts.

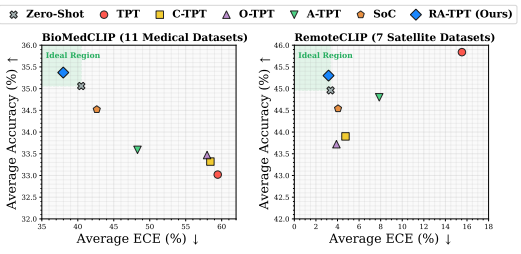


Figure 4: **Domain generalization performance on BioMedCLIP (left) and RemoteCLIP (right).** Average accuracy and ECE for the evaluated methods. Points closer to the upper-left correspond to higher accuracy and lower ECE.

Performance with CLIP-ViT-B/16. On the ten standard datasets, TPT attains the highest average accuracy (65.1%) but has substantially higher ECE (11.50) than RA-TPT (3.99 ECE, 64.7% accuracy); a single global post-hoc temperature, fit on one domain, does not transfer across heterogeneous domains (Appendix N). Across ImageNet variants, RA-TPT attains the lowest average ECE (4.56) while maintaining 58.5% average accuracy. CLIP-RN50 results (Table 9 and Appendix C.1) show a similar accuracy–ECE trade-off: TPT attains 57.5% and 43.5% average accuracy on standard datasets and ImageNet variants, respectively, but its ECE rises to 11.79 and 16.75. RA-TPT reduces these ECE values to 5.07 and 6.34 while attaining 56.8% and 42.1% accuracy.

Domain-Specific Foundation Models. We further evaluate RA-TPT using BioMedCLIP (ViT-B/16) on 11 medical datasets and RemoteCLIP (RN50) on seven remote-sensing datasets. On BioMedCLIP,

RA-TPT improves average accuracy from 35.06% for zero-shot inference to 35.37% and reduces average ECE from 40.48 to 37.98; TPT yields 33.02% accuracy and 59.45 ECE. The remaining ECE highlights the difficulty of calibrated adaptation in this domain. On RemoteCLIP, RA-TPT obtains the lowest average ECE (3.17) with 45.27% average accuracy, compared with 44.97% and 3.36 for zero-shot inference, and 45.84% and 15.53 for TPT. Full tabular results are provided in Appendices C.2 and C.3.

Generality beyond VLMs. Although RA-TPT is introduced for prompt tuning, its objective is defined on the logit vector $\bar{\mathbf{z}}$ and applies to other adapted parameters. For standard classifiers we update normalization affine parameters rather than prompts, retaining $\mathcal{L}_{\text{sep}}$, $\mathcal{L}_{\text{reg}}$, and the projection rule (which is parameter-agnostic; Appendix F). On CIFAR-10-C, CIFAR-100-C, and ImageNet-C (Hendrycks & Dietterich, 2019), using WideResNet-28-10 (Zagoruyko & Komodakis, 2016) and ResNet-50 (He et al., 2016) backbones (Croce et al., 2021), RA-TPT matches CoTTA (Wang et al., 2022) (Table 3) on CIFAR-10-C (13.77 vs. 13.76) and ImageNet-C (3.06 vs. 3.07), and is within 0.5 ECE on CIFAR-100-C, as a single-model method. Tent (Wang et al., 2021) is higher-accuracy but far less calibrated on ImageNet-C (10.70 ECE). This shows the logit-space objective extends beyond VLM prompt tuning; full setup, accuracy, temperature scaling, and α-sensitivity are in Appendix F.

Table 3: ECE (↓) on standard non-VLM corruption benchmarks (severity 5, 15 corruptions). RA-TPT matches CoTTA at comparable accuracy (Appendix F).

Method	CIFAR-10-C	CIFAR-100-C	ImageNet-C
Source	34.00	35.54	15.60
Tent	23.01	34.55	10.70
CoTTA	**13.76**	**16.22**	3.07
RA-TPT	13.77	16.63	**3.06**

5 Ablation Analysis

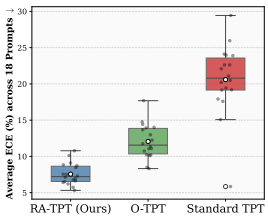


Figure 5: **Robustness to Prompt Initialization.**

Component analysis. Table 4 evaluates each component on ViT-B/16 across nine standard datasets excluding ImageNet. Entropy minimization alone reaches 64.64% accuracy but has 11.60 ECE. Adding $\mathcal{L}_{\text{reg}}$ reduces ECE to 8.10, while combining $\mathcal{L}_{\text{sep}}$ and $\mathcal{L}_{\text{reg}}$ yields 4.29 ECE. Adding gradient projection gives 64.33% accuracy and 4.17 ECE, suggesting complementary effects from separation, regularization, and projection in this aggregate setting.

Prompt robustness. Figure 5 evaluates 18 prompt templates on ResNet-50 across Aircraft, DTD, and UCF101. RA-TPT attains the lowest mean ECE and the smallest observed variation among the evaluated methods. Per-template results are given in Appendix M.

Calibration objective. Table 5 ablates $\mathcal{L}_{reg}$ on ViT-B/16 across nine standard datasets excluding ImageNet. The $\mathcal{L}_{mean}$-only, $\mathcal{L}_{disp}$-only, and unweighted variants underperform the routed alternatives. Linear routing obtains 64.18% accuracy and 4.39 ECE, while quadratic routing (ω^2) in Equation 14 obtains 64.33% and 4.17 ECE. This supports the empirical utility of uncertainty-dependent routing in the evaluated setting.

Table 4: **Loss ablation study** on ViT-B/16. Avg results.

Components				Average	
$\mathcal{L}_{ent}$	$\mathcal{L}_{sep}$	$\mathcal{L}_{reg}$	Grad	Acc.	ECE
✓	✗	✗	✗	**64.64**	11.60
✓	✗	✓	✗	63.79	8.10
✓	✗	✓	✓	63.78	5.57
✗	✓	✗	✗	64.06	8.67
✗	✓	✓	✗	63.82	4.29
✗	✓	✓	✓	64.33	**4.17**

Table 5: **Calibration ($\mathcal{L}_{reg}$).** Impact of routing parameters.

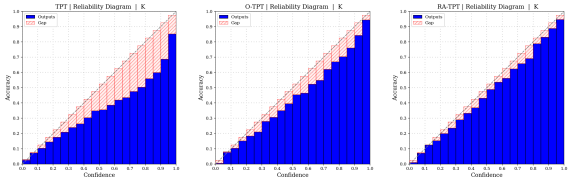
Formulation ($\mathcal{L}_{reg}$)	**Acc. (↑)**	**ECE (↓)**
$\mathcal{L}_{mean}$ only	63.68	5.29
$\mathcal{L}_{disp}$ only	63.84	4.48
Unweighted Sum	63.72	4.51
Linear (ω)	64.18	4.39
Squared (ω^2)	**64.33**	**4.17**

Table 6: **Separability ($\mathcal{L}_{sep}$).** Impact of margin dynamics.

Formulation ($\mathcal{L}_{sep}$)	**Acc. (↑)**	**ECE (↓)**
Anchor Loss Only	63.91	4.78
Static Margin	63.81	4.52
w/o Coherence $\mathcal{C}$	63.70	4.37
w/o weighting $(1-\omega)$	63.89	4.63
Full Formulation	**64.33**	**4.17**

Separability objective. Table 6 ablates $\mathcal{L}_{sep}$ on ViT-B/16 across nine standard datasets excluding ImageNet. The $\mathcal{L}_{anchor}$-only variant obtains 4.78 ECE. A static margin, removing the coherence gate $\mathcal{C}$, or omitting the $(1-\omega)$ weight increases ECE relative to the full formulation. The full formulation obtains the best aggregate accuracy (64.33%) and ECE (4.17) among these variants.

Comprehensive calibration. Figure 7 reports SCE, MCE, ACE, Brier score, and NLL. RA-TPT obtains the lowest aggregate values among the evaluated methods for these metrics, indicating that



(a) TPT (b) O-TPT (c) RA-TPT (Ours)

Figure 6: **Reliability Diagrams.** The red hatched area indicates the gap between predicted confidence and empirical accuracy. RA-TPT shows closer alignment with the ideal diagonal than TPT and O-TPT in this evaluation.

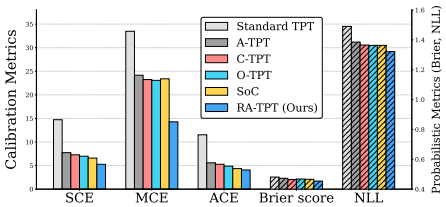


Figure 7: **Comprehensive Calibration Analysis.** Evaluation across five metrics. The left axis maps to calibration metrics, while the right axis maps to probabilistic metrics.

ECE improvements are accompanied by improvements under several calibration and proper-scoring measures. Details and definitions are provided in Appendix I.

Reliability diagrams. Figure 6 compares predicted confidence and empirical accuracy for TPT, O-TPT, and RA-TPT on ImageNet-Sketch with ViT-B/16. TPT exhibits visible overconfidence in high-confidence bins, while O-TPT reduces part of this gap. RA-TPT shows closer alignment with the ideal diagonal in this evaluation. Further analyses are provided in Appendices O.1 and O.2.

Gradient Projection Strategy. Table 7 compares conflict-resolution strategies (Section 3.4). Randomly dropping gradients when $\langle \mathbf{g}_r, \mathbf{g}_s \rangle < 0$ yields 4.96 ECE. Reverse projection, which projects $\mathbf{g}_s$, attains the lowest ECE (4.10) but lower accuracy. Compared with hard projection, soft projection achieves the highest reported accuracy (64.33%) and 4.17 ECE. These results support preserving $\mathbf{g}_s$ as the primary update direction in the evaluated configuration.

Table 7: **Ablation of Gradient Projection Strategy.** Average results across 9 standard datasets (excl. ImageNet) on ViT-B/16.

Conflict Resolution Strategy	Acc. (↑)	ECE (↓)
Random Drop on Conflict	63.72	4.96
Reverse Projection (Project $\mathbf{g}_s$)	63.44	**4.10**
Hard Orthogonal Projection ($\xi \to \infty$)	64.11	4.24
Soft Projection (Ours: Eq. 18)	**64.33**	<u>4.17</u>

Computational overhead. Table 8 reports efficiency on ImageNet-Sketch with RN50. O-TPT's orthogonality objective operates on the full class-embedding matrix $\mathbf{W} \in \mathbb{R}^{d \times K}$ via a dense $K \times K$ Gram ($\mathcal{O}(K^2 d)$ compute) and empirically peaks at 39.7 GB VRAM. RA-TPT's logit-space regularizer is only $\mathcal{O}(K)$ in the class count and matches TPT's 17.9 GB peak, at 542 ms latency versus 491 ms for TPT. On this benchmark, RA-TPT obtains 35.1% accuracy and 1.18 ECE, versus 35.3% and 13.67 for TPT. All methods still incur the common VLM forward and backward cost; detailed complexity analysis is in Appendix L.

Table 8: **Deployment Overhead.** Evaluated on ImageNet-Sketch (RN50). The additional RA-TPT logit-space regularizer has $\mathcal{O}(K)$ complexity, avoiding the $\mathcal{O}(K^2 d)$ pairwise embedding-matrix operation of O-TPT.

Method	Optimization Space		Hardware / Batch		Performance (I-S)	
	Target	Complexity	Peak VRAM	Lat.	Acc. (↑)	ECE (↓)
TPT	Entropy (p)	$\mathcal{O}(K)$	**17.9 GB**	**491ms**	**35.3**	13.67
O-TPT	Orthogonality ($\mathbf{W}$)	$\mathcal{O}(K^2 \cdot d)$	39.7 GB	534ms	34.1	5.06
RA-TPT	Logit-space (z)	$\mathcal{O}(\mathbf{K})$	**17.9 GB**	542ms	<u>35.1</u>	**1.18**

RA-TPT also degrades gracefully with fewer views: on ImageNet-V2 at a single view ($N = 1$) it stays calibrated (3.08 ECE vs. 15.88 for TPT), though $N = 1$ gives no gain over zero-shot. Full augmentation and adaptation-step studies are in Appendices J and Q.

6 Conclusions

We presented RA-TPT to improve the calibration-accuracy trade-off in Vision-Language Models. Identifying entropy-driven excessive predictive concentration under shift, we introduce three coupled components: an uncertainty-aware bounded margin with an instance-dependent hinge cap, uncertainty-routed centered-logit dispersion regularization with a complementary mean-logit penalty, and soft gradient projection to reduce local update conflicts. Our analysis offers supporting motivation, for the bounded hinge, the dispersion penalty (via a loose confidence bound), and local conflict attenuation, rather than formal guarantees. Experiments across 31 benchmarks show RA-TPT substantially improves ECE over entropy-based baselines and attains calibration comparable to calibration-aware baselines at competitive accuracy in CLIP-style adaptation. Its pure logit-space formulation further transfers to standard vision-only classifiers, matching strong corruption-robustness baselines.

AI USE STATEMENT

In this work, we used generative AI tools only for writing assistance (grammar correction and phrasing of the manuscript). Generative AI tools were not used for research ideation, experimental design, data analysis, or the derivation of the theoretical results presented in this paper. All AI-assisted text has been reviewed by the authors, and we take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work studies calibration of test-time prompt tuning for vision-language models using publicly available benchmark datasets. No human subjects were involved and no personally identifiable information was collected. We note that improved calibration is directly motivated by safety concerns in high-stakes deployment (e.g., medical imaging and remote sensing); nevertheless, our method does not eliminate the possibility of confident errors, and deployed systems should retain human oversight.

REPRODUCIBILITY STATEMENT

Full implementation details, hyperparameter settings, and the evaluation protocol are described in Section 4 and Appendix H. Complete proofs of all theoretical claims are provided in Appendix B. Dataset details and splits are documented in Appendix G. Source code, pre-trained weights, and evaluation scripts will be released publicly upon publication.

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 *International Conference on Learning Representations (ICLR)*, 2026.

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.

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 (ECCV)*, pp. 446–461. Springer, 2014.

Pingjun Chen. Knee osteoarthritis severity grading dataset; mendeley data, v1, 2018.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. Ieee, 2009.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *NeurIPS*, 2024.

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.

Leo Fillioux, Omprakash Chakraborty, Ismail Ben Ayed, Paul-Henry Cournède, Stergios Christodoulidis, Maria Vakalopoulou, and Jose Dolz. SoC: Semantic orthogonal calibration for test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS IEEE International Geoscience and Remote Sensing Symposium*, pp. 204–207. IEEE, 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 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 (ICCV)*, 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.

Hyeonseo Jang, Jaebyeong Jeon, Joong-Won Hwang, and Kibok Lee. Improving calibration in test-time prompt tuning for vision-language models via data-free flatness-aware prompt pretraining. 2026.

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 (ICML)*, pp. 4904–4916. PMLR, 2021.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *CVPR*, 2024.

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.

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, 2018.

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

Thomas Köhler, Attila Budai, Martin F Kraus, Jan Odstrčilik, Georg Michelson, and Joachim Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems (CBMS)*, pp. 95–100. IEEE, 2013.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, Dec 2013. doi: 10.1109/ICCVW.2013.77.

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.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

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.

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

Masoud Nickparvar. Brain tumor MRI dataset. `https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset`, 2021. Accessed: 2026-05-06.

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

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning (ICML)*, pp. 16888–16905. PMLR, 2022.

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

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

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, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Paul A. Samuelson. How deviant can you be? *Journal of the American Statistical Association*, 63 (324):1522–1525, 1968. doi: 10.1080/01621459.1968.10480944.

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.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Xiao Chaowei. Test-time prompt tuning for vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 34275–34289, 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.

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. 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. In *International Conference on Learning Representations (ICLR)*, 2021.

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 (NeurIPS)*, 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 *ICML*, 2024.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 23631–23644, 2022.

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.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 5824–5836, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, 2016.

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

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022b.

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.

RA-TPT: Stopping Logit Inflation at Test Time

A Limitations and Future Directions

Limitations. RA-TPT is designed to improve the empirical calibration–accuracy trade-off in one-step prompt adaptation, but it has several limitations.

First, improved calibration can be accompanied by a reduction in top-1 accuracy relative to unconstrained entropy-minimization TPT. In our experiments, this trade-off is typically modest, but it is not uniformly favorable on every dataset or domain. In particular, lower ECE alone does not establish that a system is safe for high-stakes use: calibration is one component of reliability and does not eliminate confident errors, distributional failures, or downstream decision risks.

Second, RA-TPT requires separate gradients for the separability and dispersion objectives. The resulting additional backward-pass cost increases latency relative to standard TPT, even though the logit-space regularizer itself is inexpensive. This cost can be consequential in resource-constrained or real-time settings. Future work could investigate shared-gradient approximations, lower-cost conflict estimates, or alternatives that reduce the number of backward passes.

Third, the method depends on the zero-shot model providing a sufficiently useful pseudo-label and on augmented views providing a meaningful uncertainty signal. Under severe shifts for which the initial vision–language alignment is uninformative. For example, large cross-sensor shifts such as optical-to-SAR imagery, the pseudo-label anchor may reinforce an incorrect prediction and the entropy-derived routing signal may be unreliable. Stronger multimodal priors, source-free adaptation objectives, or external uncertainty estimates may be needed in such settings.

Finally, the current study focuses on CLIP-style prompt adaptation and primarily evaluates one adaptation update per test input. The empirical behavior of RA-TPT under repeated multi-step updates, correlated test streams, and non-VLM test-time adaptation settings requires further study.

Future directions. A promising direction is to extend the approach from image-level classification to dense prediction. In remote sensing, medical imaging, and agricultural monitoring, different spatial regions of one image can experience different degrees of shift and uncertainty. A patch-level formulation that controls local predictive concentration and reports spatial uncertainty could support more informative segmentation or detection outputs. Such an extension should evaluate both calibration and task-specific decision quality rather than relying on ECE alone.

B Theoretical Analysis

This appendix establishes four local properties of RA-TPT: conditional rescaling by the coherence gate, bounded contribution of the hinge-margin term, decomposition of logit energy into centered and mean components, and local conflict attenuation by soft projection. These results characterize individual components of the update. They do not imply a general ECE bound, preserved top-1 accuracy, or convergence under repeated nonlinear adaptation.

B.1 Coherence Gate: Conditional Rescaling

The coherence factor $\mathcal{C}(\bar{\mathbf{p}})$ downweights the margin term when augmented views disagree. Its effect is conditional on the observed views; it is not an unbiased estimator of an ungated gradient.

Proposition B.1 (Conditional gradient rescaling)**.** *Let $\bar{\mathbf{g}}$ denote a random gradient quantity and let $\mathcal{C} = \mathcal{C}(\bar{\mathbf{p}}) \in (0, 1]$ be the coherence factor. Conditional on $\mathcal{C}$,*

$$\operatorname{Var}(\mathcal{C}\bar{\mathbf{g}} \mid \mathcal{C}) = \mathcal{C}^2 \operatorname{Var}(\bar{\mathbf{g}} \mid \mathcal{C}), \qquad \mathbb{E}[\mathcal{C}\bar{\mathbf{g}} \mid \mathcal{C}] = \mathcal{C}\,\mathbb{E}[\bar{\mathbf{g}} \mid \mathcal{C}]. \tag{22}$$

Thus, conditional on the observed coherence value, $\mathcal{C}$ rescales the mean gradient and its variance.

Proof. Conditional on $\mathcal{C}$, the quantity $\mathcal{C}$ is a scalar constant. The two identities in (22) follow from the scaling properties of conditional expectation and variance. Since $\mathcal{C} = \exp(-\gamma\mathcal{V})$ with $\gamma > 0$,

$$\frac{d\mathcal{C}^2}{d\mathcal{V}} = -2\gamma \exp(-2\gamma\mathcal{V}) < 0, \tag{23}$$

so the variance scaling factor decreases monotonically as view disagreement $\mathcal{V}$ increases. □

Scope. The proposition describes conditional scalar rescaling. It does not establish an unbiasedness property, a reduction in total optimization variance, or an improvement in calibration.

B.2 Adaptive Margin: Bounded Hinge Contribution

The adaptive margin uses a hinge term whose direct contribution with respect to the top-two gap vanishes after the instance-dependent cap is reached.

Proof of Claim 3.1. For fixed ω and $\mathcal{C}(\bar{\mathbf{p}})$, define

$$\kappa = \mathcal{C}(\bar{\mathbf{p}})(1-\omega) \geq 0, \qquad m = m_{\mathrm{cap}}(\omega) \geq 0. \tag{24}$$

Equation (9) becomes $\mathcal{L}_{\mathrm{margin}} = \kappa[m - \Delta]_+$. Its subgradient with respect to Δ is

$$\frac{\partial \mathcal{L}_{\mathrm{margin}}}{\partial \Delta} \in \begin{cases} \{-\kappa\}, & \Delta < m, \\ [-\kappa, 0], & \Delta = m, \\ \{0\}, & \Delta > m. \end{cases} \tag{25}$$

Therefore, conditional on ω and $\mathcal{C}(\bar{\mathbf{p}})$, the hinge contribution has zero subgradient with respect to Δ whenever $\Delta > m_{\mathrm{cap}}(\omega)$. □

Scope. This property concerns $\mathcal{L}_{\mathrm{margin}}$ only. In particular, the pseudo-label anchor in $\mathcal{L}_{\mathrm{sep}}$ can still alter the logits after hinge saturation. Thus, the result does not establish a stationary point, a bound on the full logit gap, or a confidence guarantee for the complete update.

B.3 Mean–Dispersion Decomposition

The regularizer separates the centered component of the averaged logits from their common-mode component. Recall

$$\bar{z} = \frac{1}{K}\sum_{k=1}^{K} \bar{z}_k, \qquad \tilde{\mathbf{z}} = \bar{\mathbf{z}} - \bar{z}\mathbf{1}. \tag{26}$$

Proposition B.2 (Logit-energy decomposition)**.** *The squared energy of* $\bar{\mathbf{z}}$ *decomposes as*

$$\frac{1}{K}\|\bar{\mathbf{z}}\|_2^2 = \underbrace{\frac{1}{K}\|\tilde{\mathbf{z}}\|_2^2}_{\mathcal{L}_{\mathrm{disp}}} + \underbrace{\bar{z}^2}_{\mathcal{L}_{\mathrm{mean}}}. \tag{27}$$

Moreover, for any $c \in \mathbb{R}$*,*

$$\mathcal{L}_{\mathrm{disp}}(\bar{\mathbf{z}} + c\mathbf{1}) = \mathcal{L}_{\mathrm{disp}}(\bar{\mathbf{z}}), \qquad \mathrm{softmax}(\bar{\mathbf{z}} + c\mathbf{1}) = \mathrm{softmax}(\bar{\mathbf{z}}). \tag{28}$$

Proof. Since $\bar{\mathbf{z}} = \tilde{\mathbf{z}} + \bar{z}\mathbf{1}$ and $\mathbf{1}^\top\tilde{\mathbf{z}} = 0$,

$$\begin{aligned} \|\bar{\mathbf{z}}\|_2^2 &= \|\tilde{\mathbf{z}} + \bar{z}\mathbf{1}\|_2^2 \\ &= \|\tilde{\mathbf{z}}\|_2^2 + 2\bar{z}\,\mathbf{1}^\top\tilde{\mathbf{z}} + \bar{z}^2\|\mathbf{1}\|_2^2 \\ &= \|\tilde{\mathbf{z}}\|_2^2 + K\bar{z}^2. \end{aligned} \tag{29}$$

Dividing by K shows (27). Under the shift $\bar{\mathbf{z}} \mapsto \bar{\mathbf{z}} + c\mathbf{1}$, the scalar mean becomes $\bar{z} + c$, and hence

$$(\bar{\mathbf{z}} + c\mathbf{1}) - (\bar{z} + c)\mathbf{1} = \tilde{\mathbf{z}}. \tag{30}$$

The softmax identity follows by factoring $\exp(c)$ from both the numerator and denominator. □

Interpretation. Only $\mathcal{L}_{\mathrm{disp}}$ is shift-invariant and directly relevant to the predictive distribution under a common additive logit shift. $\mathcal{L}_{\mathrm{mean}}$ is included as a complementary optimization term; it is not used in the maximum-confidence result below and is not claimed to directly modify softmax confidence.

B.4 Centered-Logit Dispersion and Maximum Confidence

We first record the elementary properties of the quadratic routing weight.

Proposition B.3 (Quadratic-routing properties)**.** *For $r(\omega) = \omega^2$ and $\omega \in [0, 1]$,*

$$r(0) = 0, \qquad r(1) = 1, \qquad 0 < r(\omega) < \omega \quad \textit{for } \omega \in (0, 1). \tag{31}$$

Thus, compared with linear routing, quadratic routing assigns a smaller weight to $\mathcal{L}_{\mathrm{disp}}$ at intermediate values of ω, while reaching unit weight at $\omega = 1$.

Proof. The endpoint identities are immediate. For $\omega \in (0, 1)$, multiplication by $\omega < 1$ gives $\omega^2 < \omega$. □

Proof of Theorem 3.1. Let $\varsigma^2 = \mathcal{L}_{\mathrm{disp}}$. Samuelson's inequality (Samuelson, 1968) gives

$$\bar{z}_{(1)} - \bar{z}_{(K)} \le 2\varsigma\sqrt{K-1} \triangleq \Delta_{\max}. \tag{32}$$

For a maximizing index $k^* = \arg\max_k \bar{z}_k$, every other logit satisfies $\bar{z}_j \ge \bar{z}_{(1)} - \Delta_{\max}$. Therefore,

$$\sum_{j=1}^{K} \exp(\bar{z}_j) \ge \exp(\bar{z}_{(1)}) + (K-1)\exp(\bar{z}_{(1)} - \Delta_{\max}). \tag{33}$$

It follows that

$$\begin{aligned} p_{\max} &= \frac{\exp(\bar{z}_{(1)})}{\sum_{j=1}^{K} \exp(\bar{z}_j)} \\ &\le \frac{1}{1 + (K-1)\exp(-\Delta_{\max})} \\ &= \frac{1}{1 + (K-1)\exp(-2\varsigma\sqrt{K-1})}. \end{aligned} \tag{34}$$

The upper bound is increasing in ς and converges to $1/K$ as $\varsigma \to 0$. □

Scope. Theorem 3.1 is a pointwise bound on maximum softmax confidence for a given centered-logit dispersion. It does not guarantee that optimization attains a prescribed dispersion, nor does it bound ECE, which also depends on prediction correctness; the complementary $\mathcal{L}_{\mathrm{mean}}$ does not enter the result. We stress that the bound is numerically loose at large class counts: for $K \in [100, 1000]$ and $\varsigma = O(1)$ the right-hand side exceeds 0.99, so it does not constrain $p_{\max}$ in absolute terms at CLIP scale. A slightly tighter constant is available by bounding $\bar{z}_{(1)} - \bar{z}_j$ relative to the mean ($\bar{z}_{(1)} - \bar{z} \le \varsigma\sqrt{K-1}$) rather than via the full range $\Delta_{\max} = 2\varsigma\sqrt{K-1}$, but the qualitative conclusion is unchanged. We therefore use Theorem 3.1 as motivation for penalizing dispersion, not as a numerical certificate on confidence.

B.5 Gradient Projection: Local Conflict Attenuation

The projection modifies the regularization gradient $\mathbf{g}_r = \nabla_\Theta \mathcal{L}_{\mathrm{reg}}$ when it locally conflicts with the separability gradient $\mathbf{g}_s = \nabla_\Theta \mathcal{L}_{\mathrm{sep}}$. We first state the result for the idealized case $\epsilon = 0$; the implementation uses $\epsilon > 0$ for numerical stability.

Proposition B.4 (Residual local conflict)**.** *Let*

$$s = \langle \mathbf{g}_r, \mathbf{g}_s \rangle, \qquad \phi = \sigma(-\xi s), \qquad \mathbf{g}_r^* = \mathbf{g}_r - \phi \frac{s}{\|\mathbf{g}_s\|_2^2} \mathbf{g}_s. \tag{35}$$

Then

$$\langle \mathbf{g}_r^*, \mathbf{g}_s \rangle = s(1 - \phi). \tag{36}$$

For $s < 0$,

$$|\langle \mathbf{g}_r^*, \mathbf{g}_s \rangle| = \frac{|s|}{1 + \exp(\xi|s|)} < |s|. \tag{37}$$

Proof. By bilinearity,

$$\begin{aligned}\langle \mathbf{g}_r^*, \mathbf{g}_s \rangle &= \left\langle \mathbf{g}_r - \phi \frac{s}{\|\mathbf{g}_s\|_2^2} \mathbf{g}_s, \mathbf{g}_s \right\rangle \\ &= s - \phi s = s(1-\phi). \end{aligned} \tag{38}$$

For $s < 0$, $\phi = \sigma(\xi|s|)$, so

$$1 - \phi = \frac{1}{1 + \exp(\xi|s|)}.$$

Taking absolute values gives (37). □

Implementation form. With $\epsilon > 0$, as used in Equation (18), the exact residual inner product is

$$\langle \mathbf{g}_r^*, \mathbf{g}_s \rangle = s \left(1 - \phi \frac{\|\mathbf{g}_s\|_2^2}{\|\mathbf{g}_s\|_2^2 + \epsilon} \right). \tag{39}$$

For $s < 0$, the magnitude of this residual remains smaller than $|s|$, because

$$0 < 1 - \phi \frac{\|\mathbf{g}_s\|_2^2}{\|\mathbf{g}_s\|_2^2 + \epsilon} < 1.$$

As $\epsilon / \|\mathbf{g}_s\|_2^2 \to 0$, the expression approaches the idealized result in Proposition B.4.

Scope. The proposition concerns the local alignment of the two gradients. It does not establish improved accuracy or calibration, nor convergence under repeated adaptation.

B.6 Why Project $\mathbf{g}_r$ Rather Than $\mathbf{g}_s$?

RA-TPT preserves the separability gradient $\mathbf{g}_s$ and modifies only the regularization gradient $\mathbf{g}_r$. A reverse alternative projects $\mathbf{g}_s$ as

$$\mathbf{g}_s^* = \mathbf{g}_s - \sigma(-\xi \langle \mathbf{g}_s, \mathbf{g}_r \rangle) \frac{\langle \mathbf{g}_s, \mathbf{g}_r \rangle}{\|\mathbf{g}_r\|_2^2 + \epsilon} \mathbf{g}_r. \tag{40}$$

Under local conflict, this modifies the self-training direction itself. Neither choice provides a general guarantee of superior classification accuracy or calibration; their empirical trade-off is reported in Table 7.

B.7 Embedding Geometry and Orthogonality Constraints

We record a limited property of exact orthogonality. It does not imply that finite-strength orthogonality regularization necessarily degrades recognition or calibration in practice.

Proposition B.5 (Exact orthogonality changes pairwise similarities)**.** *Let $\mathbf{W} \in \mathbb{R}^{K \times d}$ contain unit-normalized class embeddings as rows. If*

$$\|\mathbf{W}\mathbf{W}^\top - \mathbf{I}_K\|_F^2 \to 0, \tag{41}$$

then

$$\mathbf{w}_i^\top \mathbf{w}_j \to 0 \qquad \text{for all } i \neq j. \tag{42}$$

Proof. For $i \neq j$, the (i,j)-th entry of $\mathbf{W}\mathbf{W}^\top - \mathbf{I}_K$ is $\mathbf{w}_i^\top \mathbf{w}_j$. Since convergence of the squared Frobenius norm to zero implies entrywise convergence to zero, every off-diagonal inner product converges to zero. □

Interpretation. Exact orthogonality changes pairwise text-feature similarities, including between semantically related classes. Its practical effect depends on the data, regularization strength, optimization trajectory, and target shift. Accordingly, we treat its impact as empirical and report fine-grained and domain-specific comparisons in Appendix R.2.

C Experimental Tables

C.1 Detailed Analysis of CLIP-RN50 Results

Table 9 provides the comprehensive evaluation of accuracy and Expected Calibration Error (ECE) across all 14 datasets using the ResNet-50 backbone. While the main text summarizes these findings, a granular dataset-level analysis reveals the specific failure modes of baseline methods and the structural advantages of Reliability-Aware Test-Time Prompt Tuning (RA-TPT).

The Pathology of Standard Entropy Minimization. The zero-shot CLIP baseline exhibits relatively strong calibration (average ECE of 5.70 on standard datasets and 7.43 on variants) but suffers from low target-domain accuracy. Standard TPT aggressively recovers this accuracy (improving from 55.6% to 57.5% on standard datasets) but triggers severe miscalibration. This is starkly visible on fine-grained and texture-heavy datasets: on DTD, TPT's ECE explodes from 9.82 to 25.70, and on EuroSAT (SAT), it degrades from 15.07 to 22.68. Under severe distribution shifts (ImageNet-A), TPT reaches an ECE of 31.04. This empirically validates the theoretical diagnosis in the main text: unbounded entropy minimization induces runaway logit inflation, forcing the model into extreme overconfidence regardless of the actual prediction correctness. Similar behavior is observed in R-TPT, which attempts to filter adversarial gradients but fails to constrain the absolute logit scale, resulting in an equally poor average ECE of 16.57 on ImageNet variants.

Limitations of Prior Calibration-Aware Methods. Recent methods mitigate this by constraining the textual embedding matrix $\mathbf{W}$ (C-TPT, O-TPT) or by reshaping confidence (the concurrent SoC). SoC is the strongest baseline on RN50, lowering the average standard-dataset ECE to 5.34 and the variant ECE to 6.56, with O-TPT at 5.82 and 8.93. All remain inflated under heavy distribution shift, however: on ImageNet-A the ECE is 24.25 for O-TPT and 19.60 for SoC, and on DTD it is 16.45 and 14.59 respectively. Because these methods apply uniform geometric or global confidence pressure without evaluating per-sample uncertainty, they cannot selectively suppress confidence on genuinely ambiguous out-of-distribution inputs.

The RA-TPT Advantage. RA-TPT achieves a strong accuracy-calibration trade-off by operating entirely in the 1D logit space via uncertainty-routed logit regularization.

- *Standard datasets:* RA-TPT attains the lowest average ECE (5.07), ahead of the strongest baseline SoC (5.34), while keeping competitive accuracy (56.8%). It sharply reduces the miscalibration on the hardest datasets, cutting DTD from 16.45 (O-TPT) to 8.74 and EuroSAT from 15.59 to 10.94.
- *Distribution shifts:* On the ImageNet variants RA-TPT reaches the lowest average ECE (6.34), narrowly ahead of SoC (6.56) and well below O-TPT (8.93) and TPT (16.75). On ImageNet-Sketch it obtains an ECE of 1.18 at 35.1% accuracy, comparable to the best-performing tuning methods; on ImageNet-A it reduces ECE to 18.66, the lowest of all methods (SoC 19.60, O-TPT 24.25, TPT 31.04).

These results indicate that dynamically bounding logit separation by instance difficulty (ω), rather than rigidly forcing orthogonal text embeddings, yields a better accuracy-calibration trade-off and the lowest worst-case miscalibration, without sacrificing the discriminative gains of test-time adaptation.

C.2 Detailed Analysis of BioMedCLIP Results

Table 10 provides the full per-dataset evaluation on 11 specialized medical imaging datasets using the BioMedCLIP (Zhang et al., 2025) foundation model. Medical imaging presents a unique challenge for test-time adaptation because the visual features are highly fine-grained and visually homogeneous compared to natural images. This environment serves as a severe stress test for adaptation reliability.

The Failure of Entropy Minimization on Medical Data. Unlike standard visual benchmarks where TPT generally improves raw classification accuracy at the expense of calibration, medical domains trigger a complete collapse of the entropy-based paradigm. Standard TPT degrades the average zero-shot accuracy (dropping from 35.06% to 33.02%) while simultaneously causing ECE to skyrocket from 40.48 to 59.45. By indiscriminately minimizing entropy on unfamiliar medical textures, TPT confidently locks onto spurious features or noise, severely degrading both predictive performance and structural reliability.

Table 9: **Calibration (ECE ↓) and Accuracy (Acc. ↑)** with **CLIP-RN50** across ten standard datasets and four ImageNet variants. **Bold** = best average values compared to second best baseline; underline = second best; shaded = our method. Standard deviation is reported in D.

Method	Metric	Ten Standard Datasets										Avg.	ImageNet Variants				Avg.
		INet	DTD	FLW	Food	Air.	Pets	C101	UCF	SAT	Car		I-A	I-V2	I-R	I-S	
Zero-Shot	Acc.	58.1	39.9	60.9	73.8	15.6	83.7	85.9	58.4	24.2	55.7	55.6	21.7	51.4	55.9	33.3	40.6
	ECE	2.12	9.82	3.21	2.49	6.34	6.02	4.22	3.06	15.07	4.63	5.70	21.30	3.35	1.95	3.14	7.43
TPT	Acc.	60.3	41.4	62.7	75.0	17.7	84.8	86.9	59.7	28.4	58.5	57.5	25.2	54.6	59.1	35.3	43.5
	ECE	11.40	25.70	13.43	5.27	15.33	3.58	4.90	12.08	22.68	3.56	11.79	31.04	13.18	9.12	13.67	16.75
R-TPT	Acc.	58.1	39.6	61.1	75.7	15.5	83.6	85.5	58.2	24.1	55.4	55.7	21.6	51.5	56.0	33.3	40.6
	ECE	11.56	24.79	12.32	5.21	12.35	4.04	4.94	9.76	30.72	1.38	11.71	29.98	13.54	9.79	12.99	16.57
C-TPT	Acc.	59.9	42.1	65.1	74.4	16.8	83.6	86.6	59.4	27.6	56.1	57.2	22.2	53.4	56.9	34.3	41.7
	ECE	3.06	20.84	4.07	1.92	11.11	2.71	2.21	4.09	15.27	2.00	6.73	22.78	5.09	1.34	6.46	8.92
A-TPT	Acc.	58.2	40.7	61.2	72.4	15.7	83.3	84.5	57.7	22.6	54.3	55.1	23.7	54.5	58.4	35.1	42.9
	ECE	2.46	6.05	6.02	7.46	7.47	7.20	7.09	3.10	16.02	12.16	7.50	27.65	12.32	6.49	13.82	15.07
O-TPT	Acc.	58.5	41.7	65.7	74.5	16.9	83.2	86.7	58.9	28.0	55.3	56.9	22.6	53.2	57.1	34.1	41.7
	ECE	3.14	16.45	2.61	1.20	8.33	3.27	3.28	2.43	15.59	1.92	5.82	24.25	3.82	2.58	5.06	8.93
SoC	Acc.	58.9	41.9	65.7	74.6	16.9	83.0	86.5	59.6	26.3	55.6	56.9	21.5	52.4	56.2	33.2	40.8
	ECE	3.19	14.59	2.20	1.31	7.78	3.21	2.61	2.56	13.69	2.30	5.34	19.60	2.48	1.39	2.77	6.56
RA-TPT	**Acc.**	58.8	41.6	65.3	74.3	16.9	84.2	85.7	58.5	26.1	56.4	56.8	24.1	52.0	57.0	35.1	**42.1**
(Ours)	**ECE**	2.26	8.74	2.49	2.93	6.42	5.28	3.90	2.16	10.94	5.21	**5.07**	18.66	3.01	2.52	1.18	**6.34**

Table 10: **Calibration (ECE ↓) and Accuracy (Acc. ↑) on Medical Domains.** Evaluation using the BioMedCLIP foundation model across 11 specialized medical imaging datasets. Bold values indicate the best average performance.

Method	Metric	BTMRI	BUSI	CHMNIST	COVID19	CTKidney	Derma.	KneeXray	Kvasir	LungCol.	OCT.	RETINA	Avg.
Zero-Shot	Acc.	16.13	59.74	30.65	27.32	42.40	31.02	26.02	54.58	50.08	30.20	17.50	35.06
	ECE	72.56	21.17	42.60	48.50	29.51	31.62	30.97	30.73	20.62	62.65	54.33	40.48
TPT	Acc.	14.79	45.34	34.31	30.47	45.99	18.30	19.20	54.33	55.64	27.30	17.59	33.02
	ECE	81.88	52.85	57.83	66.55	46.17	65.98	73.25	39.59	33.82	69.41	66.65	59.45
C-TPT	Acc.	15.02	46.18	32.78	30.45	46.25	18.90	19.44	54.08	55.61	29.40	18.45	33.32
	ECE	80.71	51.20	58.69	66.48	45.94	64.97	71.08	39.63	33.74	64.91	65.36	58.43
O-TPT	Acc.	14.97	46.61	32.11	30.47	46.20	19.15	20.23	54.00	55.57	29.90	19.00	33.47
	ECE	80.59	49.69	59.07	66.43	45.95	64.08	70.30	39.37	33.67	64.24	64.24	57.97
A-TPT	Acc.	14.61	49.57	30.64	30.85	32.10	36.05	26.63	53.33	46.25	28.60	20.82	33.59
	ECE	76.08	37.63	54.39	53.44	46.34	27.34	48.62	33.12	25.92	66.43	62.03	48.30
SoC	Acc.	15.61	55.51	31.54	26.25	45.80	28.18	26.27	54.33	48.41	29.70	18.14	34.52
	ECE	73.30	23.67	57.45	49.72	25.86	34.90	34.57	31.00	21.39	63.65	53.58	42.64
RA-TPT	Acc.	16.21	60.17	31.60	28.61	44.50	28.30	26.81	57.00	48.30	29.00	18.62	**35.37**
	ECE	72.77	21.47	39.87	45.74	29.69	28.86	28.21	27.97	17.86	59.89	45.47	**37.98**

Insufficiency of Prior Adaptation Methods. Geometric regularizers (O-TPT, A-TPT) enforce orthogonality or angular diversity on the embedding matrix, but forcing uniform separation between medical classes that share strong semantic overlap distorts the pretrained medical knowledge: O-TPT and A-TPT fail to recover the zero-shot accuracy (33.47% and 33.59%) and remain heavily miscalibrated (57.97 and 48.30 ECE). The concurrent SoC is the strongest baseline, staying closest to zero-shot (34.52% accuracy, 42.64 ECE), yet it too degrades both metrics relative to zero-shot (35.06%, 40.48) rather than improving them.

RA-TPT on Medical Data. RA-TPT is the only evaluated method that improves over zero-shot on both metrics in this domain. By operating in logit space with an instance-adaptive margin, it avoids the embedding distortion that harms the geometric baselines, raising average accuracy to 35.37% (above the 35.06% zero-shot) while lowering average ECE from 40.48 to 37.98. The gain is largest on RETINA, where ECE falls from 54.33 to 45.47, versus 66.65 for TPT. We stress that this is a relative improvement, not a solution: at 37.98 average ECE all methods, including RA-TPT, remain severely miscalibrated in absolute terms, and on several datasets (e.g., BTMRI, BUSI, CTKidney) RA-TPT does not improve over zero-shot. Reliable test-time calibration on specialized medical foundation models thus remains an open problem; RA-TPT is best read as the only method here that does not make zero-shot reliability worse, not as a deployable safety guarantee.

Table 11: **Calibration (ECE ↓) and Accuracy (Acc. ↑) on Remote Sensing Domains.** Evaluation using the RemoteCLIP foundation model across 7 specialized Earth Observation datasets. Bold indicates the best average ECE.

Method	Metric	MLRSNet	PatternNet	RESISC45	AID	UCM	EuroSAT	RSICD	Avg.
Zero-Shot	Acc.	41.81	43.91	48.63	48.62	49.80	33.96	48.05	44.97
	ECE	2.37	3.76	5.10	1.65	3.95	3.42	3.27	3.36
TPT	Acc.	42.12	46.22	51.22	48.72	50.20	36.10	46.31	45.84
	ECE	16.55	16.88	13.10	7.51	17.86	8.68	18.15	15.53
C-TPT	Acc.	41.77	42.26	49.54	46.54	49.45	33.26	44.51	43.90
	ECE	2.66	4.82	2.03	7.60	3.96	2.86	9.25	4.74
O-TPT	Acc.	41.60	42.31	48.91	46.14	49.75	33.37	43.98	43.72
	ECE	1.36	3.70	0.78	6.26	3.85	3.14	8.23	3.90
A-TPT	Acc.	41.05	46.00	49.90	47.61	49.45	32.64	46.98	44.80
	ECE	11.24	11.59	15.80	3.74	5.89	3.74	3.15	7.88
SoC	Acc.	41.25	44.79	50.77	47.47	48.75	31.76	47.02	44.54
	ECE	2.34	3.73	6.01	2.52	5.74	4.78	3.25	4.05
RA-TPT (Ours)	Acc.	42.40	45.00	48.20	49.20	50.70	33.32	48.10	45.27
	ECE	3.04	2.81	4.92	2.24	3.84	2.95	2.39	**3.17**

C.3 Detailed Analysis of RemoteCLIP Results

Table 11 details the adaptation performance on 7 specialized Earth Observation (EO) datasets using the domain-specific RemoteCLIP (Liu et al., 2024) foundation model. Remote sensing imagery introduces distinct challenges for vision-language models: images are captured from a top-down perspective, lack canonical object orientation, exhibit extreme scale variations, and feature dense, cluttered backgrounds (e.g., distinguishing between visually similar "dense residential" and "medium residential" zones).

The Hazard of Unbounded Entropy in Earth Observation. The zero-shot RemoteCLIP model establishes a robust baseline, demonstrating particularly strong inherent calibration (average ECE of 3.36). However, applying standard TPT to this domain reveals a dangerous pathology. While TPT successfully elevates the average accuracy to 45.84%, it degrades sharply the model's calibration, sending the average ECE skyrocketing to 15.53. On datasets like RSICD and UCM, the ECE reaches approximately 18%. By indiscriminately minimizing entropy over highly cluttered satellite imagery, TPT forces the model into unwarranted overconfidence based on spurious background textures. In safety-critical EO applications such as disaster response or automated land-use mapping, this extreme miscalibration renders the model's predictions untrustworthy.

Partial Fixes from Prior Methods. Methods that regularize the embedding space (O-TPT, C-TPT) constrain overconfidence but degrade discrimination: by rigidly enforcing orthogonality across conceptually overlapping geographic classes, both O-TPT (43.72%) and C-TPT (43.90%) fall below the zero-shot accuracy (44.97%). The concurrent SoC preserves calibration well (4.05 ECE) but also slips below zero-shot accuracy (44.54%), so no baseline improves on the foundation model in both accuracy and calibration.

RA-TPT on Remote Sensing. RA-TPT is the only method that improves on the foundation model in both metrics on average: it raises accuracy to 45.27% (above the 44.97% zero-shot) and lowers the average ECE to 3.17, the best of all evaluated methods (O-TPT 3.90, SoC 4.05). Unlike TPT, which buys a small accuracy gain (45.84%) at the cost of a large calibration collapse (15.53 ECE), RA-TPT retains the strong zero-shot calibration while still adapting, making it a reliable choice for high-clutter Earth-observation settings, though the accuracy gain over zero-shot is modest.

D STANDARD DEVIATION

D.1 ANALYSIS OF PERFORMANCE STABILITY AND VARIANCE

Test-time adaptation relies on stochastic view generation, so stability across runs matters for reliable deployment. Tables 12 and 13 report the per-seed standard deviation of Accuracy and ECE over three seeds; the "Avg." columns are the mean of the per-dataset deviations. Averaged over the ten standard datasets, the entropy-based methods are the least stable: TPT reaches an ECE deviation of ± 0.28 on RN50 and ± 0.19 on ViT-B/16, with per-dataset values as large as ± 0.74 (RN50, Aircraft) and ± 0.48 under ImageNet-A. RA-TPT is substantially more stable than these entropy methods (± 0.14 on RN50, ± 0.12 on ViT-B/16) and comparable to the calibration-aware baselines (C-TPT, O-TPT, SoC). We attribute this reduced sensitivity to stochastic augmentation noise, relative to unbounded entropy minimization, to the Augmentation Coherence Gate $\mathcal{C}$, which dampens the margin gradient when augmented views disagree ($\mathcal{V}(\bar{\mathbf{p}})$ large); we do not claim uniformly lowest variance across methods.

Table 12: **Performance Stability** ($\pm$ Std. Dev.) across three random seeds for standard datasets and ImageNet variants in CLIP ViT-B/16. "Avg." columns report the mean of the per-dataset standard deviations.

Method	Metric	Ten Standard Datasets											ImageNet Shifts				
		INet	DTD	FLW	Food	Air.	Pets	C101	UCF	SAT	Car	Avg.	I-A	I-V2	I-R	I-S	Avg.
Baseline	Acc.	±.03	±.05	±.03	±.05	±.02	±.04	±.01	±.02	±.02	±.03	±.03	±.08	±.05	±.06	±.07	±.06
	ECE	±.03	±.02	±.06	±.02	±.02	±.01	±.01	±.06	±.04	±.03	±.03	±.12	±.06	±.05	±.07	±.08
TPT	Acc.	±.07	±.24	±.10	±.02	±.26	±.02	±.07	±.25	±.25	±.21	±.15	±.35	±.28	±.18	±.30	±.28
	ECE	±.04	±.34	±.29	±.02	±.23	±.05	±.16	±.22	±.33	±.20	±.19	±.42	±.31	±.22	±.38	±.33
R-TPT	Acc.	±.20	±.04	±.10	±.18	±.12	±.08	±.39	±.27	±.12	±.06	±.16	±.30	±.26	±.20	±.28	±.26
	ECE	±.03	±.13	±.32	±.11	±.15	±.08	±.08	±.13	±.33	±.07	±.14	±.38	±.29	±.24	±.34	±.31
C-TPT	Acc.	±.10	±.42	±.24	±.14	±.36	±.12	±.18	±.14	±.08	±.16	±.19	±.18	±.14	±.12	±.16	±.15
	ECE	±.02	±.14	±.14	±.18	±.10	±.05	±.07	±.04	±.11	±.07	±.09	±.20	±.15	±.10	±.18	±.16
A-TPT	Acc.	±.22	±.30	±.11	±.03	±.16	±.04	±.03	±.05	±.13	±.07	±.11	±.22	±.18	±.14	±.20	±.18
	ECE	±.02	±.32	±.10	±.02	±.06	±.07	±.05	±.32	±.21	±.12	±.13	±.28	±.22	±.18	±.26	±.24
O-TPT	Acc.	±.98	±.13	±.02	±.05	±.17	±.44	±.03	±.08	±.26	±.05	±.22	±.20	±.16	±.13	±.18	±.17
	ECE	±.08	±.15	±.22	±.03	±.12	±.06	±.03	±.02	±.23	±.03	±.10	±.24	±.18	±.14	±.22	±.20
SoC	Acc.	±.25	±.18	±.08	±.04	±.16	±.12	±.05	±.06	±.18	±.06	±.12	±.21	±.17	±.13	±.19	±.18
	ECE	±.05	±.21	±.15	±.02	±.09	±.06	±.04	±.12	±.22	±.08	±.10	±.26	±.20	±.16	±.24	±.22
RA-TPT	**Acc.**	±.12	±.32	±.22	±.06	±.15	±.06	±.06	±.14	±.33	±.05	±.15	±.16	±.12	±.11	±.14	±.13
	ECE	±.02	±.24	±.12	±.04	±.20	±.17	±.05	±.01	±.21	±.12	±.12	±.18	±.14	±.12	±.15	±.15

Statistical significance and worst-case reliability. Because several per-dataset ECE gaps are comparable in scale to these seed deviations, we assess significance against the strongest concurrent baseline, SoC, with a Wilcoxon signed-rank test over all 14 datasets. Calibration is statistically indistinguishable (ECE: $p = 0.86$; SoC is lower on 8 of 14 datasets, RA-TPT on 6), so we do not claim a per-dataset calibration advantage over SoC. RA-TPT's advantage is instead twofold and significant. First, it is Pareto-superior: at matched calibration it attains significantly higher accuracy

Table 13: **Performance Stability** ($\pm$ Std. Dev.) across three random seeds for standard datasets and ImageNet variants in CLIP RN50. "Avg." columns report the mean of the per-dataset standard deviations.

Method	Metric	Ten Standard Datasets											ImageNet Shifts				
		INet	DTD	FLW	Food	Air.	Pets	C101	UCF	SAT	Car	Avg.	I-A	I-V2	I-R	I-S	Avg.
Baseline	Acc.	$\pm.08$	$\pm.04$	$\pm.13$	$\pm.06$	$\pm.15$	$\pm.06$	$\pm.02$	$\pm.02$	$\pm.02$	$\pm.09$	$\pm.07$	$\pm.10$	$\pm.06$	$\pm.07$	$\pm.08$	$\pm.08$
	ECE	$\pm.04$	$\pm.06$	$\pm.13$	$\pm.02$	$\pm.03$	$\pm.11$	$\pm.01$	$\pm.03$	$\pm.05$	$\pm.07$	$\pm.06$	$\pm.14$	$\pm.08$	$\pm.06$	$\pm.09$	$\pm.09$
TPT	Acc.	$\pm.29$	$\pm.19$	$\pm.53$	$\pm.13$	$\pm.72$	$\pm.12$	$\pm.37$	$\pm.15$	$\pm.29$	$\pm.28$	$\pm.31$	$\pm.40$	$\pm.32$	$\pm.25$	$\pm.35$	$\pm.33$
	ECE	$\pm.14$	$\pm.23$	$\pm.57$	$\pm.12$	$\pm.74$	$\pm.04$	$\pm.42$	$\pm.25$	$\pm.25$	$\pm.05$	$\pm.28$	$\pm.48$	$\pm.36$	$\pm.30$	$\pm.42$	$\pm.39$
R-TPT	Acc.	$\pm.10$	$\pm.26$	$\pm.10$	$\pm.12$	$\pm.16$	$\pm.14$	$\pm.30$	$\pm.13$	$\pm.07$	$\pm.25$	$\pm.16$	$\pm.34$	$\pm.28$	$\pm.22$	$\pm.30$	$\pm.29$
	ECE	$\pm.10$	$\pm.21$	$\pm.11$	$\pm.10$	$\pm.14$	$\pm.08$	$\pm.06$	$\pm.11$	$\pm.43$	$\pm.05$	$\pm.14$	$\pm.42$	$\pm.34$	$\pm.28$	$\pm.38$	$\pm.36$
C-TPT	Acc.	$\pm.11$	$\pm.11$	$\pm.14$	$\pm.34$	$\pm.20$	$\pm.11$	$\pm.10$	$\pm.17$	$\pm.06$	$\pm.10$	$\pm.14$	$\pm.20$	$\pm.16$	$\pm.14$	$\pm.18$	$\pm.17$
	ECE	$\pm.04$	$\pm.09$	$\pm.04$	$\pm.06$	$\pm.15$	$\pm.05$	$\pm.07$	$\pm.04$	$\pm.08$	$\pm.05$	$\pm.07$	$\pm.22$	$\pm.18$	$\pm.14$	$\pm.20$	$\pm.18$
A-TPT	Acc.	$\pm.23$	$\pm.03$	$\pm.06$	$\pm.02$	$\pm.27$	$\pm.11$	$\pm.18$	$\pm.41$	$\pm.03$	$\pm.04$	$\pm.14$	$\pm.24$	$\pm.20$	$\pm.16$	$\pm.22$	$\pm.20$
	ECE	$\pm.11$	$\pm.65$	$\pm.24$	$\pm.02$	$\pm.25$	$\pm.10$	$\pm.18$	$\pm.19$	$\pm.01$	$\pm.05$	$\pm.18$	$\pm.30$	$\pm.26$	$\pm.20$	$\pm.28$	$\pm.26$
O-TPT	Acc.	$\pm.41$	$\pm.25$	$\pm.06$	$\pm.06$	$\pm.08$	$\pm.13$	$\pm.00$	$\pm.13$	$\pm.25$	$\pm.04$	$\pm.14$	$\pm.22$	$\pm.18$	$\pm.15$	$\pm.20$	$\pm.19$
	ECE	$\pm.07$	$\pm.28$	$\pm.15$	$\pm.07$	$\pm.08$	$\pm.08$	$\pm.20$	$\pm.26$	$\pm.21$	$\pm.03$	$\pm.14$	$\pm.26$	$\pm.20$	$\pm.16$	$\pm.24$	$\pm.22$
SoC	Acc.	$\pm.30$	$\pm.15$	$\pm.06$	$\pm.04$	$\pm.18$	$\pm.12$	$\pm.08$	$\pm.22$	$\pm.15$	$\pm.04$	$\pm.13$	$\pm.23$	$\pm.19$	$\pm.15$	$\pm.21$	$\pm.20$
	ECE	$\pm.09$	$\pm.35$	$\pm.18$	$\pm.05$	$\pm.16$	$\pm.09$	$\pm.19$	$\pm.22$	$\pm.11$	$\pm.04$	$\pm.15$	$\pm.28$	$\pm.23$	$\pm.18$	$\pm.26$	$\pm.24$
RA-TPT	**Acc.**	$\pm.10$	$\pm.25$	$\pm.13$	$\pm.05$	$\pm.41$	$\pm.11$	$\pm.29$	$\pm.08$	$\pm.04$	$\pm.08$	$\pm.15$	$\pm.18$	$\pm.14$	$\pm.12$	$\pm.16$	$\pm.15$
	ECE	$\pm.04$	$\pm.05$	$\pm.28$	$\pm.05$	$\pm.15$	$\pm.26$	$\pm.24$	$\pm.02$	$\pm.11$	$\pm.16$	$\pm.14$	$\pm.20$	$\pm.16$	$\pm.14$	$\pm.18$	$\pm.17$

(Acc.: $p < 0.01$, higher on 11 of 14 datasets) and improves both average accuracy and average ECE over SoC (Table 2). Second, it is the most reliable in the worst case, with the lowest maximum per-dataset ECE of any method (7.87, versus 11.17 for SoC, 13.72 for O-TPT, and 21.42 for TPT); RA-TPT never exhibits the severe single-dataset miscalibration (e.g., SoC's 11.17 on EuroSAT) that inflates competing averages.

E Uncertainty-Routed Regularization versus Simple Alternatives

RA-TPT's uncertainty-routed regularizer ($\mathcal{L}_{\text{reg}}$) constrains predictive concentration, which raises a natural question: could simpler, well-established techniques achieve the same effect? In this section we replace our uncertainty-routed formulation with three naive calibration alternatives: Unsupervised Temperature Scaling, Logit Clipping, and Uniform Norm Regularization.

To isolate the impact of the uncertainty-routed regularizer, we keep $\mathcal{L}_{\text{sep}}$ (the adaptive margin) and the soft gradient projection fixed across all variants in this ablation, swapping only the confidence regularizer. This shows that sample-agnostic penalties fail in the unsupervised test-time adaptation (TTA) regime: they either penalize the model uniformly, regardless of instance difficulty, or they destroy the gradient signal needed for adaptation.

E.1 Theoretical Limitations of Naive Alternatives

1. Unsupervised Temperature Scaling. Standard temperature scaling (Guo et al., 2017) requires a labeled validation set to optimize the temperature T. In the unsupervised TTA setting, one alternative is to fix a higher temperature ($T > \tau_{\text{CLIP}}$) to flatten the logits. *Why it fails:* temperature scaling is a post-hoc flattening of the softmax surface and does not change the gradient dynamics of the prompt tokens. With entropy minimization still active, the optimizer pushes the pre-temperature logits $\bar{\mathbf{z}}$ further toward infinity to compensate for the larger T. The model stays overconfident and the prompt geometry still diverges.

2. Logit Clipping (Clamping). A seemingly direct way to prevent logit inflation is to hard-clamp the logits to a bound $[-C, C]$ before the softmax. *Why it fails (gradient starvation):* once a confident in-distribution (ID) sample reaches the threshold C, its gradient becomes exactly zero, halting adaptation for that sample and preventing the prompt from learning fine-grained, beneficial features. Clipping

is also a static threshold: it applies the same bound to an easy ID image and to a highly ambiguous out-of-distribution (OOD) image, violating the requirement for instance-adaptive calibration.

3. Uniform Norm Regularization (Weight Decay on Logits). Instead of RA-TPT's uncertainty-routed regularizer, one could apply a uniform L_2 penalty to the logit vector for every sample: $\mathcal{L}_{\text{norm}} = \lambda \|\bar{\mathbf{z}}\|^2$. *Why it fails (ID accuracy collapse):* this is equivalent to applying our mean-logit stabilizer $\mathcal{L}_{\text{mean}}$ without the uncertainty gate, pulling all logits toward zero indiscriminately. For confident, correct ID samples this destroys discriminative power and degrades top-1 accuracy. RA-TPT avoids this by routing the dispersion penalty $\mathcal{L}_{\text{disp}}$ and the mean-logit stabilizer $\mathcal{L}_{\text{mean}}$ through the uncertainty gate ω, so that the magnitude penalty acts only when the model exhibits genuine ambiguity.

E.2 Empirical Validation

Table 14 evaluates these alternatives against RA-TPT with the ViT-B/16 backbone. We report the average across 10 datasets (the 9 standard datasets of Table 2 excluding ImageNet, plus ImageNet-V2) alongside three diverse domains: ImageNet-V2 (natural shift), DTD (texture shift), and EuroSAT (severe satellite-domain shift).

As predicted, the brute-force alternatives can occasionally edge out small gains on isolated datasets (e.g., Uniform Norm on DTD accuracy), but fail elsewhere. Unsupervised Temperature Scaling and Logit Clipping both suffer large calibration explosions on EuroSAT and DTD, exceeding 17% and 18% ECE respectively, confirming that they cannot stop logit inflation under severe domain shift. Uniform Norm Regularization instead crushes logit magnitudes indiscriminately, dropping EuroSAT accuracy from 44.10% to 39.46%. RA-TPT is the only formulation that improves both metrics jointly, reaching the highest average accuracy (64.00%) and the lowest average ECE (4.10%) across the evaluated domains.

Table 14: **Ablation of simple alternatives** (ViT-B/16). We compare RA-TPT's uncertainty-routed regularizer $\mathcal{L}_{\text{reg}}$ against three naive calibration alternatives while keeping the adaptive margin $\mathcal{L}_{\text{sep}}$ and the soft projection active. RA-TPT achieves the best balance, with the highest average accuracy and lowest average ECE across the 10 evaluated datasets.

Method Formulation	Average (10 Datasets)		ImageNet-V2		DTD		EuroSAT	
	Acc. (↑)	ECE (↓)	Acc. (↑)	ECE (↓)	Acc. (↑)	ECE (↓)	Acc. (↑)	ECE (↓)
Unsupervised Temp Scaling ($T = 2.0$)	63.37	8.63	**61.76**	9.33	46.58	18.81	40.98	17.17
Logit Clipping ($C = 5.0$)	63.91	8.73	62.02	9.73	**46.75**	18.23	41.08	17.07
Uniform Norm Reg ($\lambda\|\bar{\mathbf{z}}\|^2$)	63.32	4.91	60.91	4.07	45.08	9.58	39.46	9.41
RA-TPT ($\mathcal{L}_{\text{reg}}$)	**64.00**	**4.10**	61.20	**3.23**	45.83	**7.55**	**44.10**	**4.12**

F Generality to Non-VLM Test-Time Adaptation

Because RA-TPT's objective acts on the logit vector $\bar{\mathbf{z}}$ rather than on prompts, the same losses apply to any classifier by choosing which parameters they optimize. On a standard convolutional network we adapt the batch-normalization affine parameters at test time, keeping $\mathcal{L}_{\text{sep}}$, $\mathcal{L}_{\text{reg}}$, and the soft projection identical to the VLM setting.

Setup. We evaluate on CIFAR-10-C, CIFAR-100-C, and ImageNet-C (Hendrycks & Dietterich, 2019) at severity 5 (15 corruptions), using WideResNet-28-10 (Zagoruyko & Komodakis, 2016) for CIFAR and ResNet-50 (He et al., 2016) for ImageNet, obtained off-the-shelf via RobustBench (Croce et al., 2021). Baselines are Tent (Wang et al., 2021) (entropy minimization over BN parameters) and CoTTA (Wang et al., 2022) (a mean-teacher method with augmentation-averaged pseudo-labels and stochastic restore); RA-TPT adapts the same BN parameters. Adaptation is episodic per corruption (the model is reset to source before each corruption). Tent minimizes BN-affine entropy with its standard optimizer (Adam on CIFAR; SGD-momentum, lr 2.5×10^{-4}, on ImageNet, as Adam is unstable for entropy minimization on the 1000-class head); CoTTA and RA-TPT use AdamW at lr 10^{-4} (CoTTA is within 0.2 ECE of its recommended 10^{-3}). The CLIP-tuned margin cap ($\alpha = 4$) does not transfer to these backbones, whose logit scale differs, so we set it by the gap rule $\alpha = k \cdot g$: g

Table 15: Generality on non-VLM corruption benchmarks (severity 5). Accuracy (Acc. ↑) and ECE (↓); optimizers and learning rates are given in the text. RA-TPT matches CoTTA at comparable accuracy while being a single-model method.

	CIFAR-10-C		CIFAR-100-C		ImageNet-C	
Method	Acc.	ECE	Acc.	ECE	Acc.	ECE
Source	56.48	34.00	34.20	35.54	18.08	15.60
Tent	75.05	23.01	54.29	34.55	37.14	10.70
CoTTA	79.57	**13.76**	55.49	**16.22**	31.56	3.07
RA-TPT	79.55	13.77	56.18	16.63	31.61	**3.06**

Table 16: Fair temperature scaling (one temperature fit on held-out in-distribution/validation data, applied unchanged across all corruptions). ECE (↓), severity 5.

	CIFAR-10-C		CIFAR-100-C		ImageNet-C	
Method	raw	+TS	raw	+TS	raw	+TS
Tent	23.0	16.3	34.5	26.5	10.70	6.42
CoTTA	13.8	**4.55**	16.2	6.42	3.07	5.51
RA-TPT	13.8	**4.3**	16.6	**6.1**	3.06	6.3

is the source model's mean top-2 logit gap (computed without labels) and k is a single scale constant, selected once by minimizing ECE on a small held-out set of validation corruptions disjoint from the reported 15. This gives $k = 7$, fixed across all datasets ($\alpha = 40.6,\ 16.8,\ 3.99$ on CIFAR-10-C, CIFAR-100-C, ImageNet-C, reflecting their different logit scales).

Results. RA-TPT matches CoTTA on ImageNet-C (3.06 vs. 3.07 ECE) and CIFAR-10-C (13.77 vs. 13.76), and is within 0.5 ECE on CIFAR-100-C (Table 15), at comparable accuracy and lower method complexity (a single model, no mean teacher or augmentation averaging). Tent attains higher accuracy on ImageNet-C (37.1%) but is far less calibrated (10.70 ECE vs. 3.06), illustrating the accuracy–calibration trade-off. Per-corruption results are in Tables 17–19. The same objective that improves calibration for prompt tuning is thus effective on the normalization parameters of a convolutional classifier.

Comparison with post-hoc temperature scaling. Temperature scaling (TS), applied here as a single temperature fit on held-out validation data and used unchanged across all corruptions, is a strong post-hoc baseline, so we ask how it interacts with adaptation (Table 16). Its effect is dataset-dependent. On ImageNet-C, RA-TPT is already well-calibrated (raw ECE 3.06) and TS gives no further benefit, worsening both RA-TPT and CoTTA, while the overconfident entropy baseline improves but stays higher (Tent 10.70 to 6.42). On CIFAR-10/100-C, all methods retain residual overconfidence that TS reduces (RA-TPT on CIFAR-10-C, 13.8 to 4.3), though RA-TPT's raw ECE is already below the entropy baseline even after TS (Tent+TS 16.3). We therefore do not claim RA-TPT's calibration is unreachable by a post-hoc temperature; rather, the two are complementary: RA-TPT attains lower raw ECE than entropy-based adaptation, and any residual overconfidence can be further reduced by TS.

Robustness to α and the role of $\mathcal{L}_{\text{mean}}$. RA-TPT is stable across a wide range of the margin cap: on CIFAR-10-C, ECE varies by under 0.6% as α ranges over 1–80, and on ImageNet-C it stays within 3.0–3.7% over a $50\times$ range of α. On these standard-softmax backbones the logits are of moderate magnitude ($O(1\text{–}10)$), so the complementary mean-logit term $\mathcal{L}_{\text{mean}}$ is essentially inactive: replacing the full regularizer with the centered-logit dispersion term $\mathcal{L}_{\text{disp}}$ alone changes ECE by $< 0.1\%$ on both CIFAR benchmarks, confirming that *in this regime* confidence control is governed by dispersion. This balance is scale-dependent and set by the uncertainty gating, not fixed: under CLIP's low temperature ($\tau = 0.01$) the pre-softmax logits are far larger ($\bar{z} \approx 20\text{–}34$; see the logit-magnitude plots in Appendix P), so the $(1 - \omega^2)$-weighted $\mathcal{L}_{\text{mean}}$ term carries a share of the regularizer gradient on confident inputs. The two settings are therefore consistent rather than contradictory: the same

Table 17: Per-corruption accuracy (%, ↑) and ECE (%, ↓) on CIFAR-10-C at severity 5 (episodic, WRN-28-10).

	Source		Tent		CoTTA		RA-TPT	
Corruption	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
Gaussian noise	27.68	61.56	71.14	26.61	71.97	19.47	71.82	19.61
Shot noise	34.27	54.44	70.10	27.71	73.84	18.18	73.77	18.17
Impulse noise	27.08	54.69	59.88	37.19	63.70	25.36	63.66	25.36
Defocus blur	53.05	35.95	81.41	17.05	87.19	8.51	87.21	8.42
Glass blur	45.67	43.01	62.49	34.46	64.73	24.63	64.78	24.52
Motion blur	65.24	26.66	79.04	19.51	85.81	9.40	85.86	9.33
Zoom blur	58.00	32.00	84.34	14.27	87.80	7.58	87.67	7.68
Snow	74.91	18.84	74.52	23.60	82.74	11.43	82.73	11.45
Frost	58.68	33.20	75.27	22.89	82.60	11.34	82.60	11.36
Fog	74.00	19.68	79.72	18.82	84.79	10.18	84.77	10.22
Brightness	90.69	6.69	88.57	10.45	91.62	5.38	91.68	5.35
Contrast	53.29	34.60	79.14	19.61	87.37	6.96	87.48	7.24
Elastic transform	73.42	18.67	70.05	27.66	76.26	16.01	76.24	15.90
Pixelate	41.55	48.07	77.89	20.03	80.37	13.02	80.33	13.07
JPEG compression	69.71	21.90	72.19	25.28	72.69	18.96	72.69	18.87
Average	**56.48**	**34.00**	**75.05**	**23.01**	**79.57**	**13.76**	**79.55**	**13.77**

Table 18: Per-corruption accuracy (%, ↑) and ECE (%, ↓) on CIFAR-100-C at severity 5 (episodic, WRN-28-10).

	Source		Tent		CoTTA		RA-TPT	
Corruption	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
Gaussian noise	12.61	54.89	41.92	44.50	40.81	21.04	41.81	21.72
Shot noise	13.77	55.50	43.70	43.18	41.91	20.39	43.53	20.74
Impulse noise	4.75	56.09	35.22	49.40	34.16	23.14	35.16	23.79
Defocus blur	32.31	36.03	65.00	25.97	67.55	11.89	67.84	12.17
Glass blur	18.00	48.99	40.00	46.20	42.57	21.64	43.19	22.23
Motion blur	46.00	25.14	64.09	26.62	65.79	12.66	66.35	12.80
Zoom blur	41.94	27.52	65.75	25.24	68.22	11.39	68.40	11.84
Snow	49.96	22.34	56.17	32.94	57.49	16.58	58.16	16.87
Frost	38.92	33.65	57.10	32.14	57.77	15.96	58.31	16.36
Fog	46.20	27.22	59.72	30.13	60.43	14.26	61.05	14.69
Brightness	69.75	11.95	67.34	24.83	72.29	10.07	72.58	10.30
Contrast	25.91	43.13	64.17	28.33	69.62	11.61	69.97	12.32
Elastic transform	48.16	22.52	52.21	35.22	52.86	17.15	53.51	17.53
Pixelate	25.46	43.02	57.71	31.54	57.93	15.18	58.70	15.46
JPEG compression	39.30	25.14	44.31	41.97	43.01	20.31	44.15	20.70
Average	**34.20**	**35.54**	**54.29**	**34.55**	**55.49**	**16.22**	**56.18**	**16.63**

gated objective emphasizes dispersion when logits are small and adds magnitude shrinkage when they are large.

Per-corruption results. Tables 17, 18, and 19 give the full per-corruption Accuracy and ECE for all methods on CIFAR-10-C, CIFAR-100-C, and ImageNet-C, respectively.

G Dataset Details and Evaluation Protocol

Our empirical protocol stress-tests RA-TPT across 31 heterogeneous vision benchmarks, purposefully selected to capture a wide spectrum of semantic granularities, natural distribution shifts, and specialized imaging modalities. Rather than detailing each dataset individually, we group the benchmarks by their core evaluation objectives below. The specific class counts and test set sizes for the natural imagery datasets are summarized in Table 20.

Table 19: Per-corruption accuracy (%, ↑) and ECE (%, ↓) on ImageNet-C at severity 5 (episodic, ResNet-50). Tent uses SGD with momentum (canonical protocol; see Appendix F).

	Source		Tent		CoTTA		RA-TPT	
Corruption	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
Gaussian noise	2.88	19.91	23.56	10.68	16.16	1.77	16.46	1.38
Shot noise	3.52	23.20	24.52	12.09	17.02	2.67	17.28	2.62
Impulse noise	2.44	21.52	23.62	10.50	15.68	2.16	15.86	1.73
Defocus blur	18.24	5.08	19.22	15.96	15.32	4.25	15.24	4.19
Glass blur	10.34	13.10	18.70	15.62	15.44	3.38	15.32	3.05
Motion blur	14.52	14.50	33.44	9.74	25.88	2.47	26.44	1.80
Zoom blur	21.94	16.00	45.40	6.22	38.66	2.41	38.58	3.18
Snow	16.50	22.92	43.38	6.68	34.10	1.36	34.36	1.28
Frost	22.58	16.11	36.96	10.57	31.66	2.64	32.42	2.74
Fog	23.60	18.59	54.38	3.98	47.84	4.04	47.90	4.98
Brightness	58.86	5.38	66.88	4.01	64.96	2.59	64.38	2.43
Contrast	5.38	7.48	11.36	39.89	16.96	7.18	17.24	6.42
Elastic transform	17.00	25.21	51.04	5.51	44.04	2.53	43.64	3.15
Pixelate	20.88	14.84	55.92	4.21	48.80	2.92	47.84	2.77
JPEG compression	32.54	10.15	48.72	4.89	40.90	3.62	41.16	4.21
Average	**18.08**	**15.60**	**37.14**	**10.70**	**31.56**	**3.07**	**31.61**	**3.06**

In-Domain and Fine-Grained Recognition. To evaluate adaptation stability when confronted with subtle intra-class variances, diverse textures, and specialized object geometries, we utilize canonical ImageNet (INet) (Deng et al., 2009) alongside 9 fine-grained datasets. This includes biological classifications (OxfordPets (Pets) (Parkhi et al., 2012), Flowers102 (FLW) (Nilsback & Zisserman, 2008)), specialized vehicle/object recognition (StanfordCars (Car) (Krause et al., 2013), FGVC-Aircraft (Air.) (Maji et al., 2013), Caltech101 (C101) (Fei-Fei et al., 2004)), human action contexts (UCF101 (UCF) (Soomro et al., 2012)), and texture/material variations (DTD (Cimpoi et al., 2014), Food101 (Food) (Bossard et al., 2014)), and Remote sensing benchmark (Eurosat (SAT) (Helber et al., 2018)).

Out-of-Distribution (OOD) Shifts. A critical measure of reliability is a model's behavior under domain drift. We assess structural safety using four severe ImageNet variants: ImageNet-A (I-A) (Hendrycks et al., 2021b) (naturally occurring adversarial examples), ImageNet-V2 (I-V2) and ImageNet-R (I-R) (Hendrycks et al., 2021a) (temporal, source, and artistic shifts), and ImageNet-Sketch (I-S) (Wang et al., 2019) (structural and texture-stripped representations).

Cross-Modality Foundation Models. Beyond natural imagery, we rigorously test whether our logit-space calibration mechanisms generalize to entirely different visual domains. For medical image understanding, we utilize BioMedCLIP (Zhang et al., 2025) to evaluate 11 specialized datasets spanning 9 distinct modalities: computed tomography (CTKidney (Islam et al., 2022)), dermatoscopy (DermaMNIST (Derma.) (Codella et al., 2018; Tschandl et al., 2018)), endoscopy (Kvasir (Pogorelov et al., 2017)), histopathology (LC25000 (LungCol.) (Borkowski et al., 2019), CHMNIST (Kather et al., 2016)), MRI (BTMRI (Nickparvar, 2021)), optical coherence tomography (OCTMNIST (OCT.) (Kermany et al., 2018)), ultrasound (BUSI (Al-Dhabyani et al., 2020)), X-ray (COVID-QU-Ex (COVID19) (Tahir et al., 2021), KneeXray (Chen, 2018)), and fundus imaging (RETINA (Köhler et al., 2013; Porwal et al., 2018)).

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 evaluate EuroSAT (Helber et al., 2018) within this dedicated remote sensing suite. These datasets present unique challenges, such as arbitrary visual orientations, varying spatial resolutions, and extreme domain shifts.

Table 20: **Dataset Statistics.** Number of classes (K) and test set size (N) for the generic, fine-grained, and distribution shift benchmarks evaluated with CLIP.

Dataset	Classes (K)	Test Set Size (N)
Generic & Fine-Grained Classification		
ImageNet (Deng et al., 2009)	1000	50,000
Caltech101 (Fei-Fei et al., 2004)	100	2,465
OxfordPets (Parkhi et al., 2012)	37	3,669
StanfordCars (Krause et al., 2013)	196	8,041
Flowers102 (Nilsback & Zisserman, 2008)	102	2,463
Food101 (Bossard et al., 2014)	101	30,300
FGVC-Aircraft (Maji et al., 2013)	100	3,333
DTD (Cimpoi et al., 2014)	47	1,692
EuroSAT (Helber et al., 2018)	10	8,100
UCF101 (Soomro et al., 2012)	101	3,783
Natural Distribution Shifts		
ImageNet-A (Hendrycks et al., 2021b)	200	7,500
ImageNet-V2 (Hendrycks et al., 2021a)	1000	10,000
ImageNet-R (Hendrycks et al., 2021a)	200	30,000
ImageNet-Sketch (Wang et al., 2019)	1000	50,889

H Extended Implementation Details and Baselines

This section details the optimization settings, baseline evaluation protocols, and architectural choices used in our experiments to ensure fair comparison and reproducibility.

Architecture and Optimization. We adopt pre-trained CLIP models with both Vision Transformer (ViT-B/16) and ResNet-50 (RN50) backbones for natural distribution shifts, extending to BiomedCLIP and RemoteCLIP for specialist domains. Consistent with the Test-Time Prompt Tuning (TPT) framework (Shu et al., 2022), prompt embeddings are initialized as hard text prompts (e.g., "`a photo of a <class>`"). Adaptation is performed strictly instance-by-instance: for each test image we generate 63 augmented views (total batch size 64) and take a single AdamW step with learning rate 1×10^{-3}. The backbones remain frozen; only the context tokens are updated. After the forward and backward pass the model state is rolled back, ensuring true zero-shot inference with no sequential batch dependence. All experiments were run on a single NVIDIA RTX Pro 6000 GPU and averaged over three independent runs. All optimization and objective settings are summarized in Table 21.

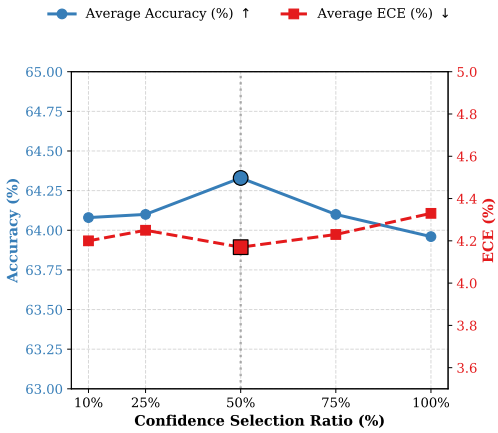


Figure 8: **Confidence Selection Ratio Ablation.** Average performance across 9 standard datasets excluding ImageNet (ViT-B/16). Unlike standard TPT, which requires aggressive 10% filtering to avoid collapse, RA-TPT's logit regularization safely allows a 50% retention ratio, balancing rich semantic gradients against bounded logit inflation.

Evaluation Protocol and Baseline Reproducibility. Fully unsupervised test-time adaptation assumes the target distribution is unknown at inference, so dataset-specific hyperparameter tuning is not permitted, as it implicitly leaks target-domain information. We therefore re-implement all baselines (TPT (Shu et al., 2022), C-TPT (Yoon et al., 2024), O-TPT (Sharifdeen et al., 2025), A-TPT (Ahamed et al., 2026), R-TPT (Sheng et al., 2025), and SoC (Fillioux et al., 2026)) on the standard TPT backbone and, for each, use the single dataset-agnostic configu-

Table 21: Complete RA-TPT hyperparameters, fixed across all 31 benchmarks and selected once on a 5-shot ImageNet subset. β weights the mean-logit stabilizer and is fixed at 0.3; see Appendix K for its insensitivity.

Setting	Value
Margin cap α	4
Margin relaxation λ	1.5
Coherence weight γ	5
Projection sharpness ξ	10
Mean-logit weight β	0.3
Learning rate	1×10^{-3}
Optimizer	AdamW
Adaptation steps T	1
Augmented views N	63 (+1 original)
View retention (RA-TPT)	top 50%
View retention (entropy baselines)	top 10%

Table 22: **Matched view-retention study** (CLIP-ViT-B/16, nine standard datasets; ImageNet excluded). Average Accuracy (↑) and ECE (↓) when every method uses the same confidence-based retention ratio. RA-TPT attains the lowest ECE at both 10% and 50%, so its advantage is not an artifact of the retention setting. **Bold**: best ECE per column.

	Top-10% retention		**Top-50% retention**	
Method	**Acc.**	**ECE**	**Acc.**	**ECE**
TPT	64.64	11.60	64.63	11.00
O-TPT	63.76	4.98	63.78	4.70
SoC	63.57	4.58	63.71	4.65
RA-TPT	64.08	**4.20**	64.33	**4.17**

ration reported in its paper, applied unchanged across all datasets. Where a repository additionally provides per-dataset settings, we do not adopt them, since this would violate the unsupervised setting; consequently our baseline numbers reflect a uniform, un-tuned protocol and can be slightly higher in average ECE than repository-specific results. RA-TPT is held to the same standard: a single fixed configuration (Table 21) selected on a held-out ImageNet subset and frozen across all 31 benchmarks. We also note the concurrent FPP (Jang et al., 2026), which we exclude from empirical comparison because its required custom pre-training phase falls outside the off-the-shelf zero-shot TTA setting.

Confidence Selection Ratio Ablation. Standard TPT and its geometric variants retain only the top 10% most confident views for entropy minimization. This aggressive filtering is a necessary heuristic for prior methods: lacking structural bounds on logit magnitude, ingesting noisy augmentations triggers immediate calibration collapse. In Figure 8 we ablate this ratio across 9 datasets with ViT-B/16. Because RA-TPT introduces an uncertainty-routed logit regularizer that constrains predictive concentration, it is no longer fragile to noisy views: retaining 50% of the views gives the best balance, leveraging a richer semantic gradient signal (average accuracy 64.33%) while keeping calibration bounded (ECE 4.17%). We adopt this 50% ratio as the RA-TPT default in all experiments.

Matched view-retention control. The main results use each method at its published-optimal retention (top 10% for the entropy baselines, top 50% for RA-TPT). To confirm that RA-TPT's advantage does not stem from this asymmetry, we evaluated baselines at both retention levels on the nine standard datasets (Table 22). Average ECE is nearly invariant to the ratio for all methods (changes below 0.6 ECE), and RA-TPT attains the lowest ECE at both 10% (4.20) and 50% (4.17). The advantage therefore follows from the objective, not the view budget. The separation is clearest in the worst case: on EuroSAT at the matched 50% setting, O-TPT and SoC still reach 13.19 and 12.56 ECE, whereas RA-TPT stays at 4.28. RA-TPT is also the only method whose accuracy improves with higher retention ($64.08 \rightarrow 64.33$), consistent with the tolerance of its bounded regularizer to lower-confidence views.

I Comprehensive Calibration Metrics

To provide a rigorous evaluation of model reliability beyond standard Expected Calibration Error (ECE), we assess our framework across a diverse suite of calibration and proper scoring metrics.

Notation. Let n represent the total number of test samples, K the number of target classes, and B the total number of confidence bins (We used 20 number of bins for all the experiments). We define $\mathcal{B}_b$ as the subset of samples assigned to the b-th bin. For a given sample i, y_i denotes the ground-truth label, and $p_{i,k}$ represents the predicted softmax probability for class k. The empirical accuracy and the average predicted confidence within a specific bin b are denoted as $\mathrm{acc}(\mathcal{B}_b)$ and $\mathrm{conf}(\mathcal{B}_b)$, respectively. The indicator function is given by $\mathbf{1}[\cdot]$.

Expected Calibration Error (ECE). It quantifies the expected absolute difference between a model's predicted confidence and its true accuracy. It partitions predictions into B equally spaced intervals and computes the weighted average of the bin-wise errors:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{n} \big|\mathrm{acc}(\mathcal{B}_b) - \mathrm{conf}(\mathcal{B}_b)\big|. \tag{43}$$

Static Calibration Error (SCE). Unlike ECE, which weights errors by the number of samples per bin, SCE computes the simple arithmetic mean of the calibration errors across all B fixed-width bins. This ensures that severe miscalibration in low-density bins is not masked by larger bins:

$$\mathrm{SCE} = \frac{1}{B} \sum_{b=1}^{B} \big|\mathrm{acc}(\mathcal{B}_b) - \mathrm{conf}(\mathcal{B}_b)\big|. \tag{44}$$

Maximum Calibration Error (MCE). It measures the worst-case calibration disparity across all confidence bins. This metric is particularly critical for safety-oriented applications where maximum bounded error is a strict requirement:

$$\mathrm{MCE} = \max_{b \in \{1,\dots,B\}} \big|\mathrm{acc}(\mathcal{B}_b) - \mathrm{conf}(\mathcal{B}_b)\big|. \tag{45}$$

Adaptive Calibration Error (ACE). To account for heavily skewed sample distributions across standard confidence intervals, ACE dynamically defines bin boundaries such that each bin contains an identical number of samples ($|\mathcal{B}_b| = n/B$):

$$\mathrm{ACE} = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{n} \big|\mathrm{acc}(\mathcal{B}_b) - \mathrm{conf}(\mathcal{B}_b)\big|. \tag{46}$$

Brier Score (BS). The Brier Score serves as a strict proper scoring rule. It evaluates the mean squared difference between the predicted probability distribution and the one-hot encoded ground-truth vector:

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \big(p_{i,k} - \mathbf{1}[y_i = k]\big)^2. \tag{47}$$

Negative Log-Likelihood (NLL). NLL, or standard cross-entropy, directly penalizes the model for assigning low probabilities to the correct class, making it highly sensitive to overconfident incorrect predictions:

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log p_{i,y_i}. \tag{48}$$

Summary of Results. Table 23 details the performance of RA-TPT against baseline adaptation methods across these diverse metrics. For interpretability, SCE, MCE, and ACE are reported as percentages (%), whereas the Brier Score and NLL are reported on their standard probabilistic scales. As demonstrated, RA-TPT systematically outperforms recent geometric methods (such as SoC, O-TPT, A-TPT, and C-TPT) across all metric categories, visually corroborating the reliability diagrams in the main text and confirming that our approach fundamentally improves the structural integrity of the predictive distribution.

Table 23: **Comprehensive Calibration Metrics.** Full per-dataset evaluation of test-time adaptation methods using the CLIP-ViT-B/16 backbone. Static Calibration Error (SCE), Maximum Calibration Error (MCE), and Adaptive Calibration Error (ACE) are reported as percentages (%), while Brier Score (BS) and Negative Log-Likelihood (NLL) are reported as probabilistic values. Bold values indicate the best average performance.

Method	Metric	INet	DTD	FLW	Food	Air.	Pets	C101	UCF	SAT	Car	Avg.
TPT	SCE	8.20	18.85	15.97	4.05	22.16	12.72	7.11	15.46	16.93	5.90	14.73
	MCE	17.86	37.78	29.73	8.78	46.03	27.95	90.55	28.56	34.33	13.18	33.48
	ACE	10.08	21.46	13.44	3.92	17.11	5.56	4.48	12.97	21.33	4.98	11.53
	BS	0.460	0.756	0.450	0.240	0.920	0.200	0.120	0.460	0.760	0.450	0.482
	NLL	1.330	2.460	1.770	0.690	3.530	0.580	0.310	1.350	1.870	1.010	1.490
C-TPT	SCE	3.51	11.67	7.13	4.62	5.41	5.19	18.55	3.38	10.75	2.73	7.29
	MCE	6.36	25.45	17.26	8.86	12.17	24.72	82.18	13.16	28.79	13.49	23.24
	ACE	3.16	12.40	5.33	3.90	4.66	1.73	4.48	2.12	13.26	1.71	5.28
	BS	0.440	0.720	0.440	0.250	0.860	0.173	0.110	0.460	0.750	0.450	0.465
	NLL	1.230	2.260	1.680	0.680	3.010	0.410	0.260	1.280	1.830	1.010	1.365
A-TPT	SCE	3.44	9.13	6.53	4.65	7.46	5.65	22.59	2.75	12.92	2.25	7.74
	MCE	8.17	32.31	16.18	9.38	18.11	24.70	86.93	7.44	30.61	7.80	24.16
	ACE	2.46	9.04	4.51	3.67	7.54	1.89	5.55	2.10	17.51	1.76	5.59
	BS	0.450	0.710	0.410	0.250	0.870	0.178	0.120	0.480	0.800	0.460	0.473
	NLL	1.240	2.230	1.670	0.690	3.050	0.430	0.260	1.340	1.910	1.030	1.385
O-TPT	SCE	2.98	8.26	5.45	5.88	5.58	5.22	18.95	2.68	12.34	2.56	6.99
	MCE	8.18	19.48	13.43	11.13	13.23	25.13	85.92	11.01	29.17	13.90	23.07
	ACE	2.38	7.86	4.48	4.85	4.46	2.40	7.30	2.32	13.80	1.90	4.88
	BS	0.450	0.703	0.413	0.260	0.862	0.174	0.114	0.480	0.760	0.460	0.468
	NLL	1.360	2.200	1.650	0.693	3.040	0.250	0.330	1.850	1.030	1.230	1.363
SoC	SCE	2.31	7.48	4.89	5.81	5.56	5.17	19.84	3.95	8.05	2.83	6.59
	MCE	5.41	19.55	15.52	10.51	17.17	24.96	85.49	19.90	22.68	12.85	23.40
	ACE	1.90	6.78	3.57	4.72	4.11	2.29	3.97	2.85	11.24	2.14	4.36
	BS	0.450	0.699	0.417	0.257	0.863	0.174	0.111	0.477	0.755	0.456	0.466
	NLL	1.235	2.193	1.659	0.698	3.006	0.400	0.241	1.327	1.843	1.023	1.363
RA-TPT (Ours)	SCE	3.51	8.07	5.70	3.98	7.39	6.41	6.14	3.65	3.59	4.40	**5.28**
	MCE	8.43	22.34	13.37	8.06	22.30	23.84	8.96	11.45	14.80	9.24	**14.28**
	ACE	2.67	7.58	3.98	2.49	5.60	2.94	4.28	2.19	4.49	4.47	**4.07**
	BS	0.450	0.690	0.430	0.230	0.860	0.170	0.120	0.450	0.690	0.450	**0.454**
	NLL	1.260	2.120	1.620	0.670	2.880	0.460	0.260	1.250	1.670	1.020	**1.321**

J AUGMENTATION EFFICIENCY

In Section 5 of the main text, we noted that RA-TPT remains highly calibrated even when the number of augmented views (N) is reduced to a single, unaugmented pass ($N = 1$). This property is critical for real-world deployment on edge devices, where the forward-pass bottleneck of generating 64 augmented views per image introduces unacceptable latency. Figure 9 visualizes the performance of standard TPT, O-TPT, and RA-TPT across varying augmentation batch sizes on the ViT-B/16 architecture.

Experimental Protocol and Filtering Override. Standard implementations of TPT and O-TPT utilize an entropy-based filtering mechanism (e.g., retaining only the top 10% most confident augmented views for the backward pass). However, applying a percentage-based filter creates a confounding variable when ablating the absolute number of views (N), and completely breaks the algorithm at constrained regimes (e.g., a 10% filter on $N = 4$ yields < 1 view). Therefore, to isolate the true effect of augmentation batch size, *we completely disabled the confidence selection mechanism for this entire analysis*. Across all methods and all values of N from 128 down to 1, every algorithm utilized 100% of the generated views. This provides a strict, unfiltered stress test of each method's inherent stability.

The Degradation of Entropy Minimization at Low N. As shown in Figure 9 (right), standard TPT's Expected Calibration Error (ECE) degrades sharply as N decreases, jumping from 11.23 at $N = 64$ to a severe 15.88 ECE at $N = 1$. Without the stabilizing consensus of a large augmented batch (and lacking the safety net of confidence filtering), standard TPT applies unbounded entropy-minimization gradients to single, noisy visual representations. This forces the prompt to memorize the specific stochastic artifacts of that single view, triggering rapid logit inflation and catastrophic overconfidence. While O-TPT's geometric constraints partially mitigate this spike (rising to 5.45 ECE at $N = 1$), it still exhibits noticeable degradation.

Boundary Condition at $N = 1$: Robustness via Logit Regularization. RA-TPT exhibits an exceptionally flat ECE curve, maintaining a low 3.08 ECE even with a single, unaugmented view ($N = 1$). As noted in our theoretical analysis, the coherence gate relies on view disagreement. At this $N = 1$ boundary condition, inter-view variance is trivially zero ($\mathcal{V} = 0$), causing the gate to remain wide open ($\mathcal{C} = 1$). In this regime, $\mathcal{L}_{\text{sep}}$ loses its variance-reduction property and applies the unfiltered gradient of the single view.

Unlike standard TPT, RA-TPT does not rely on cross-view consensus to prevent logit inflation: at N=1 its dispersion penalty and uncertainty-capped margin keep ECE bounded (3.08 on ImageNet-V2), whereas TPT collapses to 15.88. We note, however, that a single unaugmented view carries no adaptation signal beyond zero-shot: at N=1 RA-TPT's accuracy matches zero-shot (60.8) and its ECE is marginally higher (3.08 vs. 2.79). N=1 therefore marks a graceful stability floor, not a recommended operating point. The practical implication is that RA-TPT degrades gracefully as N is reduced, so practitioners can use substantially fewer than 64 views to cut adaptation cost while retaining calibrated behavior, whereas entropy-based methods require aggressive augmentation and filtering to avoid collapse.

K HYPERPARAMETER SENSITIVITY ANALYSIS

In Section 4 of the main text, we established that RA-TPT operates using a fixed set of five objective hyperparameters, $\{\alpha, \lambda, \gamma, \xi\}$, together with a complementary mean-logit stabilizer weight β (fixed at 0.3; Table 21). We select these values once and freeze them across all 31 evaluated benchmarks. Here, we provide the rigorous sensitivity analysis demonstrating the stability of this formulation.

Experimental Setup. To avoid overfitting to the test distributions during tuning, we performed this ablation on a miniaturized, class-balanced subset of the ImageNet validation set, sampling exactly 5 images per class (5,000 images total) using the ViT-B/16 backbone. We independently swept the four components that shape the objective: the Coherence Weight (γ), the margin target (α), the margin relaxation weight (λ), and the soft gradient projection weight (ξ). During each sweep, all other parameters were frozen at their optimal defaults ($\gamma = 5$, $\alpha = 4$, $\lambda = 1.5$, $\xi = 10$). The stabilizer weight β is not a primary objective parameter; as it only scales the complementary mean-logit term,

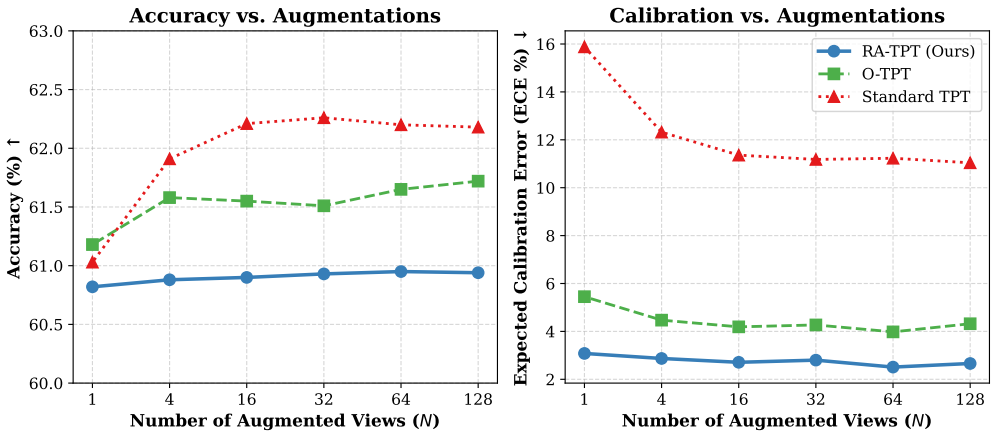


Figure 9: **Effect of Augmentation Batch Size (N).** Performance on the ImageNet-V2 (ViT-B/16). To ensure a controlled evaluation of edge-case latency, confidence filtering was disabled for all methods across all values of N. While TPT's ECE sharply degrades as $N \to 1$, RA-TPT maintains an exceptionally flat calibration curve due to its bounded logit dispersion.

we hold it fixed at $\beta = 0.3$ and found performance to be largely insensitive to it, so no per-component sweep is required.

Results and Stability. The results, visualized in Figure 10, demonstrate that RA-TPT is exceptionally stable. Across wide variations in hyperparameter magnitudes, the classification accuracy remains tightly bounded (fluctuating by less than 0.3%), while the Expected Calibration Error (ECE) strictly remains below 3.0%. Crucially, this indicates that our framework does not require fragile tuning to achieve its competitive calibration. For instance, γ and ξ naturally settle into stable saturation regimes ($\gamma \in [0.5, 5]$ and $\xi > 5$). The complementary stabilizer β is similarly forgiving: varying it around its default leaves both accuracy and ECE essentially unchanged. The structural constraints of the logit regularizer and the Coherence Gate self-regulate the optimization trajectory, ensuring that even sub-optimal hyperparameter initializations yield safely calibrated test-time adaptation.

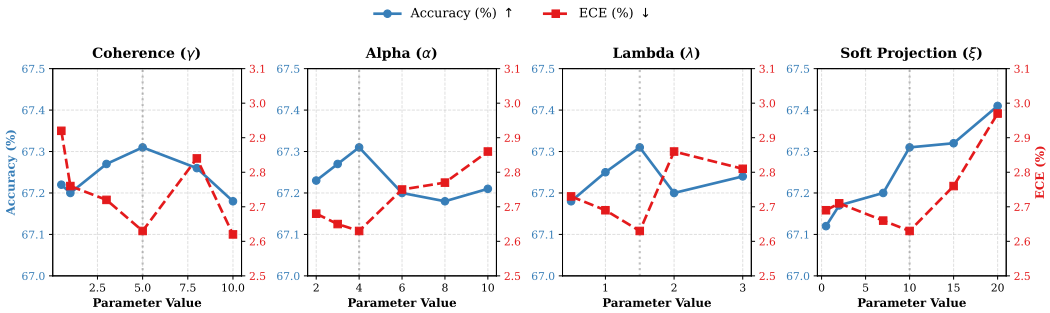


Figure 10: **Hyperparameter Sensitivity on ImageNet (5-shot subset, ViT-B/16).** The dual-axis plot illustrates the stability of RA-TPT across the four objective-shaping components $\{\alpha, \lambda, \gamma, \xi\}$ (the complementary stabilizer β is held fixed at 0.3; see text). The solid blue line denotes Accuracy (↑) and the dashed red line denotes ECE (↓). The vertical dotted gray line marks our default configuration ($\gamma = 5$, $\alpha = 4$, $\lambda = 1.5$, $\xi = 10$), which achieves the optimal balance of 67.31% Accuracy and 2.63% ECE. The tightly bounded y-axis limits underscore the framework's inherent resistance to hyperparameter brittleness.

L Computational Cost and Memory Footprint Analysis

In Section 5 we reported that RA-TPT reduces the peak VRAM footprint by 55% relative to O-TPT (39.7 GB → 17.9 GB), bringing test-time adaptation within a single 24 GB GPUs. Here we separate

the two distinct axes of cost, which are compute (FLOPs/latency) and peak memory and we explain each; conflating them obscures where the savings actually come from.

Setup. Let N_{gen} be the number of augmented views generated in the forward pass (typically $N_{\text{gen}} = 64$), K the number of classes, d the textual embedding dimension, and M the number of learnable context tokens. In all methods the visual encoder Φ_I and text encoder Φ_T are frozen; the only trainable tensor is the M-token prompt (2,048 parameters out of 76.71M total), and it lies on the text branch.

Forward pass (identical across methods). Every method passes the N_{gen} views through Φ_I and computes the class embeddings $\mathbf{W} \in \mathbb{R}^{d\times K}$ through Φ_T. Since the prompt length and backbone (e.g., ViT-B/16) are shared, forward FLOPs are identical at 3284.51 GFLOPs. Crucially, because Φ_I is frozen and carries no learnable parameter upstream, the view forwards are *not* part of the prompt's autograd graph: their activations are transient working memory, released after the pass, and no backward graph is retained through the visual encoder.

Compute scaling with K (backward pass). The methods differ in how their regularizer is computed.

- **Embedding-space baselines (O-TPT, A-TPT, C-TPT, SoC):** these enforce structural penalties directly on the text-feature matrix $E \in \mathbb{R}^{K\times d}$. O-TPT, for instance, minimizes $\|EE^\top - I_K\|_2^2$ (Sharifdeen et al., 2025), an orthogonality constraint that couples all K class embeddings through a dense $K \times K$ Gram matrix, resulting in an $\mathcal{O}(K^2 d)$ *compute* cost that scales quadratically in the number of classes.

- **RA-TPT (ours):** we operate on low-dimensional logit statistics rather than $\mathbf{W}$-space geometry. The coherence gate $\mathcal{C}(\bar{\mathbf{p}})$ costs $\mathcal{O}(N \cdot K)$; the logit regularizers ($\mathcal{L}_{\text{mean}}, \mathcal{L}_{\text{disp}}$) are $\mathcal{O}(K)$ on the mean-logit vector $\bar{\mathbf{z}} \in \mathbb{R}^K$; and the soft gradient projection is $\mathcal{O}(Md)$ on the prompt gradients. With $N \ll K$ and negligible Md, RA-TPT's regularizer is $\mathcal{O}(K)$.

This $\mathcal{O}(K)$ versus $\mathcal{O}(K^2 d)$ gap is a *compute* difference: it governs how per-step FLOPs grow with the class count, not the peak memory. In particular, the $K \times K$ Gram matrix itself is tiny, about 4 MB at K=1000 in fp32, so it is not, on its own, a memory bottleneck.

Peak memory. The 17.9 GB peak shared by TPT and RA-TPT is dominated by the transient forward activations of encoding the N_{gen} views through the frozen Φ_I, together with the retained text-branch backward graph over the K class prompts (the only path to the learnable prompt). Both are architecturally identical across methods. RA-TPT adds no embedding-space state: its regularizers act on already-reduced scalar logit statistics, so no dense per-class intermediates are retained. O-TPT's higher 39.7 GB peak instead comes from its $\|EE^\top - I_K\|_2^2$ objective, which couples all K class embeddings and therefore forces autograd to retain the text-encoder backward graph over all K class prompts *jointly*, rather than from the size of the $K \times K$ Gram matrix, which is only $\approx$4 MB. Operating downstream in logit space is therefore what removes O-TPT's excess memory, and is why RA-TPT fits within 24 GB.

Parity with standard TPT despite higher view retention. RA-TPT reaches the *same* 17.9 GB peak as standard TPT and only a marginal latency increase (542 ms vs. 491 ms), even though it retains a larger fraction of confident views (50% vs. 10%) and computes two gradient terms for the projection. The reason is that confidence-based view selection is a *post-hoc* operation applied *after* the shared forward pass: all methods first generate and encode the full set of $N_{\text{gen}} = 64$ views through the frozen encoders to compute per-view confidence, and the retention ratio only determines which of these *already-computed* per-view logits contribute to the adaptation loss. Because the visual encoder is frozen and outside the autograd graph, changing the retention ratio does not change the retained backward graph at all; therefore, 10% and 50% retention yield the same 17.9 GB peak. RA-TPT's only overhead relative to TPT lies on the backward side: obtaining the separation and regularization gradients g_{sep} and g_{reg} for the soft projection requires one additional backward traversal plus an $\mathcal{O}(Md)$ projection over the prompt gradients, which accounts for the +51 ms. Because the N_{gen}-view forward dominates wall-clock time, this second backward adds roughly 10% rather than doubling latency.

M Extended Prompt Robustness Analysis

In Section 5 of the main text, we briefly established that RA-TPT is highly robust to the choice of the initial prompt template. Here, we provide a detailed elaboration of this vulnerability in standard Test-Time Prompt Tuning (TPT) and analyze how our framework structurally reduces template sensitivity.

Experimental Setup. Zero-shot Vision-Language Models are notoriously sensitive to the phrasing of the text prompt. To evaluate adaptation stability, we tested 18 diverse prompt templates (e.g., *"a photo of a..."*, *"a blurry photo of a..."*, *"a low contrast photo of the..."*) across three structurally distinct datasets: FGVC-Aircraft (fine-grained object classification), DTD (texture recognition), and UCF101 (action recognition), utilizing the ResNet-50 backbone. For each template, we compute the pointwise average ECE across the three datasets to measure the method's overall reliability under that specific linguistic initialization. The resulting distributions are visualized in the boxplot in Figure 5.

The Fragility of Standard Entropy Minimization. As illustrated in Figure 5, standard TPT exhibits extreme sensitivity to the initial prompt. Its average ECE across the 18 templates fluctuates strongly, with the interquartile range stretching across a large margin and maximum ECEs frequently exceeding 25% on average. The theoretical mechanism behind this failure stems from the unbounded nature of entropy minimization. If a sub-optimal prompt (e.g., *"a bad photo of a..."*) induces a misaligned or flattened initial logit distribution, the standard TPT objective immediately applies maximum gradient pressure to force the model to become confident. Because there is no structural bound on logit magnitude, the optimizer blindly inflates the logits of the nearest class, trapping the prompt in a highly overconfident, miscalibrated local minimum.

Insufficiency of Geometric Constraints. O-TPT attempts to mitigate this by enforcing orthogonality on the textual embeddings ($\mathbf{W}$). While this reduces the worst-case miscalibration compared to standard TPT, it still exhibits high variance. Because geometric regularizers are fundamentally *magnitude-agnostic*, they cannot stop the logit magnitudes from scaling up if the initial template places the text embeddings in a poor geometric configuration relative to the visual features.

RA-TPT: Robustness via Coherence Gating. RA-TPT demonstrates strong robustness, achieving the lowest mean ECE and a tight variance across the 18 templates. This stability is not accidental, but a direct consequence of our formulation:

1. **The Coherence Gate ($\mathcal{C}$):** If an unusual prompt template yields unstable or contradictory predictions across the N augmented views, our augmentation variance penalty $\mathcal{V}(\bar{\mathbf{p}})$ spikes, forcing the coherence gate $\mathcal{C}$ toward zero. This automatically throttles the margin gradient, preventing the model from chasing noise induced by a bad linguistic initialization.
2. **Uncertainty-routed regularization:** Regardless of where the prompt initializes, penalizing centered-logit dispersion ($\mathcal{L}_{\text{disp}}$) bounds the attainable maximum confidence, limiting how far the optimizer can concentrate the prediction; the mean-logit term ($\mathcal{L}_{\text{mean}}$) serves as an optimization stabilizer.

Consequently, RA-TPT safely adapts the representations while keeping deployment reliability largely insensitive to the user's initial choice of text template.

Evaluated Prompt Templates. To ensure strict reproducibility of our robustness ablation, we detail the linguistic initializations used in Figure 5. Following standard practices for evaluating vision-language models, we constructed 18 diverse templates by altering the phrasing, article usage, and descriptive adjectives. Let `[CLASS]` denote the ground-truth target class name. The evaluated templates are listed in Table 24.

Per-Dataset Breakdown. While Figure 5 visualizes the pointwise average across the three evaluated domains, we provide the complete, unaggregated per-dataset Expected Calibration Error (ECE $\downarrow$) in Table 25. This breakdown shows that RA-TPT's robustness is not merely an artifact of averaging, but a consistent structural property across fine-grained objects (Aircraft), textures (DTD), and action recognition (UCF101).

Interestingly, the accuracy across different prompt templates remains relatively stable for all evaluated methods. For example, on the Aircraft dataset, standard TPT exhibits an accuracy standard deviation of ± 0.95, O-TPT shows ± 2.05, and our proposed RA-TPT maintains a tight ± 0.90.

Table 24: **List of Evaluated Prompt Templates.** The 18 linguistic initializations used to evaluate adaptation stability in Section 5 and Appendix M.

Index	Template (Indefinite Article)	Index	Template (Definite Article)
1	"a photo of a `[CLASS]`"	10	"a photo of the `[CLASS]`"
2	"a blurry photo of a `[CLASS]`"	11	"a blurry photo of the `[CLASS]`"
3	"a black and white photo of a `[CLASS]`"	12	"a black and white photo of the `[CLASS]`"
4	"a low contrast photo of a `[CLASS]`"	13	"a low contrast photo of the `[CLASS]`"
5	"a high contrast photo of a `[CLASS]`"	14	"a high contrast photo of the `[CLASS]`"
6	"a bad photo of a `[CLASS]`"	15	"a bad photo of the `[CLASS]`"
7	"a good photo of a `[CLASS]`"	16	"a good photo of the `[CLASS]`"
8	"a photo of a small `[CLASS]`"	17	"a photo of the small `[CLASS]`"
9	"a photo of a big `[CLASS]`"	18	"a photo of the big `[CLASS]`"

Table 25: **Per-Dataset Prompt Robustness (ECE ↓).** Detailed breakdown of Expected Calibration Error across the 18 evaluated prompt templates. RA-TPT consistently prevents the severe miscalibration spikes seen in standard TPT and O-TPT across all three domains.

Prompt Idx.	FGVC-Aircraft			DTD			UCF101		
	TPT	O-TPT	RA-TPT (Ours)	TPT	O-TPT	RA-TPT (Ours)	TPT	O-TPT	RA-TPT (Ours)
1	15.33	8.33	**6.42**	25.70	16.45	**8.74**	12.08	2.43	**2.16**
2	19.67	10.63	**8.63**	29.72	14.78	**10.85**	16.86	6.78	**3.08**
3	24.97	9.48	**9.02**	33.55	22.26	**15.27**	19.35	**5.00**	6.10
4	21.22	12.94	**7.04**	29.41	18.51	**14.01**	21.73	10.56	**5.48**
5	29.10	11.50	**10.42**	36.10	23.81	**16.97**	23.12	9.04	**4.96**
6	15.62	9.63	**7.17**	27.30	15.54	**11.23**	15.36	7.99	**2.98**
7	20.25	10.72	**6.95**	31.86	18.72	**13.21**	15.72	5.82	**2.51**
8	14.04	6.52	**5.37**	27.35	20.17	**11.45**	16.15	3.64	**2.73**
9	20.02	13.30	**9.14**	33.30	23.18	**14.54**	14.54	4.54	**2.44**
10	15.32	9.36	**7.27**	20.52	12.54	**7.01**	9.38	3.58	**2.92**
11	19.15	11.68	**8.08**	27.16	15.86	**9.93**	14.29	6.56	**2.26**
12	25.22	10.80	**10.09**	28.07	**9.01**	10.61	18.38	5.15	**4.70**
13	20.89	16.37	**7.77**	22.33	16.71	**7.84**	20.09	10.20	**4.26**
14	27.31	14.73	**10.37**	24.44	17.37	**8.78**	20.19	9.68	**2.74**
15	15.05	16.80	**6.54**	25.27	11.20	**9.98**	13.42	5.32	**2.74**
16	20.61	11.94	**7.03**	27.19	15.26	**9.79**	13.65	3.48	**1.79**
17	14.56	9.25	**5.61**	23.54	17.50	**7.99**	14.68	3.82	**2.34**
18	18.26	14.49	**8.68**	26.21	19.49	**9.82**	12.97	4.96	**2.27**

N Comparison with Post-Hoc Temperature Scaling

To isolate the calibration benefit of RA-TPT's instance-aware adaptation from simple global confidence softening, we compare against a **Zero-Shot CLIP + Temperature Scaling (TS)** reference. The temperature scalar is fit on the 5-shot ImageNet subset; a sweep selects $T = 1.2$, applied uniformly across the evaluation suite. Because TS is a global post-hoc division of the logits, it preserves the argmax ordering and therefore does not alter the zero-shot classification accuracy; we thus report only ECE here (RA-TPT accuracy is in Table 2).

We deliberately do *not* temperature-scale the other baselines: each competing method is evaluated as published, and TS is included as a separate, clearly-labeled post-hoc calibration reference rather than as a competing test-time method. As Table 26 shows, a single global temperature, improved performance on DTD (4.29) and on Aircraft (2.53), but fails to transfer: on the remaining seven datasets it is substantially worse than RA-TPT (e.g., Pets 8.19 vs. 3.19, Cars 10.02 vs. 4.21, UCF101 6.25 vs. 2.24), yielding an average ECE of 6.59% against RA-TPT's **4.17%**. This reflects the heterogeneity a single scalar cannot address: distribution-shift miscalibration is *sample-heterogeneous*, so a single temperature that fixes one domain over- or under-softens others. RA-TPT instead adapts confidence per sample, achieving consistent calibration across domains rather than a domain-specific compromise.

Table 26: **Calibration Comparison: Temperature Scaling vs. RA-TPT.** Expected Calibration Error (ECE % ↓) across nine standard datasets with CLIP-ViT-B/16. Accuracy is omitted as global temperature scaling does not change the zero-shot classification accuracy; RA-TPT accuracy is reported in Table 2. **Bold** = lower ECE.

Method	DTD	UCF101	Aircraft	Pets	Caltech	Cars	Food	Flowers	EuroSAT	Average
Zero-Shot + TS (T=1.2)	**4.29**	6.25	**2.53**	8.19	9.74	10.02	6.14	7.00	5.15	6.59
RA-TPT (Ours)	7.87	**2.24**	5.33	**3.19**	**4.58**	**4.21**	**2.36**	**3.45**	**4.28**	**4.17**

O Reliability Analysis

O.1 Extended Reliability Analysis: ViT-B/16

To evaluate the calibration behavior of our framework, we provide reliability diagrams for the ViT-B/16 backbone across four distinct evaluation settings: ImageNet-V2 (natural distribution shift), Food101 (fine-grained object recognition), EuroSAT (a specialized remote-sensing domain), and Oxford-Pets (fine-grained animal classification). These diagrams plot expected sample accuracy against predicted confidence across 20 uniform bins. Perfect calibration is the dashed diagonal ($y = x$); the red hatched regions denote the calibration gap (overconfidence, where predictive confidence exceeds empirical accuracy).

Overconfidence in Baselines. The reliability diagrams for standard TPT show systematic overconfidence: the optimizer pushes predictions into the highest-confidence bins while their empirical accuracy falls short of the diagonal. This is most severe on the shifted and specialized domains (ImageNet-V2 and EuroSAT). O-TPT reduces these gaps through its geometric regularizers but remains miscalibrated on the same shifted domains, indicating that magnitude-agnostic geometric constraints only partially curb overconfidence under distribution shift. On the well-separated fine-grained domain (Oxford-Pets), by contrast, the baselines are already comparatively well calibrated, so the miscalibration to be corrected is concentrated in the shifted and specialized settings.

RA-TPT's Calibrated Adaptation. RA-TPT keeps the calibration error low across all four settings, closely tracing the ideal diagonal, with the clearest reductions in the calibration gap on the shifted and specialized domains (ImageNet-V2, EuroSAT) where the baselines remain overconfident; on the already well-calibrated Oxford-Pets it stays in the same low-error regime as the strongest baseline. This corroborates our diagnosis in the main text: by decoupling class separability from confidence and penalizing centered-logit dispersion ($\mathcal{L}_{\text{disp}}$), RA-TPT bounds predictive concentration under entropy-free adaptation, so the model reports high confidence mainly when the adapted representations support it.

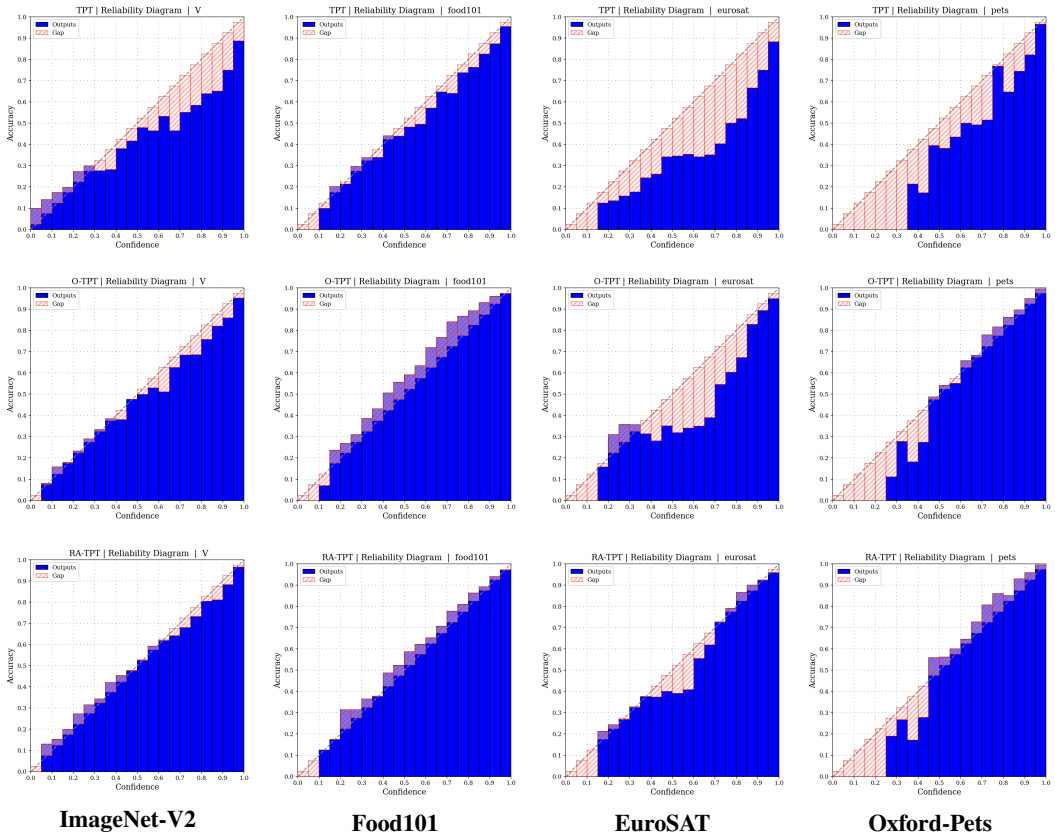


Figure 11: **ViT-B/16 Reliability Diagrams.** Evaluated across ImageNet-V2, Food101, EuroSAT, and Oxford-Pets. **Top Row:** Standard TPT. **Middle Row:** O-TPT. **Bottom Row:** RA-TPT (Ours). The red hatched area is the calibration gap (overconfidence). TPT and O-TPT are overconfident on the shifted and specialized domains (ImageNet-V2, EuroSAT); RA-TPT bounds the logit dispersion and keeps confidence aligned with accuracy across all four settings, with the largest gains where the baselines are most miscalibrated.

O.2 Extended Reliability Analysis: ResNet-50

To show that RA-TPT's calibration behavior is not specific to Transformers, we replicate the analysis on a convolutional backbone (ResNet-50). Figure 12 shows calibration across four distinct domains: Oxford Flowers-102 (fine-grained botany), Oxford-IIIT Pets (fine-grained animals), UCF101 (human action recognition), and ImageNet-R (artistic and synthetic renditions).

Failure of Baselines on Convolutional Features. CNNs have different inductive biases and feature geometries than Vision Transformers. As shown in the top row of Figure 12, the pathology of standard TPT persists: unconstrained entropy minimization drives logit inflation and severe overconfidence (large red hatched regions) on Flowers-102, UCF101, and ImageNet-R. O-TPT (middle row) reduces these gaps but stays miscalibrated on the shifted renditions (ImageNet-R) and action domain (UCF101), indicating that enforcing orthogonality on the textual prompts ($\mathbf{W}$) does not fully cure overconfidence when the visual features ($\mathbf{v}$) come from a CNN. On the fine-grained Oxford-Pets domain the baselines are already moderately calibrated, so the correction needed is again concentrated in the shifted settings.

Architecture-Agnostic Stability of RA-TPT. The bottom row of Figure 12 shows the generality of our approach. RA-TPT keeps calibration error low across the four domains and follows the $y = x$ diagonal most closely on Flowers-102, UCF101, and ImageNet-R, the domains where the baselines remain overconfident. Because RA-TPT operates in the logit space ($\mathbf{z}$) via the adaptive margin and uncertainty-routed dispersion regularization, it is largely invariant to the upstream spatial-extraction mechanism of the backbone. Whether the logits come from the self-attention maps of a ViT or the hierarchical convolutions of a ResNet-50, RA-TPT bounds the attainable maximum confidence, indicating that calibrated adaptation does not rely on architecture-specific heuristics.

P Logit Magnitude and Error Density Analysis

P.1 Internal Dynamics and Overconfidence: ViT-B/16

To further demystify the internal dynamics that drive calibration collapse in standard Test-Time Prompt Tuning (TPT), and to validate the corrective mechanisms of RA-TPT, we analyze both the post-adaptation top-two logit magnitudes across the test stream and the confidence distribution of the model's incorrect predictions. The analyses presented below utilize the ViT-B/16 backbone across three diverse datasets: Flowers-102, DTD, and Stanford Cars.

Preventing Unbounded Logit Inflation. The top row of Figure 13 shows the post-adaptation magnitude of the top-two pre-softmax logits for each test image; because prompts are reset after every image, the horizontal axis is the test-sample index, not a single-sample optimization trajectory. Across the test stream, standard TPT and O-TPT consistently produce heavily inflated top-two logit magnitudes. This empirical observation directly supports our theoretical claim: because standard entropy minimization lacks an absolute structural bound, the optimizer aggressively scales up the logit norm to artificially sharpen the output distribution. In contrast, RA-TPT operates with dynamically bounded logit dispersion, consistently restricting the top-two pre-softmax magnitude to the lowest levels among the compared methods, indicating that our uncertainty-routed dispersion regularizer substantially reduces optimization-induced inflation.

Reducing the Danger Zone (Confidence on Errors). The ultimate metric for deployment safety is not just overall accuracy, but whether a model knows when it is wrong. The bottom row of Figure 13 plots the density of errors as a function of softmax confidence. We explicitly highlight the "Danger Zone" (softmax confidence > 0.85), which represents severe failures: instances where the model makes an incorrect prediction with extreme, absolute certainty.

Standard TPT exhibits a dangerous pathology, placing a large density of its incorrect predictions in the > 0.85 Danger Zone: because it forcibly minimizes entropy regardless of a sample's ambiguity, it becomes highly confident even when wrong. RA-TPT substantially reduces this failure mode. By decoupling the margin from logit dispersion, RA-TPT shifts the error-density curve toward lower confidence, so that its incorrect predictions concentrate in the lower-confidence regime (e.g., 0.2 to 0.5) and the mass in the Danger Zone is sharply reduced. This makes RA-TPT's confidence scores more usable for triggering human fallback when the model is uncertain.

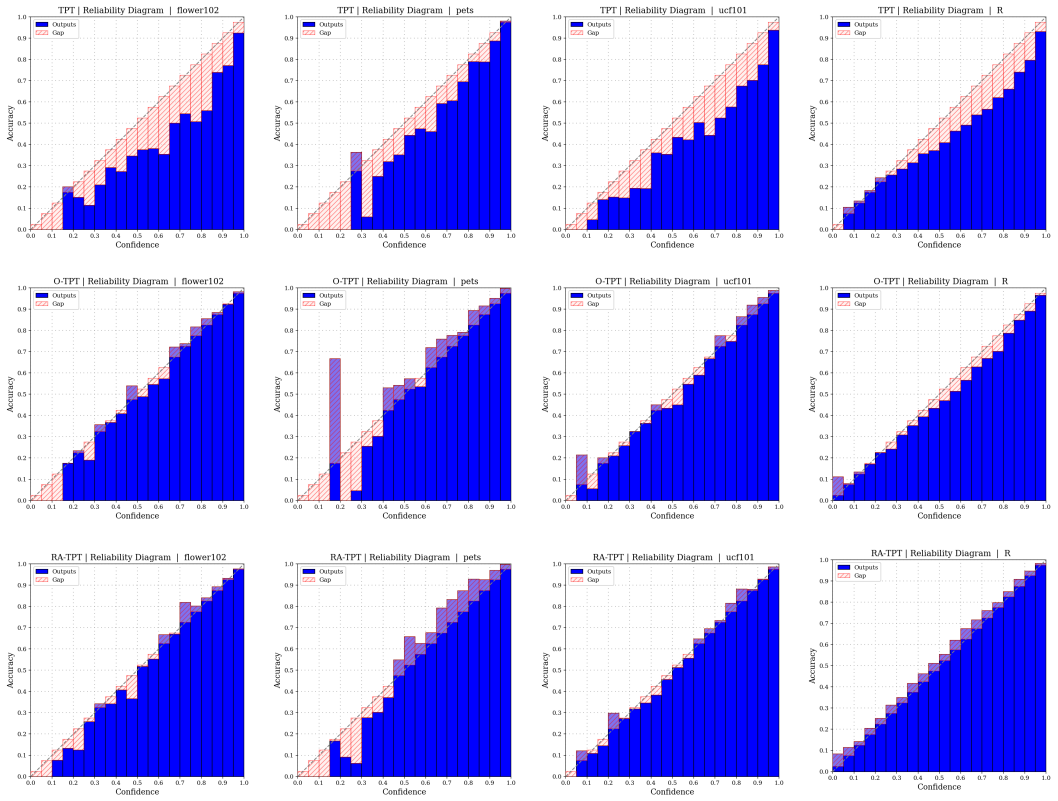
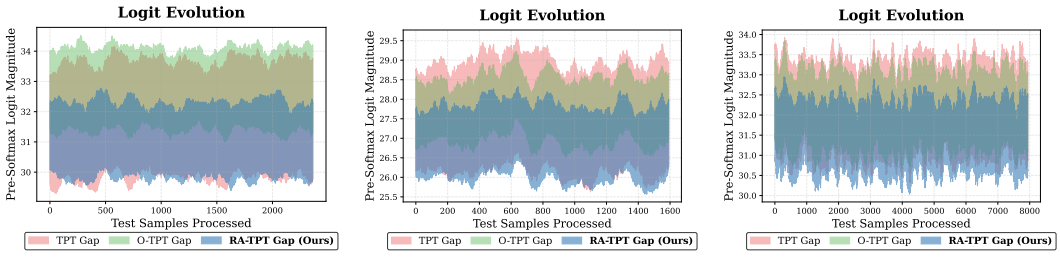
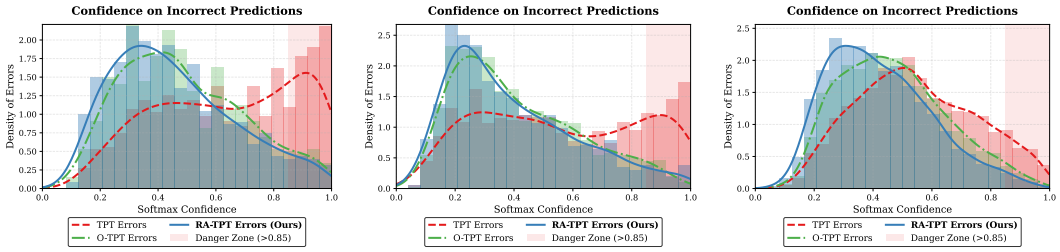


Figure 12: **ResNet-50 Reliability Diagrams.** Evaluated across Flowers-102, Oxford-Pets, UCF101, and ImageNet-R. **Top Row:** Standard TPT. **Middle Row:** O-TPT. **Bottom Row:** RA-TPT (Ours). The red hatched area is the calibration gap. TPT and O-TPT are overconfident on Flowers-102, UCF101, and ImageNet-R (driven by convolutional logit inflation); RA-TPT structurally bounds the logit spread and follows the diagonal most closely on these shifted domains, while all methods are comparably calibrated on the easier Oxford-Pets.



(a) Flowers-102 (b) DTD (c) Stanford Cars



(d) Flowers-102 (e) DTD (f) Stanford Cars

Figure 13: **Internal Dynamics and Safety (ViT-B/16). Top Row (a-c):** Post-adaptation top-two pre-softmax logit magnitude per test image (prompts reset per image; horizontal axis is the sample index). RA-TPT bounds the magnitude, preventing the unbounded inflation seen in TPT and O-TPT. **Bottom Row (d-f):** Confidence on incorrect predictions. TPT produces a large density of high-confidence errors (the Danger Zone, > 0.85). RA-TPT shifts the error distribution to lower confidence levels, supporting safer failure modes.

P.2 Internal Dynamics and Overconfidence: ResNet-50

To examine whether the mechanics of calibration collapse and RA-TPT's corrections generalize across architectures, we replicate our internal-dynamics analysis using the ResNet-50 backbone. Figure 14 shows the post-adaptation top-two logit magnitudes and error confidence distributions across six diverse datasets: FGVC-Aircraft, Stanford Cars, Caltech-101, DTD, Flowers-102, and UCF101.

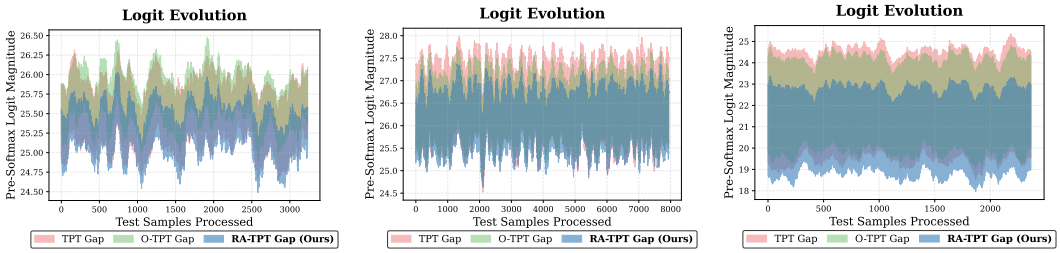
Consistent Bounding of Logit Magnitude. Across all six datasets, standard TPT and O-TPT suffer from runaway logit inflation (Rows 1 and 3 of Figure 14). RA-TPT consistently maintains the lowest top-two pre-softmax logit magnitudes, indicating that our dispersion regularizer effectively regulates convolutional features just as it does transformer self-attention maps.

A Consistent Pattern: Inflation and Overconfidence. Across these six domains a consistent empirical pattern emerges: greater unconstrained logit inflation is associated with a higher density of high-confidence errors in the Danger Zone (softmax confidence > 0.85). Structured domains with less zero-shot ambiguity (e.g., Stanford Cars) tend to induce milder inflation and a smaller high-confidence error mass, whereas texture-heavy or highly ambiguous domains (e.g., DTD, Flowers-102) push the unconstrained baselines into larger inflation and denser high-confidence errors.

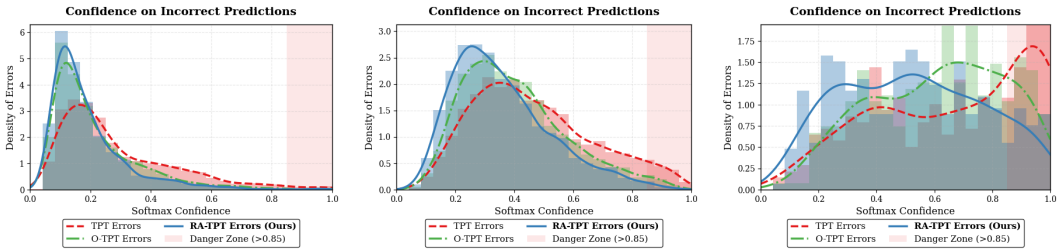
Regardless of a dataset's baseline ambiguity, RA-TPT shifts the error distribution toward lower confidence. By constraining logit dispersion, it weakens this association, reducing the Danger Zone and helping the model signal uncertainty when its prediction is wrong.

P.3 Internal Dynamics under Natural Distribution Shifts

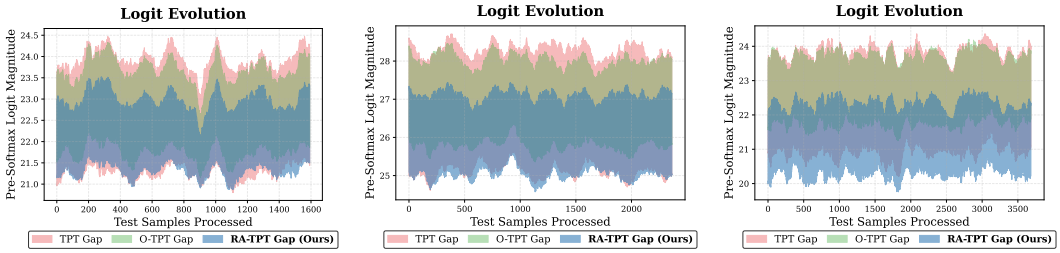
To further validate the robustness of RA-TPT's calibration mechanisms, we extend our internal dynamics analysis to severe natural distribution shifts. Figure 15 visualizes the post-adaptation top-two logit magnitudes and error confidence distributions across three challenging ImageNet variants: ImageNet-A (natural adversarial examples), ImageNet-R (artistic renditions), and ImageNet-V2 (reproduced test set) with ResNet-50 backbone.



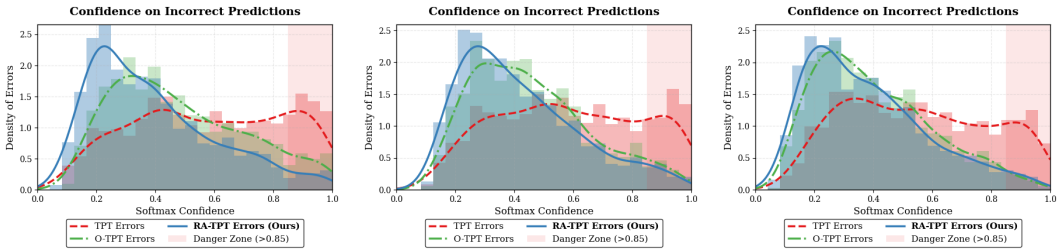
(a) Aircraft (Logit) (b) Stanford Cars (Logit) (c) Caltech-101 (Logit)



(d) Aircraft (Errors) (e) Stanford Cars (Errors) (f) Caltech-101 (Errors)



(g) DTD (Logit) (h) Flowers-102 (Logit) (i) UCF101 (Logit)

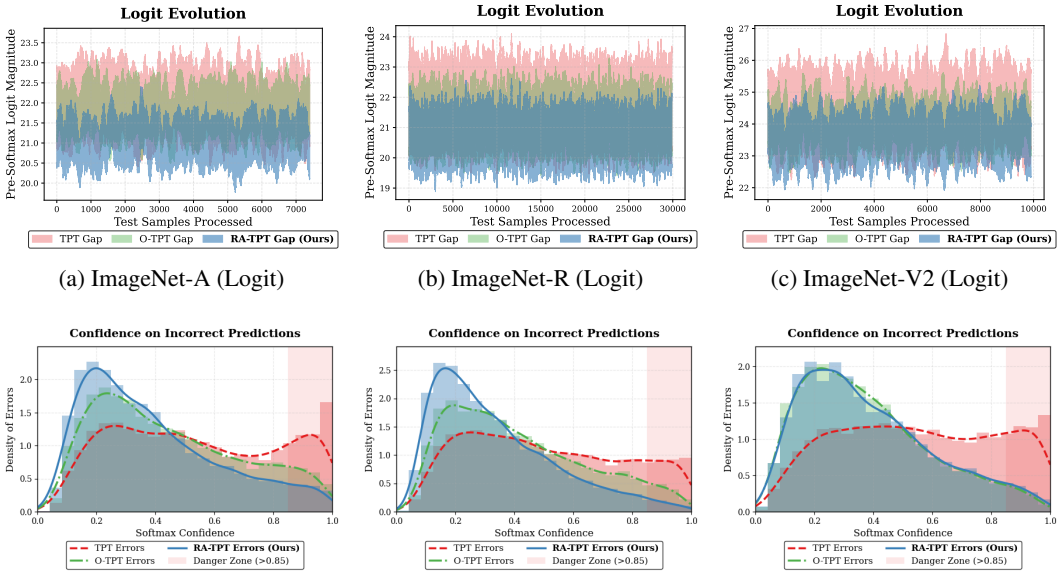


(j) DTD (Errors) (k) Flowers-102 (Errors) (l) UCF101 (Errors)

Figure 14: **Internal Dynamics and Safety (ResNet-50).** Post-adaptation top-two logit magnitude and error density across six datasets. **Rows 1 & 3:** RA-TPT bounds logit inflation compared to the runaway scaling of TPT/O-TPT. **Rows 2 & 4:** This bounding reduces the mass in the > 0.85 Danger Zone (red shaded area). Notably, texture-heavy domains (DTD, Flowers) show larger baseline inflation and Danger Zone density than more structured domains, consistent with an association between logit magnitude and high-confidence errors.

As illustrated in the top row of Figure 15, the baseline methods (TPT and O-TPT) continue to exhibit unconstrained logit inflation on these challenging distribution shifts. Consequently, this unchecked magnitude scaling populates the critical > 0.85 Danger Zone with highly confident but incorrect predictions (bottom row).

RA-TPT bounds the top-two pre-softmax logit magnitude across all three ImageNet variants, preventing the runaway scaling seen in the baselines. By constraining logit dispersion during adaptation, RA-TPT shifts the confidence distribution of incorrect predictions to the left, sharply reducing the Danger Zone. This indicates that our framework maintains consistent reliability and safer failure modes even when the model encounters adversarial or artistically shifted natural images.



(a) ImageNet-A (Logit) (b) ImageNet-R (Logit) (c) ImageNet-V2 (Logit)

(d) ImageNet-A (Errors) (e) ImageNet-R (Errors) (f) ImageNet-V2 (Errors)

Figure 15: **Internal Dynamics on Distribution Shifts.** Post-adaptation top-two logit magnitude and error density evaluated on ImageNet-A, ImageNet-R, and ImageNet-V2. **Top Row:** RA-TPT bounds logit magnitude across the test stream under natural distribution shifts. **Bottom Row:** By constraining logit dispersion, RA-TPT reduces the accumulation of high-confidence errors in the > 0.85 Danger Zone (red shaded area).

Q Impact of Multiple Adaptation Steps on Calibration

Standard Test-Time Prompt Tuning (TPT) relies on unconstrained entropy minimization. While a single gradient step ($T = 1$) often yields gains, taking multiple sequential steps on the same instance can drive severe logit inflation: the optimizer repeatedly sharpens the prediction to maximize confidence, degrading calibration. RA-TPT is designed and deployed as a single-step method ($T = 1$); here we probe its behavior under repeated updates as a robustness check rather than a deployment setting.

We evaluate 1 to 5 adaptation steps on ImageNet-V2, ImageNet-A, and ImageNet-R with the ResNet-50 backbone, comparing against standard TPT and O-TPT.

Empirical observations. At a single step ($T = 1$), RA-TPT is the best-calibrated method on all three datasets, and it remains the best at every step count (Table 27). However, RA-TPT is not immune to multi-step drift: over five steps its ECE rises on ImageNet-V2 ($3.01 \rightarrow 10.57$) and ImageNet-R ($2.52 \rightarrow 10.40$). The key difference from the baselines is the rate. TPT and O-TPT collapse catastrophically over the same range (ImageNet-A: TPT $31.04 \rightarrow 64.04$, O-TPT $24.25 \rightarrow 46.05$; ImageNet-V2: TPT $\rightarrow 39.83$), roughly an order of magnitude larger than RA-TPT's drift, whereas on ImageNet-A RA-TPT instead *improves* step-over-step ($18.66 \rightarrow 11.31$). We therefore do not claim monotone multi-step improvement: on ImageNet-V2 and ImageNet-R the ECE does rise, but it stays small and bounded rather than diverging, which is consistent with our single-step ($T = 1$) deployment choice.

Why the drift stays bounded. Two coupled dynamics explain the residual increase on ImageNet-V2 and ImageNet-R. First, once the bounded margin $\mathcal{L}_{\text{margin}}$ saturates at its cap (Claim 3.1), the self-training anchor $\mathcal{L}_{\text{anchor}} = -\log \bar{p}_{k^*}$ continues to concentrate the prediction, inducing a slow residual drift in the logit dispersion; the uncertainty-routed dispersion penalty $\mathcal{L}_{\text{disp}}$ opposes this drift but does not eliminate it under repeated updates (see the analysis in Appendix B). Second, multiple steps on a single instance create a pseudo-label feedback loop: if the top-1 k^* is incorrect but the model is already confident (low ω), the anchor dominates the weakly weighted dispersion penalty (ω^2 small) and reinforces the wrong label, slowly raising ECE on the misclassified subpopulation. Unlike standard TPT, where this loop diverges (e.g., 39.83% on ImageNet-V2), RA-TPT's dispersion penalty keeps the drift far smaller.

Dataset divergence: escaping the feedback loop on ImageNet-A. The opposite trend on ImageNet-A illustrates how instance-adaptive routing (ω) counteracts this confirmation bias on genuinely hard samples. ImageNet-A consists of naturally adversarial images on which the zero-shot model is highly uncertain, so ω stays high across all T steps. Under our fixed setting $(\alpha, \lambda) = (4, 1.5)$, the margin cap $m_{\text{cap}} = \max(0, \alpha - \lambda\omega)$ then contracts toward its lower bound $\alpha - \lambda = 2.5$ rather than vanishing; the $m_{\text{cap}} \to 0$ row of Table 1 is a conceptual boundary of the general mechanism (it would require $\lambda \geq \alpha$), not a regime our configuration enters. The bounded margin term therefore saturates early and, with the anchor's concentration pressure down-weighted at high ω, is outweighed by the strong ω^2-weighted dispersion penalty $\mathcal{L}_{\text{disp}}$, which reduces predictive concentration. Consequently, on the hardest shift, where standard methods collapse most severely, RA-TPT reduces ECE step-over-step rather than inflating it.

Table 27: **Effect of Multiple TTA Steps on Accuracy and Calibration.** Evaluated on ResNet-50. Standard methods suffer severe ECE inflation as steps increase; RA-TPT keeps the drift small and bounded (and improves on ImageNet-A). RA-TPT stays the best-calibrated method at every step, and is deployed at $T = 1$; multiple steps are shown as a robustness probe.

Dataset	Method	Step 1		Step 2		Step 3		Step 4		Step 5	
		Acc. (↑)	ECE (↓)	Acc. (↑)	ECE (↓)	Acc. (↑)	ECE (↓)	Acc. (↑)	ECE (↓)	Acc. (↑)	ECE (↓)
ImageNet-V2	TPT	54.58	13.18	54.26	24.07	54.53	31.34	54.15	36.71	54.20	39.83
	O-TPT	53.20	3.82	53.45	11.19	53.63	15.09	52.87	17.51	52.72	17.85
	RA-TPT	52.00	**3.01**	51.83	**3.54**	51.95	**6.28**	51.63	**8.34**	51.74	**10.57**
ImageNet-A	TPT	25.20	31.04	26.23	44.49	27.04	53.69	27.25	59.89	27.13	64.04
	O-TPT	22.60	24.25	23.34	35.75	23.54	40.91	23.17	43.32	23.09	46.05
	RA-TPT	24.10	**18.66**	22.42	**17.36**	22.45	**14.86**	21.96	**12.69**	21.76	**11.31**
ImageNet-R	TPT	59.10	9.12	58.80	20.53	58.69	28.19	58.45	33.17	58.43	36.09
	O-TPT	57.10	2.58	56.13	12.76	55.84	18.00	55.87	21.32	56.15	23.81
	RA-TPT	57.00	**2.52**	57.22	**4.74**	57.15	**6.89**	57.00	**8.66**	57.04	**10.40**

R Extended Qualitative Analysis

To visually complement our quantitative results, we provide an extended qualitative analysis of Test-Time Prompt Tuning (TPT), Orthogonality-constrained TPT (O-TPT), and our Reliability-Aware formulation (RA-TPT). The cases below are representative illustrations rather than statistical evidence: their purpose is to make the failure modes of Figure 1(d) concrete and to show that RA-TPT's dispersion control behaves consistently across different visual domains and architectures. Aggregate calibration results are reported in the main tables and Appendix D.

R.1 Fine-Grained Action Recognition (UCF-101 / ResNet-50)

A common limitation of test-time adaptation methods is architecture dependency. To illustrate that our ω^2 uncertainty router behaves consistently across both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), we extend the qualitative analysis to fine-grained action recognition on UCF-101 using a ResNet-50 backbone (Figure 16).

Action recognition introduces temporal and spatial ambiguities not present in static natural imagery. In the top row of Figure 16, standard TPT struggles to distinguish fine-grained subclasses (e.g., misclassifying a "Punching Bag" as a "Speed Bag" at 81.7% confidence). Both O-TPT and RA-TPT recover the correct class, though RA-TPT reports a better-calibrated confidence.

The distinction is clearer in the bottom row. Under severe visual ambiguity ("Mixing" vs. "Cutting"), TPT's unconstrained entropy minimization produces a near-absolute confidence (96.3%) in an incorrect prediction. O-TPT reduces this confidence only moderately and remains overconfident. RA-TPT dampens the logit vector under this ambiguity, yielding a lower-confidence prediction (27.1%) that more appropriately signals uncertainty.



Figure 16: **Extended Qualitative Analysis (UCF-101 / ResNet-50).** RA-TPT recovering fine-grained action classes while maintaining appropriate uncertainty under ambiguity.

R.2 Semantic and Structural Distortion in Fine-Grained Classification

A common approach to preventing feature collapse during test-time adaptation is enforcing orthogonal constraints on the prompt embedding matrix $\mathbf{W}$ (e.g., O-TPT). We hypothesize that enforcing $\mathbf{W}\mathbf{W}^\top \to \mathbf{I}$ can distort the semantic space: fine-grained classification involves overlapping semantic geometries, and forcing closely related concepts toward orthogonality may artificially separate them, which can exacerbate misclassification and leave the model poorly calibrated.

To illustrate this hypothesis across different types of visual entanglement, we evaluate RA-TPT against standard TPT and O-TPT on a ResNet-50 backbone across two fine-grained domains: organic subjects (Oxford-IIIT Pet) and rigid manufactured structures (FGVC-Aircraft).

Organic Subjects (Oxford-IIIT Pet) We first examine organic, deformable subjects. As shown in Figure 17, in these examples O-TPT does not fully regularize logit inflation, and under high ambiguity the orthogonal constraint can shift the prediction to an unrelated class (e.g., from Russian Blue to Abyssinian). By bounding the logit dispersion rather than constraining the embedding geometry, RA-TPT corrects these errors without inducing such shifts.

Rigid Geometric Structures (FGVC-Aircraft) To check that our uncertainty-routed regularizer does not overfit to natural textures, we extend the evaluation to FGVC-Aircraft. This domain is challenging due to small inter-class structural differences (e.g., near-identical fuselages differing only by engine cowlings), so the ideal class embeddings of related variants naturally cluster together.

As shown in Figure 18, the effect of orthogonal constraints is visible in this example. While O-TPT reduces the extreme overconfidence of standard TPT, its warped latent space leaves the model on the incorrect aircraft variant (Figure 18, Top Left) and it does not recover the correct structural geometry. RA-TPT corrects the error and routes the features back to the correct variant.

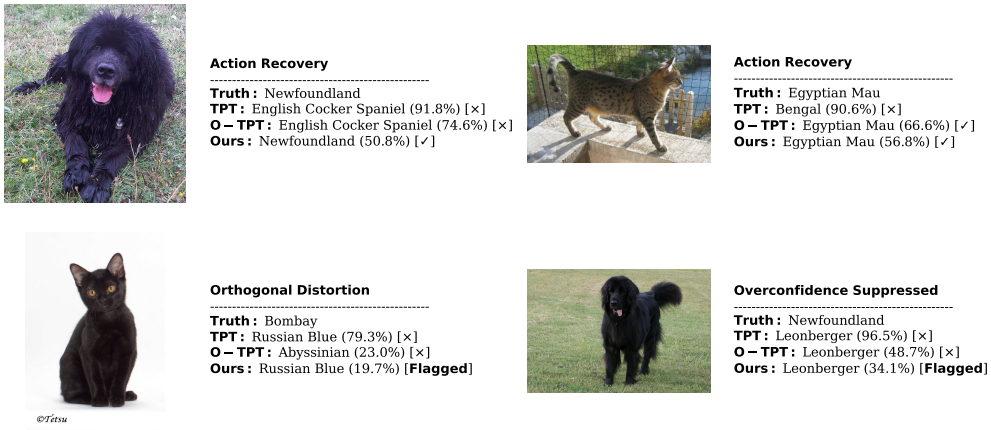


Figure 17: **Architecture-Agnostic Recovery and Semantic Distortion (Oxford-IIIT Pet / ResNet-50). Top Row:** On visually entangled categories, unconstrained TPT reaches severe overconfidence ($> 90\%$). While O-TPT occasionally recovers the class (Top Right), it sometimes fails to correct the trajectory (Top Left). RA-TPT recovers the correct class while maintaining calibrated confidence. **Bottom Row:** Under severe ambiguity, enforcing orthogonality ($\mathbf{W}\mathbf{W}^\top \to \mathbf{I}$) in O-TPT either mirrors the original failure (Bottom Right) or shifts the prediction to an unrelated class (Bottom Left). RA-TPT avoids this shift and corrects the logit inflation.

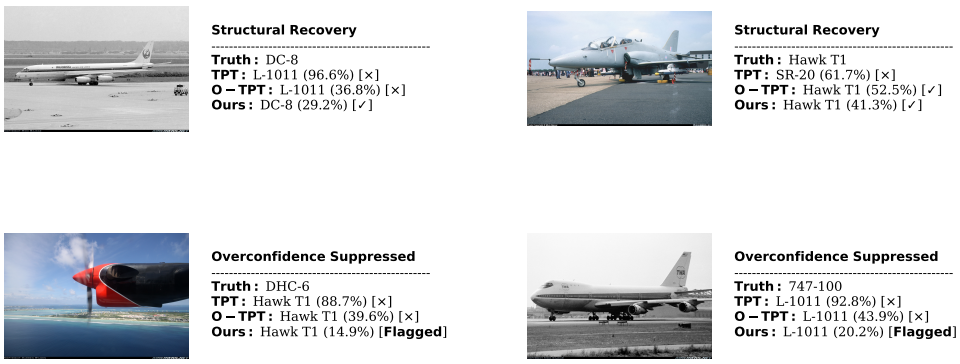


Figure 18: **Structural Distortion and Recovery (FGVC-Aircraft / ResNet-50).** We compare RA-TPT against TPT and O-TPT on rigid, manufactured geometries. **Top Row:** On highly similar aircraft variants, TPT reaches extreme overconfidence (96.6%) on incorrect classes. While O-TPT reduces this confidence (Top Left), its orthogonal constraint appears to distort the latent space, leaving the model misclassified. RA-TPT recovers the correct aircraft variant while maintaining calibrated confidence. **Bottom Row:** Under severe ambiguity, both methods suppress TPT's logit inflation, but RA-TPT pushes the prediction to a lower, better-calibrated confidence ($< 25\%$).