CONCENTRATION-INVARIANT GEODESIC FUSION FOR CALIBRATED TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Backpropagation-free test-time adaptation (TTA) provides an efficient alternative for adapting pre-trained vision-language models under distribution shift; however, the underlying causes of miscalibration in these methods remain poorly understood. In this paper, we propose CRAFT, a calibration-aware adaptation framework that retains the efficiency of the backpropagation-free paradigm. We attribute the accuracy-calibration trade-off to a phenomenon we term *Geometric Collapse*: the unnormalized Euclidean fusion of textual anchors and visual prototypes departs from the unit hypersphere upon which the model was pre-trained, inflating the concentration unevenly across classes and resulting in overconfident predictions. To address this, CRAFT performs logit fusion along the geodesic (SLERP) on the unit hypersphere, which preserves the pre-trained concentration geometry, with a tuning-free instance-level weighting. Subsequently, a Fisher-Rao retraction on the probability simplex caps the Fisher-Rao concentration radius of the fused prediction at the VLM's inherent zero-shot level, mitigating overconfidence without requiring labeled data or tuned thresholds. Across 15 image recognition benchmarks encompassing cross-dataset benchmarks and natural distribution shifts, CRAFT consistently reduces the average Expected Calibration Error (ECE) while maintaining competitive accuracy.

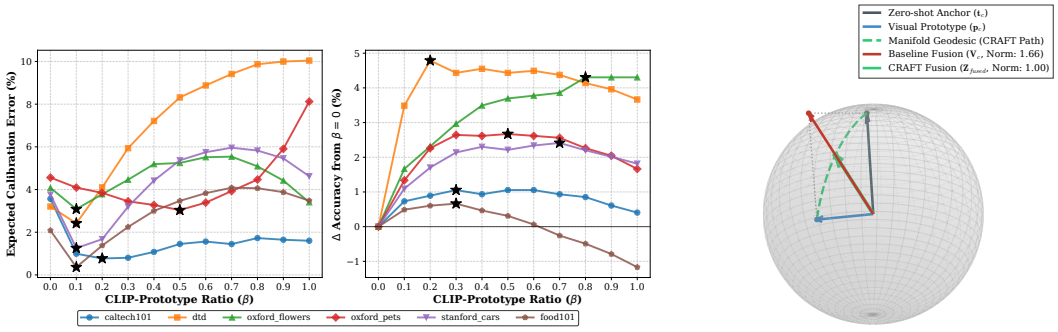
1 INTRODUCTION

Vision-Language Models (VLMs) such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) enable remarkable zero-shot recognition by contrastively aligning images and text on a shared d-dimensional unit hypersphere, where decision boundaries are defined entirely by the cosine similarity between text and image features. Under distribution shift, however, this static zero-shot boundary often fails to align with the test distribution (Shu et al., 2022; Wortsman et al., 2022), degrading not only classification accuracy but also the reliability of confidence estimates.

To adapt without the prohibitive cost of backpropagation-based fine-tuning, backpropagation-free Test-Time Adaptation (TTA) has emerged as a highly efficient alternative: pseudo-labeled test features are accumulated into online caches or prototypes (Karmanov et al., 2024; Guan et al., 2025; Wang et al., 2025; Huang et al., 2026), and the resulting prototype logits are combined with the zero-shot logits by a simple unnormalized Euclidean fusion (Guan et al., 2025; Huang et al., 2026) to keep the process latency-free. This design boosts accuracy but quietly breaks reliability. Across TTA evaluations, current methods (Zhang et al., 2024; Karmanov et al., 2024; Guan et al., 2025; Huang et al., 2026) consistently degrade Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017), the standard measure of the gap between predicted confidence and accuracy, even as target-domain accuracy improves: the adapted classifiers become confidently wrong. We term the structural cause of this trade-off *Geometric Collapse*.

In this paper, we propose CRAFT, a **c**alibration-aware **r**eliable **a**daptation **f**ramework for **t**est time, motivated by a geometric analysis of this ECE collapse. Viewing the contrastive latent space through the von Mises-Fisher (vMF) distribution, whose concentration parameter is the inverse pre-trained temperature ($\kappa = 1/\tau$), we observe that unnormalized Euclidean addition pushes the fused vector off the unit hypersphere, and its magnitude grows as the visual prototype aligns with the text anchor at test time, sharpening the predictive distribution. The overconfidence is thus not an incidental artifact of adaptation but a structural consequence of unconstrained fusion, as illustrated in Figure 1b.

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) Accuracy-Calibration Trade-off in TTA

(b) Geometric visualization

Figure 1: **Calibration error in Test-Time Adaptation and the proposed geometric solution. (a)** Increasing the logit fusion weight (β) generally improves accuracy but simultaneously degrades ECE. The stars indicate that the optimal β for minimum ECE and maximum accuracy varies across datasets, showing the difficulty of relying on a static threshold. **(b)** Geometric visualization on the unit hypersphere. Baseline Euclidean fusion (red) increases the vector norm, while CRAFT (green) uses geodesic interpolation to maintain the unit norm.

Building on this observation, CRAFT reformulates logit fusion as Hyperspherical Concentration-Invariant Fusion (HCIF), a geodesic (SLERP) (Shoemake, 1985) interpolation that keeps the fused representation on the pre-trained unit hypersphere. Since the optimal fusion weight differs across datasets (Figure 1a), the interpolation weight is set per instance from a scale-invariant statistic of the logits, with no dataset-specific tuning. Finally, an Information-Geometric Fisher-Rao Retraction (IGFR) (Das et al., 2025; Pavesi et al., 2026) on the probability simplex caps the Fisher-Rao concentration radius of the fused prediction at the zero-shot reference without altering class rankings. Together, these components restore calibration while preserving the single-forward-pass efficiency of backpropagation-free adaptation. The main contributions of this paper are:

- **We attribute the Calibration Trap** in training-free TTA to a geometric mechanism: unnormalized Euclidean logit fusion inflates the fused representation norms asymmetrically across classes. This uneven norm inflation acts as a class-dependent logit scale distortion, artificially sharpening the predictive distribution and inducing overconfidence.
- **We propose CRAFT**, which replaces additive fusion with a geodesic (SLERP) fusion that preserves the hyperspherical geometry of the pre-trained embedding space, paired with a Fisher-Rao retraction on the probability simplex that limits the Fisher-Rao distance from the uniform distribution to the zero-shot radius without altering class rankings.
- **CRAFT improves reliability at negligible cost.** Across 15 benchmarks, CRAFT consistently reduces average ECE on top of existing backpropagation-free TTA methods while maintaining comparable accuracy and single-forward-pass efficiency, and requires no dataset-specific tuning of the fusion weight.

2 Related Work

Test-Time Adaptation for Vision-Language Models Test-time adaptation (TTA) adjusts pre-trained models to distribution shifts at inference without labeled data (Wang et al., 2020). For VLMs, TPT (Shu et al., 2022) pioneered this paradigm by optimizing textual prompts via entropy minimization, and subsequent works (Feng et al., 2023; Hassan et al., 2023) refined this direction, but all require test-time backpropagation, limiting their use in latency-sensitive settings. Backpropagation-free frameworks avoid this cost by accumulating test-time knowledge into caches or prototypes: TDA (Karmanov et al., 2024) maintains key-value caches of confident predictions, and PTA (Huang et al., 2026) updates class-wise prototypes with a confidence-weighted EMA. Both, however, fuse the zero-shot and prototype logits by unnormalized Euclidean addition, which, as we analyze in Sec. 4.1, leaves the unit hypersphere and degrades calibration even as accuracy improves. CRAFT addresses this at the fusion step while keeping the efficient prototype paradigm.

Calibration in Test-Time Adaptation Beyond accuracy, a line of work improves the calibration of backpropagation-based TTA, where entropy minimization causes overconfidence: C-TPT (Yoon et al., 2024) disperses text features, O-TPT (Sharifdeen et al., 2025) enforces orthogonality between them, SoC (Fillioux et al., 2026) relaxes this with a Huber-based regularizer, and FPP (Jang et al., 2026) pretrains prompts toward flat minima. All attach auxiliary losses to gradient-based prompt optimization, and thus do not transfer to the backpropagation-free setting, where calibration has received little attention. CRAFT fills this gap geometrically: it locates the miscalibration in the fusion step, where unnormalized addition induces class-dependent logit scale distortions, and restores the constant-norm geometry by geodesic (SLERP) (Shoemake, 1985) fusion, without gradients or labeled data.

3 PRELIMINARIES

3.1 CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING

Contrastive vision-language models (VLMs) like CLIP (Radford et al., 2021) achieve effective zero-shot visual understanding by aligning visual and textual representations in a shared latent embedding space. CLIP consists of a visual encoder E_V and a text encoder E_T, which are jointly trained to maximize the cosine similarity of matching image-text pairs while minimizing it for non-matching pairs across a batch. For a downstream classification task with C classes, zero-shot prediction synthesizes a linear classifier from natural language. Each class name is embedded into a handcrafted prompt template (e.g., "A photo of a {label}.") and passed through the text encoder to obtain class-specific text embeddings $\{F_t^c\}_{c=1}^C \subset \mathbb{R}^d$, while the visual encoder extracts the feature $F_g = E_V(x) \in \mathbb{R}^d$ from an input image x. The zero-shot probability for class y_c is then the softmax over cosine similarities scaled by the pre-trained temperature τ:

$$p_{\text{clip}}(y_c|x) = \frac{\exp(\cos(F_g, F_t^c)/\tau)}{\sum_{k=1}^{C} \exp(\cos(F_g, F_t^k)/\tau)}. \tag{1}$$

3.2 EXPECTED CALIBRATION ERROR (ECE)

Model calibration measures how well a model's predicted probabilities align with its empirical accuracy: a perfectly calibrated model is correct on exactly p of the samples it predicts with confidence p. Expected Calibration Error (ECE) (Naeini et al., 2015) is widely adopted as a standard metric for this alignment. Given a dataset of N samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, let $\hat{y}_i$ denote the predicted label and $\hat{p}_i$ the associated confidence for the i-th sample, obtained from Equation 1 as

$$\hat{y}_i = \arg\max_{y_c} p_{\text{clip}}(y_c|x_i), \quad \hat{p}_i = \max_{y_c} p_{\text{clip}}(y_c|x_i). \tag{2}$$

The N predictions are partitioned into M equally spaced confidence bins. With B_m the set of sample indices falling into the m-th bin, the accuracy and average confidence per bin are

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i), \quad \text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The ECE is the weighted average of the gap between accuracy and confidence across bins:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|. \tag{4}$$

4 PROPOSED METHOD

To mitigate the Calibration Trap pervasive in Test-Time Adaptation (TTA) without sacrificing the efficiency of backpropagation-free adaptation, we present the Calibration-aware Reliable Adaptation Framework for Test-time (CRAFT). In this section, we first diagnose the geometric failure point of the Euclidean fusion baseline under the von Mises-Fisher (vMF) distribution. Subsequently, we introduce a geodesic fusion mechanism that preserves the hyperspherical geometry of the pre-trained space, an instance-level weighting rule that removes dataset-specific tuning of the fusion weight, and a Fisher-Rao retraction that caps the Fisher-Rao concentration radius at the zero-shot level.

4.1 Geometric Collapse and Class-Dependent Logit Scaling

Although both the text anchor t_c and the visual prototype p_c are L_2-normalized unit vectors ($\|t_c\|_2 = \|p_c\|_2 = 1$), their additively fused evidence vector $V_c = t_c + p_c$ is not, and thus leaves the unit hypersphere on which the model was pre-trained. Given an input image feature x and a temperature parameter τ, expanding the scaled dot product yields

$$\frac{x \cdot V_c}{\tau} = \frac{x}{\tau} \cdot (t_c + p_c) = \frac{x \cdot t_c}{\tau} + \frac{x \cdot p_c}{\tau} = L_{\text{clip}} + L_{\text{proto}}, \tag{5}$$

demonstrating that this formulation simply sums the individually scaled similarities. While intuitive, this approach breaks the constant-norm assumption of the hyperspherical space. Expanding the L_2-norm of the fused vector gives

$$\|V_c\|_2 = \sqrt{2 + 2(t_c \cdot p_c)}. \tag{6}$$

This norm varies across classes. Since the cache for class c is filled with samples confidently pseudo-labeled as c, the prototype of a frequently predicted class tends to align well with its anchor, inflating $t_c \cdot p_c$ (we observe fused norms up to ≈ 1.66; Fig. 1b), whereas classes with sparse or noisy caches retain $t_k \cdot p_k \approx 0$ and $\|V_k\|_2 \approx \sqrt{2}$. This uneven norm expansion acts directly as a class-dependent logit scale, artificially amplifying the confidence of well-cached classes.

This asymmetric norm inflation is problematic: when mapping to the logit space, the effective concentration parameter becomes class-dependent, $\kappa_c = \|V_c\|_2/\tau$. Instead of maintaining a uniform temperature, the inflated κ_c of well-cached classes acts as an uneven, class-dependent logit scale. This distortion sharpens the predictive distribution beyond what the angular margins alone would justify. We view this as a plausible structural mechanism behind the overconfidence observed in Fig. 1b; a detailed analysis is given in Appendix A.1.

4.2 Hyperspherical Concentration-Invariant Fusion (HCIF)

To maintain the fused vectors on the unit hypersphere, we replace Euclidean fusion with **Spherical Linear Interpolation (SLERP)**. Let $\beta \in [0, 1]$ be an instance-level routing weight (Sec. 4.3). Let $\rho_c = t_c \cdot p_c$ be the class-wise text anchor–prototype alignment. For each class c, we calculate the specific interpolation angle $\Omega_c = \arccos(\rho_c)$. The fused evidence vector $V_c^*(\beta)$ is defined as

$$V_c^*(\beta) = \frac{\sin((1-\beta)\Omega_c)}{\sin(\Omega_c)} t_c + \frac{\sin(\beta\Omega_c)}{\sin(\Omega_c)} p_c. \tag{7}$$

By utilizing class-specific interpolation angles, this formulation ensures that the fused vector for each class satisfies $\|V_c^*(\beta)\|_2 = 1$. Since linear operations distribute over the dot product, this exact vector-space fusion reduces to a scalar combination applied element-wise in the logit space:

$$L_c^*(x) = w_{\text{text},c} L_{\text{clip},c} + w_{\text{proto},c} L_{\text{proto},c}, \qquad L_{\text{clip},c} = (x \cdot t_c)/\tau,\ L_{\text{proto},c} = (x \cdot p_c)/\tau, \tag{8}$$

where $w_{\text{text},c} = \frac{\sin((1-\beta)\Omega_c)}{\sin(\Omega_c)}$ and $w_{\text{proto},c} = \frac{\sin(\beta\Omega_c)}{\sin(\Omega_c)}$.

4.3 Routing Weight Setting via Gaussian Probabilistic Routing (GPR)

The remaining design choice is the interpolation weight β in equation 7. Entropy- or confidence-based criteria depend on the absolute scale of the logits, which shifts under domain change, so a threshold chosen on one dataset does not transfer to another. We therefore set β from a scale-invariant statistic of the logits, a rule we refer to as Gaussian Probabilistic Routing (GPR).

For each logit vector $L_m \in \{L_{\text{clip}}, L_{\text{proto}}\}$, let μ_m and σ_m denote the mean and standard deviation of its entries across the class dimension. We standardize the top-1 logit $L_m^* = \max(L_m)$ into a dimensionless Z-score,

$$Z_m = \frac{L_m^* - \mu_m}{\sigma_m}, \tag{9}$$

which is unchanged under any affine rescaling of the logits, so no dataset-specific temperature or threshold is involved. Each score is mapped to $[0, 1]$ through the Gaussian CDF Φ, and the two are combined into

$$\beta = \frac{\Phi(Z_{\text{proto}})}{\Phi(Z_{\text{clip}}) + \Phi(Z_{\text{proto}})}. \tag{10}$$

GPR is designed purely for scale invariance: Φ is only a smooth, monotone map to $[0, 1]$, with no distributional assumption on the logits. The source whose top logit stands out more sharply earns more weight; the goal is not accuracy gains but avoiding hand-tuning of β.

4.4 Information-Geometric Fisher-Rao Retraction (IGFR)

While the feature-level SLERP fusion in Sec. 4.2 keeps the fused vector close to the unit hypersphere, a second hazard persists: since prototypes are accumulated from the model's own confident predictions, the fused distribution can remain sharper than its accuracy warrants under distribution shift. Temperature scaling would fix this but requires labeled validation data. Instead, we treat the zero-shot prediction, which is empirically well calibrated, as an instance-wise reference for how confident the model is entitled to be. To compare the two predictions geometrically, we first map the fused logit L_c^* and the base CLIP logit L_{clip} to the probability simplex Δ^{C-1} via the Softmax operator, yielding P_{fused} and P_{clip}, and apply the square-root embedding, an isometry (up to a constant factor) between the probability simplex under the Fisher–Rao metric and the positive orthant of the unit hypersphere $\mathcal{S}_{\geq 0}^{C-1}$,

$$X_{\text{fused},c} = \sqrt{P_{\text{fused},c}}, \quad X_{\text{clip},c} = \sqrt{P_{\text{clip},c}} \quad \forall c \in \{1, \ldots, C\}. \tag{11}$$

Let $X_U \in \mathcal{S}_{\geq 0}^{C-1}$ denote the uniform distribution amplitude ($X_{U,c} = 1/\sqrt{C}$), the point of maximum entropy. Since the Fisher-Rao geodesic distance between two distributions corresponds (up to the same factor) to the angle between their amplitude vectors, we define the confidence radius r as the angular distance from X_U,

$$r_{\text{fused}} = \arccos(\langle X_U, X_{\text{fused}} \rangle), \quad r_{\text{clip}} = \arccos(\langle X_U, X_{\text{clip}} \rangle). \tag{12}$$

If $r_{\text{fused}} > r_{\text{clip}}$, the fused prediction is more concentrated than the zero-shot reference, and we regard the excess concentration as unwarranted. In that case, we retract X_{fused} along the great circle connecting it to X_U, stopping at radius r_{clip}: the retracted point is $X'_{\text{fused}} = \alpha X_U + \gamma X_{\text{fused}}$ with $\alpha = \frac{\sin(r_{\text{fused}} - r_{\text{clip}})}{\sin(r_{\text{fused}})}$ and $\gamma = \frac{\sin(r_{\text{clip}})}{\sin(r_{\text{fused}})}$, so that $\langle X_U, X'_{\text{fused}} \rangle = \cos(r_{\text{clip}})$ exactly; when $r_{\text{fused}} \leq r_{\text{clip}}$, the prediction is left unchanged. Since $\alpha, \gamma > 0$ and X_U is constant across classes, the retraction preserves the ordinal ranking of the probabilities: confidence is lowered without changing the Top-1 prediction, leaving accuracy untouched. Derivations of both properties are given in Appendix B.

After recovering the calibrated probability $P'_{\text{fused}} = (X'_{\text{fused}})^2$, we return to the logit space through the Aitchison Centered Log-Ratio (CLR) mapping,

$$L_{\text{final},c} = \ln(P'_{\text{fused},c}) - \frac{1}{C} \sum_{j=1}^{C} \ln(P'_{\text{fused},j}). \tag{13}$$

Since the softmax is invariant to constant shifts, any logarithmic representative yields the same probabilities; CLR is simply the canonical zero-mean choice. IGFR is deliberately one-sided, bounding the Fisher-Rao distance from the uniform distribution at the zero-shot radius and preventing unwarranted concentration increases, a favorable trade given the overconfidence-dominated errors of backpropagation-free TTA.

5 Experiments

We evaluate CRAFT across image recognition benchmarks, natural distribution shifts, and temporally correlated test streams, followed by component ablations and efficiency analysis.

5.1 Experimental Setting

Datasets and Metrics. We evaluate CRAFT on 15 image recognition benchmarks: ImageNet (Deng et al., 2009), 10 downstream datasets, together with four distribution shifts, ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a) and ImageNet-Sketch (Wang et al., 2019). For non-i.i.d. evaluation, we construct correlated label streams via a Dirichlet distribution with $\gamma \in \{0.1, 0.01, 0.001\}$ and Separate, where smaller γ means stronger correlation. We report Accuracy (%) and Expected Calibration Error (ECE, %).

Table 1: **Cross-dataset image recognition.** Accuracy and ECE on 11 benchmarks using CLIP ViT-B/16. † denotes results reproduced using the official implementation.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
Zero Shot†	Acc.	68.78	44.09	70.69	85.88	66.23	24.96	89.10	93.83	67.46	48.30	66.21	65.96
	ECE	1.52	3.45	4.18	2.09	1.65	3.99	4.84	3.53	3.17	5.93	3.75	3.46
Backpropagation-based													
TPT	Acc.	69.00	46.70	69.00	84.70	65.50	23.40	87.10	93.80	67.30	42.40	66.30	65.02
	ECE	10.60	21.20	13.50	3.98	11.30	16.80	5.77	4.51	13.00	21.50	5.16	11.57
C-TPT	Acc.	68.50	46.00	69.80	83.70	64.80	24.00	88.20	93.60	65.70	43.20	65.80	64.85
	ECE	3.15	11.90	5.04	3.43	5.04	4.36	1.90	4.24	2.54	13.20	1.59	5.13
O-TPT	Acc.	67.30	45.68	70.07	84.13	64.23	23.64	87.95	93.95	64.16	42.84	64.53	64.41
	ECE	1.97	7.88	3.87	1.46	4.93	3.68	1.90	3.80	2.34	12.98	1.78	4.24
FPP	Acc.	67.80	46.14	69.43	84.31	64.08	24.75	87.30	93.25	67.08	50.66	66.65	65.59
	ECE	2.97	7.52	2.67	1.92	3.22	7.26	2.63	5.85	3.04	5.19	2.00	4.02
Backpropagation-free													
BoostAdapter†	Acc.	69.42	45.39	71.21	86.30	67.71	27.51	89.15	94.20	70.84	49.85	69.03	67.33
	ECE	6.56	23.14	12.00	11.19	23.04	52.92	3.70	5.33	13.75	16.70	21.51	17.26
+ **CRAFT (Ours)**	Acc.	69.39	44.62	71.05	86.15	67.38	27.12	88.85	94.16	69.15	47.38	69.72	66.82
	ECE	**3.50**	**11.50**	**8.41**	**4.59**	**11.77**	**11.05**	**3.06**	**2.61**	**10.04**	**6.64**	**4.34**	**7.05**
TDA†	Acc.	69.44	44.98	71.30	86.11	67.71	25.32	89.45	94.04	71.11	63.11	67.52	68.19
	ECE	6.48	13.18	4.98	**1.49**	6.86	13.46	**2.86**	2.93	7.37	**6.71**	**1.90**	6.20
+ **CRAFT (Ours)**	Acc.	69.34	44.74	71.54	86.00	67.30	24.99	89.59	94.20	70.13	61.60	66.97	67.85
	ECE	**3.44**	**3.79**	**2.92**	1.96	**3.28**	**7.38**	3.88	**1.99**	**3.08**	11.26	2.61	**4.14**
TCA†	Acc.	64.60	44.27	70.24	79.19	61.57	22.23	85.12	90.30	63.26	57.88	55.42	63.10
	ECE	7.83	7.13	15.88	**0.80**	10.92	16.78	**3.76**	**3.24**	11.33	**3.46**	3.74	7.72
+ **CRAFT (Ours)**	Acc.	64.42	44.03	69.79	79.14	61.67	22.56	85.20	91.20	63.42	56.78	55.38	63.05
	ECE	**3.86**	**3.59**	**5.82**	2.46	**4.20**	**9.31**	4.88	4.03	**4.05**	14.06	**1.39**	**5.24**
SCA†	Acc.	72.09	55.67	75.31	86.34	69.47	28.74	90.81	94.32	71.79	57.17	66.98	69.88
	ECE	**6.89**	9.78	**6.31**	5.73	4.06	3.99	**9.84**	3.47	4.50	10.78	**7.34**	6.61
+ **CRAFT (Ours)**	Acc.	72.37	56.03	77.39	86.48	70.30	29.52	91.96	94.56	73.06	57.84	67.70	70.66
	ECE	6.90	**6.83**	6.81	**5.09**	**1.91**	**3.23**	12.74	**2.66**	**2.81**	**10.36**	9.44	**6.25**
PTA†	Acc.	70.10	47.99	74.54	86.28	68.94	25.68	91.28	94.85	73.46	60.99	68.57	69.33
	ECE	20.69	37.58	15.18	8.74	23.04	32.65	**2.92**	3.45	17.48	**12.42**	14.71	17.17
+ **CRAFT (Ours)**	Acc.	69.69	47.10	74.79	86.45	69.08	26.61	91.09	94.36	73.70	62.59	68.55	69.46
	ECE	**12.92**	**9.45**	**4.33**	**3.21**	**14.75**	**14.04**	3.11	**1.58**	**7.12**	13.07	**5.38**	**8.09**

Baselines and Architecture. We use CLIP ViT-B/16 (Radford et al., 2021) as the primary backbone. We include Zero-Shot CLIP and backpropagation-based TTA methods (TPT (Shu et al., 2022), C-TPT (Yoon et al., 2024), O-TPT (Sharifdeen et al., 2025), FPP (Jang et al., 2026)) as reference baselines. We then integrate CRAFT with five backpropagation-free methods: BoostAdapter (Zhang et al., 2024), TDA (Karmanov et al., 2024), TCA (Wang et al., 2025), SCA (Guan et al., 2025), and PTA (Huang et al., 2026).

5.2 Main Evaluation

Cross-Dataset Image Recognition. Table 1 reveals a clear accuracy–calibration trade-off in existing backpropagation-free TTA methods: adaptation can provide competitive or improved recognition accuracy while substantially increasing ECE. CRAFT consistently mitigates this degradation across all five backpropagation-free baselines, reducing their average ECE with only small changes in accuracy (-0.51 to $+0.78$ percentage points). The effect is particularly pronounced for strongly overconfident baselines; for example, CRAFT reduces ECE from 17.26% to 7.05% for BoostAdapter and from 17.17% to 8.09% for PTA, while preserving comparable accuracy. These results indicate that CRAFT largely decouples the adaptation benefit from the calibration degradation commonly introduced by prototype-based fusion.

Natural Distribution Shifts. As shown in Table 2, calibration degradation becomes more severe under natural distribution shifts, where several backpropagation-free baselines exhibit substantially larger ECE than Zero-Shot CLIP. CRAFT reduces the average ECE for all five baselines while

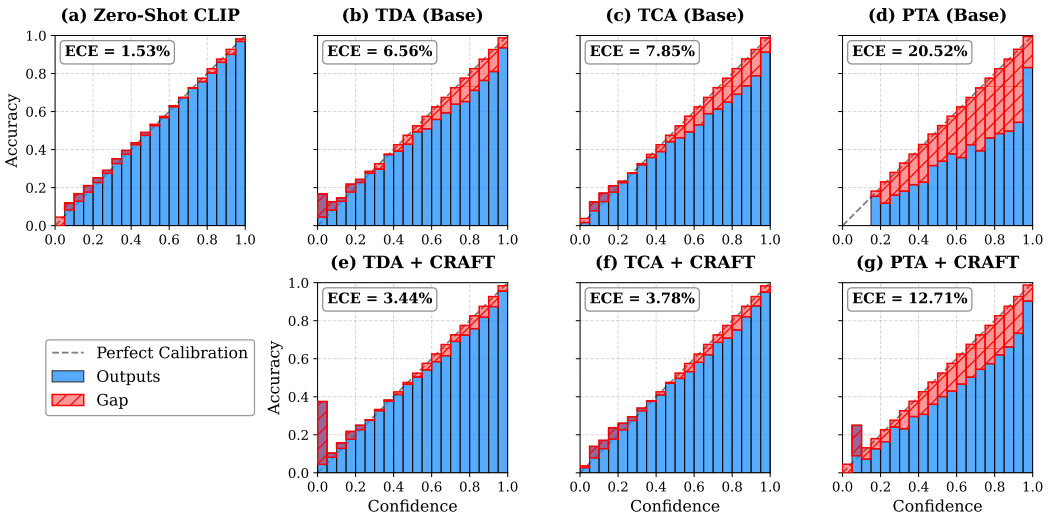


Figure 2: **Reliability diagrams** for Zero-Shot CLIP (Radford et al., 2021) and backpropagation-free TTA methods (Karmanov et al., 2024; Wang et al., 2025; Huang et al., 2026) with and without CRAFT, evaluated on the ImageNet dataset. Blue bars denote empirical accuracy within each confidence bin, red hatched regions the calibration gap, and the dashed diagonal perfect calibration.

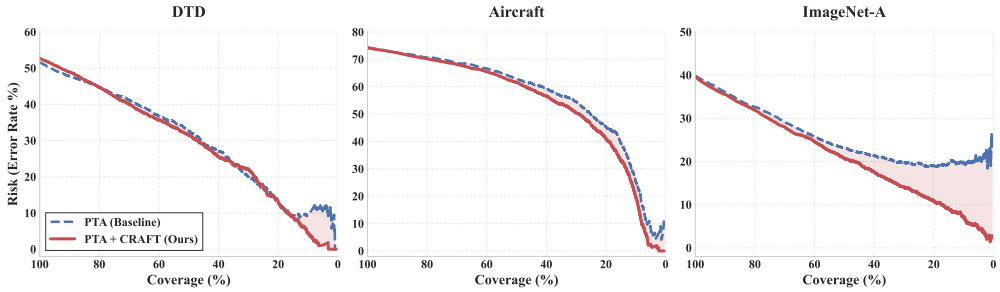


Figure 3: **Risk–coverage curves for selective prediction** on DTD, Aircraft, and ImageNet-A. Solid red denotes PTA+CRAFT and dashed blue denotes the PTA baseline; lower risk is better.

changing average accuracy by at most 0.34 percentage points. For instance, the ECE of Boost-Adapter improves from 23.88% to 12.72%, and that of PTA from 27.72% to 18.90%. This consistent reduction suggests that the calibration benefit of CRAFT is maintained even when the test distribution deviates substantially from the pre-training distribution.

Correlated Data Streams. We further examine online TTA under temporally correlated test streams in Table 3. As temporal correlation becomes stronger, the calibration of the original TTA baselines generally deteriorates; under $\gamma = 0.001$, PTA reaches an ECE of 20.01% and TDA 7.28%. CRAFT reduces these values to 9.24% and 3.99%, respectively, while inducing only minor changes in accuracy. These results show that CRAFT remains effective when the incoming stream departs from the conventional i.i.d. assumption.

Reliability Analysis. Figure 2 shows that the calibration degradation of backpropagation-free TTA is concentrated primarily in the high-confidence region, where predicted confidence substantially exceeds empirical accuracy. CRAFT narrows these gaps across the illustrated baselines, consistent with the ECE reductions in Tables 1 and 2. This suggests that CRAFT mitigates excessive confidence introduced during test-time adaptation rather than merely changing classification accuracy.

Selective Prediction. We further test whether CRAFT's confidence scores support selective prediction, where low-confidence predictions are abstained from. Across DTD, Aircraft, and ImageNet-A, CRAFT achieves lower risk than PTA as coverage decreases (Figure 3), most clearly on ImageNet-

Table 2: **Natural distribution shifts.** Accuracy and ECE on four ImageNet distribution shifts using CLIP ViT-B/16. † denotes results reproduced using the official implementation.

Method	Metric	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-Sketch	Avg
Zero Shot†	Acc.	59.37	64.21	80.56	49.48	63.41
	ECE	16.46	11.50	3.47	12.33	10.94
Backpropagation-based						
TPT	Acc.	52.60	63.00	76.70	47.50	59.95
	ECE	16.40	11.10	4.36	16.10	11.99
C-TPT	Acc.	51.60	62.70	76.00	47.90	59.55
	ECE	8.16	6.23	1.54	7.35	5.82
O-TPT	Acc.	49.90	61.70	72.60	47.10	57.82
	ECE	7.22	3.97	1.46	6.87	4.88
FPP	Acc.	52.30	61.90	76.70	47.20	59.53
	ECE	5.38	3.45	7.07	2.62	4.63
Backpropagation-free						
BoostAdapter†	Acc.	64.25	65.37	80.76	51.45	65.46
	ECE	31.49	21.25	12.39	30.39	23.88
+ **CRAFT (Ours)**	Acc.	63.77	65.21	80.56	50.95	65.12
	ECE	**17.01**	**13.67**	**4.27**	**15.94**	**12.72**
TDA†	Acc.	59.87	64.47	80.81	50.91	64.02
	ECE	19.32	14.56	3.76	16.08	13.43
+ **CRAFT (Ours)**	Acc.	59.78	64.37	80.73	50.38	63.82
	ECE	**17.27**	**12.82**	**3.52**	**13.61**	**11.81**
TCA†	Acc.	46.53	58.52	65.89	37.08	52.01
	ECE	17.86	21.54	**2.03**	17.64	14.77
+ **CRAFT (Ours)**	Acc.	46.43	58.48	66.11	37.52	52.14
	ECE	**12.69**	**13.52**	2.63	**8.83**	**9.42**
SCA†	Acc.	63.71	65.22	80.77	51.89	65.40
	ECE	15.35	**12.04**	2.36	**10.43**	10.05
+ **CRAFT (Ours)**	Acc.	64.16	65.46	80.87	52.26	65.69
	ECE	**13.18**	12.62	**1.47**	11.44	**9.68**
PTA†	Acc.	60.52	64.11	80.22	50.69	63.89
	ECE	32.10	30.16	14.32	34.29	27.72
+ **CRAFT (Ours)**	Acc.	59.79	64.15	80.35	50.57	63.72
	ECE	**23.40**	**23.62**	**7.51**	**21.05**	**18.90**

A. CRAFT thus improves confidence-based ranking, assigning higher confidence to correct predictions rather than merely reducing confidence globally; detailed AURC (Geifman et al., 2019) and Acc@C% results are in Appendix E.

5.3 Discussion

Ablation Study. Table 4 shows that HCIF+GPR and IGFR contribute complementary calibration improvements. For PTA, HCIF+GPR reduces ECE from 17.17% to 10.41%, and IGFR alone to 12.11%. Combining both further lowers ECE to 8.09%, slightly increasing accuracy from 69.33% to 69.46%. A similar pattern holds for TCA, where the full model reduces ECE from 7.72% to 5.24% with only a 0.05 percentage-point accuracy change. Notably, IGFR leaves accuracy unchanged at the reported precision, consistent with its rank-preserving design.

Computational Efficiency. Table 5 shows that CRAFT incurs modest throughput overhead (12.08–12.37%) while increasing peak memory by less than 0.5% for both TDA and PTA. Because CRAFT requires no additional model forward pass, it preserves the single-forward-pass property of the underlying backpropagation-free methods.

Table 3: **Online TTA under correlated streams.** Average Accuracy and ECE under different Dirichlet correlation levels and separate streams using CLIP ViT-B/16 with a batch size of 128.

Method	***Low*** ($\gamma = 0.1$)		***Medium*** ($\gamma = 0.01$)		***High*** ($\gamma = 0.001$)		***Separate***	
	Acc.	ECE	Acc.	ECE	Acc.	ECE	Acc.	ECE
Zero Shot	65.96	3.47	65.96	3.47	65.96	3.47	65.96	3.47
TDA	67.68	6.49	67.82	6.74	67.61	7.28	67.08	7.71
+ **CRAFT (Ours)**	67.62	**3.87**	67.17	**3.68**	67.21	**3.99**	67.05	**4.05**
TCA	62.02	9.09	62.82	9.07	62.11	10.15	62.87	9.38
+ **CRAFT (Ours)**	63.19	**5.32**	62.73	**5.11**	62.18	**5.17**	62.99	**5.91**
PTA	67.89	18.33	66.74	19.49	65.97	20.01	67.39	20.03
+ **CRAFT (Ours)**	67.96	**8.01**	66.05	**8.81**	64.78	**9.24**	64.37	**9.60**

Table 4: **Component ablation study.** Average Accuracy and ECE across the 11 cross-dataset image recognition benchmarks for TCA and PTA, with HCIF+GPR and IGFR enabled individually or jointly.

Model	HCIF + GPR	IGFR	Acc. (%)	ECE (↓)
TCA	✗	✗	**63.10**	7.72
	✓	✗	63.05	5.36
	✗	✓	**63.10**	6.84
	✓	✓	63.05	**5.24**
PTA	✗	✗	69.33	17.17
	✓	✗	**69.46**	10.41
	✗	✓	69.33	12.11
	✓	✓	**69.46**	**8.09**

Table 5: **Computational efficiency** on ImageNet-1K (Deng et al., 2009), measured on a single NVIDIA RTX 6000 Ada GPU.

Method	FPS (↑)	Mem. (MiB ↓)
CLIP	179.71	356.57
TDA	75.07	394.19
+ **CRAFT**	66.00	396.14
(Δ)	*-12.08%*	*+0.49%*
PTA	178.79	358.52
+ **CRAFT**	156.67	360.13
(Δ)	*-12.37%*	*+0.45%*

Supplementary Experiments. The appendix provides further experimental validation of CRAFT, including its generality across backbones and adaptation frameworks (Appendix D), integration with SaLS (Appendix G), evaluations on 3D point clouds (Appendix H), comparisons to L_2-normalized Euclidean fusion (Appendix I), and standard deviation analyses (Appendix J).

6 Conclusion

In this paper, we studied why backpropagation-free test-time adaptation degrades the calibration of vision-language models. We traced the degradation to a geometric cause: unnormalized Euclidean fusion of textual anchors and visual prototypes leaves the unit hypersphere, and the resulting class-dependent norm inflation acts as an uneven lowering of the softmax temperature, sharpening predictions beyond what their angular margins justify. Based on this analysis, we proposed CRAFT, which fuses logits along the geodesic on the unit hypersphere with a scale-invariant instance-level weighting, and limits the Fisher-Rao concentration radius to the model's own zero-shot level through a Fisher-Rao retraction, all without labeled data, tuned thresholds, or additional forward passes. Across 15 benchmarks covering fine-grained classification and natural distribution shifts, CRAFT consistently reduced the average calibration error on top of existing backpropagation-free methods while maintaining comparable accuracy and efficiency. Extending the geometric perspective to other fusion-based adaptation settings, such as multi-modal caches or continual test-time streams, is a natural next step.

Limitations. CRAFT anchors its confidence to the zero-shot prediction and its analysis to the vMF description of the latent space. Extending the framework beyond these choices, such as adaptive confidence references or relaxed distributional assumptions, and applying the geometric perspective to broader fusion-based adaptation settings are promising directions for future work.

REFERENCES

Soumojit Das, Nairanjana Dasgupta, and Prashanta Dutta. Geometric calibration and neutral zones for uncertainty-aware multi-class classification. *arXiv preprint arXiv:2511.20960*, 2025.

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

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2704–2714, 2023.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SJfb5jCqKm`.

Zenghao Guan, Yucan Zhou, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

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

Jameel Hassan, Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 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)*, 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)*, 2021b.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-based test-time adaptation of vision-language models. In *International Conference on Machine Learning (ICML)*, 2026.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171, 2024.

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

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Federico Pavesi, Antonio Candelieri, and Noémie Jaquier. Information theoretic bayesian optimization over the probability simplex. *arXiv preprint arXiv:2603.09793*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International conference on machine learning*, pp. 18559–18575. PMLR, 2022.

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

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1588–1597. IEEE, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

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

Zixin Wang, Dong Gong, Sen Wang, Zi Huang, and Yadan Luo. Is less more? exploring token condensation as training-free test-time adaptation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 144–154. IEEE, 2025.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961. IEEE, 2022.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1189. IEEE, 2023.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, volume 2024, pp. 8636–8660, 2024.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. *Advances in Neural Information Processing Systems*, 37:67795–67825, 2024.

A GEOMETRIC ANALYSIS OF LOGIT FUSION UNDER THE VMF MODEL

This appendix collects the analysis referenced in Sec. 4.1 and Sec. 4.2, organized in two steps. Sec. A.1 shows why the Euclidean fusion $V_c = t_c + p_c$ breaks calibration: the fused vector leaves the unit hypersphere with a class-dependent norm, which acts as a class-dependent scale distortion in the discriminative softmax. Sec. A.2 verifies that per-class SLERP keeps the fused vector at unit norm, effectively eliminating this artificial distortion.

We work on the unit hypersphere $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, on which both encoders place their L_2-normalized outputs, and describe class-conditional feature densities with the von Mises-Fisher (vMF) distribution,

$$P(x \mid y = c) = C_d(\kappa) \exp(\kappa(x \cdot t_c)), \qquad C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{14}$$

where $t_c \in \mathcal{S}^{d-1}$ is the mean direction (text anchor), $\kappa = 1/\tau$ is the concentration parameter, and I_ν is the modified Bessel function of the first kind. Throughout, $p_c \in \mathcal{S}^{d-1}$ denotes the visual prototype for class c, K the number of classes, and $\Omega_c = \arccos(t_c \cdot p_c)$ the anchor-prototype angle; since prototypes are aggregated from features predicted as class c, we assume $t_c \cdot p_c \geq 0$, i.e., $\Omega_c \in [0, \pi/2]$. The norm computations below are exact; the vMF model serves as the interpretive frame that links norms to calibration.

A.1 NORM INFLATION UNDER EUCLIDEAN FUSION

In this section, we show that Euclidean fusion inflates the fused norm unevenly across classes, and that this norm inflation acts as a class-dependent logit scale. Classes whose prototypes align well with their anchors—typically the classes the model already predicts frequently—receive systematically sharper predictions than their angular margins alone would justify. This uneven scaling is the primary mechanism behind the overconfidence observed in Sec. 4.1.

To see this, note that since $\|t_c\|_2 = \|p_c\|_2 = 1$,

$$\|V_c\|_2^2 = \|t_c\|_2^2 + \|p_c\|_2^2 + 2(t_c \cdot p_c) = 2 + 2\cos(\Omega_c), \tag{15}$$

which lies in $[2, 4]$ because $\cos(\Omega_c) \in [0, 1]$, so the fused norm $\|V_c\|_2 \in [\sqrt{2}, 2]$ depends on the class through the alignment angle Ω_c. Writing $\hat{V}_c = V_c/\|V_c\|_2 \in \mathcal{S}^{d-1}$ for the direction, the logit decomposes as

$$\kappa(x \cdot V_c) = (\kappa\|V_c\|_2)(x \cdot \hat{V}_c) = \kappa_c(x \cdot \hat{V}_c). \tag{16}$$

Therefore, unnormalized Euclidean fusion is equivalent to classifying with unit-norm directions $\hat{V}_c$, but under class-dependent scales $\kappa_c = \kappa\|V_c\|_2 \in [\sqrt{2}\kappa, 2\kappa]$.

In the discriminative classifier used during test-time adaptation, the predictive probability is computed by applying the softmax directly to these scaled logits:

$$P(y = c \mid x) = \frac{\exp\left(\kappa_c(x \cdot \hat{V}_c)\right)}{\sum_{j=1}^{K} \exp\left(\kappa_j(x \cdot \hat{V}_j)\right)}. \tag{17}$$

In zero-shot inference, the scale is uniform ($\kappa_c = \kappa$) for every class. After Euclidean fusion, however, the κ_c differ significantly. Because κ_c operates as an inverse temperature parameter, a larger κ_c effectively lowers the temperature for that specific class. Consequently, well-cached classes (where $t_c \cdot p_c$ is large) are evaluated with a systematically higher scale κ_c relative to classes with sparse or noisy caches. As the scale for the top class grows relative to the others, the predictive distribution sharpens independently of the actual angular evidence $(x \cdot \hat{V}_c)$. Under distribution shift, this geometric artifact sharpens probabilities without corresponding accuracy gains, thereby degrading calibration.

A.2 UNIT NORM OF PER-CLASS SLERP

In this section, we verify that replacing Euclidean addition with per-class SLERP removes the norm inflation identified above: the fused vector stays at unit norm for any interpolation weight, so all

classes share the same scale, eliminating the class-dependent temperature scaling effect. The per-class SLERP fusion is derived as:

$$V_c^*(\beta) = \frac{\sin((1-\beta)\Omega_c)}{\sin(\Omega_c)}\, t_c + \frac{\sin(\beta\Omega_c)}{\sin(\Omega_c)}\, p_c, \qquad \beta \in [0,1],\ \Omega_c \in (0,\pi). \tag{18}$$

Write $a = \sin((1-\beta)\Omega_c)$ and $b = \sin(\beta\Omega_c)$. Then

$$\|V_c^*(\beta)\|_2^2 = \frac{a^2 + b^2 + 2ab\cos(\Omega_c)}{\sin^2(\Omega_c)}. \tag{19}$$

Substituting $a = \sin(\Omega_c)\cos(\beta\Omega_c) - \cos(\Omega_c)\, b$ and expanding, the numerator reduces to

$$\sin^2(\Omega_c)\cos^2(\beta\Omega_c) + b^2\left(1 - \cos^2(\Omega_c)\right) = \sin^2(\Omega_c)\left(\cos^2(\beta\Omega_c) + \sin^2(\beta\Omega_c)\right) = \sin^2(\Omega_c), \tag{20}$$

so $\|V_c^*(\beta)\|_2^2 = 1$. This confirms that by utilizing class-specific interpolation angles Ω_c as described in Sec. 4.2, our HCIF mechanism preserves the unit-norm geometry. Consequently, the concentration parameter $\kappa_c = \|V_c^*\|_2/\tau$ becomes a constant $1/\tau$ across all classes, effectively eliminating the class-wise scale distortion caused by unnormalized Euclidean fusion.

B Properties of the Fisher-Rao Retraction

In this section, we verify the two claims made in Sec. 4.4: the retracted point lands exactly at the target radius r_{clip} while remaining on the unit sphere, and the retraction preserves the ordinal ranking of the class probabilities, so the Top-1 prediction is unchanged.

Recall the retraction $X'_{\text{fused}} = \alpha X_U + \gamma X_{\text{fused}}$, applied when $r_{\text{fused}} > r_{\text{clip}}$, with coefficients $\alpha = \sin(r_{\text{fused}} - r_{\text{clip}})/\sin(r_{\text{fused}})$ and $\gamma = \sin(r_{\text{clip}})/\sin(r_{\text{fused}})$.

For the first claim, since $\langle X_U, X_U\rangle = 1$ and $\langle X_U, X_{\text{fused}}\rangle = \cos(r_{\text{fused}})$ by the definition of the confidence radius,

$$\begin{aligned}
\langle X_U, X'_{\text{fused}}\rangle &= \alpha + \gamma\cos(r_{\text{fused}}) \\
&= \frac{\sin(r_{\text{fused}} - r_{\text{clip}}) + \sin(r_{\text{clip}})\cos(r_{\text{fused}})}{\sin(r_{\text{fused}})} \\
&= \frac{\sin(r_{\text{fused}})\cos(r_{\text{clip}})}{\sin(r_{\text{fused}})} = \cos(r_{\text{clip}}),
\end{aligned} \tag{21}$$

where the last step uses the identity $\sin(r_{\text{fused}} - r_{\text{clip}}) = \sin(r_{\text{fused}})\cos(r_{\text{clip}}) - \cos(r_{\text{fused}})\sin(r_{\text{clip}})$. An analogous expansion of $\|X'_{\text{fused}}\|_2^2$ using $\|X_U\|_2 = \|X_{\text{fused}}\|_2 = 1$ gives

$$\|X'_{\text{fused}}\|_2^2 = \alpha^2 + \gamma^2 + 2\alpha\gamma\cos(r_{\text{fused}}) = 1, \tag{22}$$

which follows by the same trigonometric reduction as in Sec. A.2, since the retraction is itself a spherical interpolation between X_U and X_{fused} at angle r_{fused}. The retracted point therefore stays on the unit sphere and sits exactly at Fisher-Rao radius r_{clip} from the uniform distribution, matching the zero-shot concentration radius.

For the second claim, note that $X_{U,c} = 1/\sqrt{C}$ is constant across classes, and $0 < r_{\text{clip}} < r_{\text{fused}} < \pi/2$ implies $\alpha > 0$ and $\gamma > 0$. Component-wise,

$$X'_{\text{fused},c} = \frac{\alpha}{\sqrt{C}} + \gamma\, X_{\text{fused},c}, \tag{23}$$

which is a strictly increasing affine map of $X_{\text{fused},c}$: whenever $X_{\text{fused},c} > X_{\text{fused},k}$, we also have $X'_{\text{fused},c} > X'_{\text{fused},k}$. Since all amplitudes are non-negative, squaring preserves this order, so the calibrated probabilities $P'_{\text{fused},c} = (X'_{\text{fused},c})^2$ carry the same ranking as $P_{\text{fused},c}$. The Top-1 prediction, and indeed the full ordering, is unchanged, while every gap between class probabilities is contracted toward the uniform distribution; accuracy is untouched and only confidence is reduced.

C Robustness under Non-IID Streaming Scenarios

Standard test-time adaptation often implicitly assumes that the incoming test stream is independent and identically distributed (i.i.d.). However, in real-world deployments, test samples frequently

exhibit temporal correlation (Zanella et al., 2025). To evaluate the robustness of CRAFT under distribution shifts with continuous temporal dependency, we evaluate our method on non-i.i.d. data streams controlled by a Dirichlet distribution parameterized by γ, as well as temporally separated streams (Sep.).

As shown in Table 6, baseline TTA methods such as TDA, TCA, and PTA experience an increase in Expected Calibration Error (ECE) as the correlation parameter γ decreases (indicating stronger

Table 6: **Online Test-Time Adaptation with Correlated Streams.** We report Accuracy (%) and Expected Calibration Error (ECE). Results are evaluated across different Dirichlet correlation parameters ($\gamma \in \{0.1, 0.01, 0.001\}$) and separate stream (Sep.) using CLIP-ViT-B/16 with a batch size of 128.

Method	γ	Metric	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	ImageNet	Avg
Zero Shot	-	Acc.	44.09	70.73	85.88	66.25	24.84	88.99	93.87	67.64	48.31	66.19	68.80	65.96
		ECE	3.64	4.13	2.09	1.66	3.82	4.67	3.58	3.44	5.88	3.74	1.53	3.47
TDA	0.1	Acc.	45.73	71.38	86.13	67.68	25.78	89.20	93.59	70.53	57.81	67.14	69.55	67.68
		ECE	13.21	4.96	**1.46**	7.02	13.04	**2.03**	3.25	8.43	9.98	**1.64**	6.42	6.49
	0.01	Acc.	44.95	71.71	86.08	67.45	26.71	89.17	94.04	68.94	60.16	67.40	69.42	67.82
		ECE	13.94	5.73	**1.21**	7.62	12.38	**1.63**	3.71	10.28	8.95	**1.82**	6.90	6.74
	0.001	Acc.	45.49	71.18	85.89	67.47	25.09	89.37	94.12	69.80	58.98	67.20	69.14	67.61
		ECE	14.01	6.60	**1.05**	8.45	14.35	**1.65**	3.59	10.78	8.90	2.69	8.05	7.28
	Sep.	Acc.	45.19	71.34	85.74	67.09	26.32	88.95	93.96	69.07	54.08	66.77	69.41	67.08
		ECE	14.89	6.50	**1.01**	9.14	13.17	**1.35**	4.00	10.94	12.84	**2.66**	8.36	7.71
+ CRAFT (Ours)	0.1	Acc.	45.55	71.75	86.03	67.26	25.03	89.96	94.08	69.15	58.73	66.85	69.42	67.62
		ECE	**3.69**	**2.57**	1.95	**3.18**	**7.66**	3.91	**1.84**	**3.50**	**8.42**	2.49	**3.36**	**3.87**
	0.01	Acc.	44.47	72.33	86.10	67.29	25.69	89.62	93.75	69.42	54.14	66.80	69.31	67.17
		ECE	**5.06**	**3.41**	1.90	**3.04**	**7.77**	3.36	**2.22**	**3.44**	**4.13**	2.59	**3.52**	**3.68**
	0.001	Acc.	45.13	71.09	86.09	67.23	25.45	89.59	94.08	68.27	56.37	66.82	69.22	67.21
		ECE	**3.89**	**3.27**	1.84	**3.75**	**7.40**	3.35	**1.91**	**4.88**	**6.95**	**2.37**	**4.24**	**3.99**
	Sep.	Acc.	44.47	71.13	85.91	67.03	25.69	88.64	94.20	68.62	56.01	66.57	69.23	67.05
		ECE	**4.35**	**3.38**	1.73	**4.12**	**7.10**	2.65	**2.63**	**4.94**	**6.32**	2.69	**4.67**	**4.05**
TCA	0.1	Acc.	43.21	70.19	79.25	61.48	22.78	85.55	91.86	62.63	45.56	55.27	64.41	62.02
		ECE	8.55	17.25	**0.79**	11.04	16.01	**2.59**	3.87	13.33	**14.80**	3.87	7.93	9.09
	0.01	Acc.	44.83	69.70	79.20	61.14	22.45	85.21	92.31	61.93	52.39	56.33	65.54	62.82
		ECE	8.38	18.61	**1.17**	12.29	16.66	**1.96**	**4.02**	15.32	**7.78**	5.64	7.91	9.07
	0.001	Acc.	44.53	67.31	78.93	60.69	22.87	83.76	92.48	60.86	51.93	55.70	64.10	62.11
		ECE	8.31	20.78	**1.54**	13.51	16.87	**2.00**	4.17	16.67	11.71	6.25	9.82	10.15
	Sep.	Acc.	43.39	69.94	79.07	60.49	22.45	84.51	92.89	60.43	57.99	56.20	64.22	62.87
		ECE	9.95	18.56	**1.48**	14.28	16.92	**1.93**	**3.37**	17.43	**3.13**	5.72	10.42	9.38
+ CRAFT (Ours)	0.1	Acc.	43.39	69.16	79.37	61.43	22.60	85.18	92.27	62.82	58.23	56.15	64.52	63.19
		ECE	**3.05**	**6.43**	2.50	**4.12**	**9.33**	4.17	**3.68**	**4.15**	15.77	**1.36**	**3.92**	**5.32**
	0.01	Acc.	44.11	69.28	79.30	61.34	22.84	85.52	92.93	63.15	52.02	55.22	64.31	62.73
		ECE	**2.92**	**7.95**	2.31	**4.74**	**9.01**	3.51	4.23	**5.72**	9.85	**1.69**	**4.30**	**5.11**
	0.001	Acc.	43.69	67.52	79.35	61.41	23.08	85.69	92.56	61.31	48.90	56.16	64.32	62.18
		ECE	**2.80**	**9.57**	2.24	**5.40**	**9.28**	3.92	**3.08**	**7.22**	**6.16**	**2.21**	**5.03**	**5.17**
	Sep.	Acc.	42.85	69.33	79.22	60.87	22.45	85.02	93.05	61.80	57.65	56.24	64.39	62.99
		ECE	**2.10**	**8.02**	2.10	**5.76**	**9.55**	3.67	3.52	**6.92**	16.14	**1.99**	**5.22**	**5.91**
PTA	0.1	Acc.	47.90	73.60	85.81	68.61	25.72	89.54	94.00	71.63	52.00	67.94	70.08	67.89
		ECE	37.23	15.63	9.07	23.34	32.90	4.49	4.18	19.13	19.76	15.23	20.71	18.33
	0.01	Acc.	46.21	72.82	84.36	67.54	24.61	89.40	93.67	70.20	49.47	66.39	69.52	66.74
		ECE	39.32	16.32	10.05	24.21	33.87	4.51	4.22	20.92	23.19	16.45	21.31	19.49
	0.001	Acc.	45.07	73.23	83.10	66.25	24.88	87.47	92.93	69.15	50.16	65.39	68.09	65.97
		ECE	40.75	16.14	10.92	25.32	33.49	5.96	4.98	21.69	20.49	17.62	22.71	20.01
	Sep.	Acc.	45.61	72.04	82.99	65.42	25.12	88.11	94.28	69.02	46.44	64.67	67.56	67.39
		ECE	39.65	16.58	11.03	26.06	33.84	5.23	4.25	21.85	20.65	17.94	23.21	20.03
+ CRAFT (Ours)	0.1	Acc.	48.02	73.81	85.74	68.86	25.99	90.12	94.33	71.88	50.88	68.04	69.87	67.96
		ECE	**8.64**	**5.21**	**3.63**	**15.04**	**14.57**	**2.55**	**1.64**	**9.37**	**8.97**	**5.66**	**12.80**	**8.01**
	0.01	Acc.	45.85	72.74	83.68	67.67	25.33	86.94	92.60	69.96	46.56	66.03	69.22	66.05
		ECE	**11.07**	**6.59**	**4.65**	**15.95**	**14.93**	**1.38**	**2.41**	**11.11**	**7.94**	**7.42**	**13.44**	**8.81**
	0.001	Acc.	45.73	70.97	81.59	66.03	25.63	86.94	91.57	68.70	43.17	64.76	67.49	64.78
		ECE	**11.03**	**6.72**	**5.50**	**17.07**	**14.15**	**1.12**	**3.11**	**12.51**	**6.51**	**8.81**	**15.10**	**9.24**
	Sep.	Acc.	43.15	71.09	81.55	64.95	24.46	86.50	94.04	69.50	41.62	64.39	66.82	64.37
		ECE	**12.91**	**7.10**	**5.59**	**18.23**	**16.06**	**1.93**	**1.78**	**11.79**	**5.81**	**8.77**	**15.67**	**9.60**

temporal correlation). This suggests that unnormalized logit fusion becomes increasingly vulnerable to overconfidence when the visual cache is populated with correlated, redundant features.

By integrating CRAFT, we observe a consistent reduction in ECE across all non-i.i.d. settings without a noticeable trade-off in classification accuracy. For instance, under the severe non-i.i.d. setting of PTA ($\gamma = 0.001$), CRAFT reduces the average ECE from 20.01% to 9.24%. These results indicate that preserving the hyperspherical geometry via HCIF and bounding confidence via IGFR provides a robust regularization mechanism against the rapid accumulation of biased prototypes in correlated data streams.

D Generality Across Backbones and Adaptation Frameworks

To verify the general applicability of our proposed method, we evaluate CRAFT across multiple vision-language backbones and existing backpropagation-free TTA frameworks. Tables 7 and 8 present the cross-dataset image recognition results on 11 diverse datasets using CLIP-ResNet50 and CLIP-ViT-L/14, respectively.

The empirical results demonstrate that CRAFT functions as an effective plug-and-play module. When applied to state-of-the-art backpropagation-free methods (e.g., BoostAdapter, TDA, SCA, and PTA), CRAFT consistently lowers the ECE across nearly all datasets while maintaining the accuracy gains achieved by the baseline methods. Notably, this behavior holds across different architectural scales, from ResNet50 to ViT-L/14.

This consistent calibration improvement supports our geometric analysis in Section 4: the calibration degradation observed in recent TTA methods is largely a structural artifact of unnormalized Euclidean fusion. By correcting this geometric misalignment, CRAFT provides a reliable confidence estimate irrespective of the underlying feature extractor of the baseline.

Table 7: **Cross-dataset image recognition.** We present the results of CLIP-ResNet50. † denotes results reproduced using the officially released code.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
Zero Shot†	Acc.	60.34	42.73	66.06	77.39	61.20	17.04	85.66	87.87	61.91	36.20	56.30	59.34
	ECE	2.02	5.04	4.78	3.21	3.02	3.31	7.00	3.96	5.57	5.01	4.59	4.32
Backpropagation-based													
TPT	Acc.	60.7	41.5	62.5	74.9	61.1	17.0	84.5	87.0	59.5	28.3	58.0	57.7
	ECE	11.4	25.7	13.4	5.25	9.24	16.1	3.65	5.04	12.4	22.5	3.76	11.7
C-TPT	Acc.	60.2	42.2	65.2	74.7	61.0	17.0	84.1	86.9	59.7	27.8	56.5	57.75
	ECE	3.01	19.8	4.14	1.86	2.93	10.7	2.77	2.07	3.83	15.1	1.94	6.19
O-TPT	Acc.	58.97	41.9	65.61	74.22	60.85	16.77	83.4	86.86	58.84	28.35	56.44	57.47
	ECE	3.1	16.53	2.5	1.2	3.2	8.18	3.5	2.75	2.6	14.71	1.69	5.45
Backpropagation-free													
BoostAdapter†	Acc.	60.85	43.14	64.35	77.41	62.51	18.54	86.45	87.99	65.00	30.26	60.15	59.70
	ECE	6.92	22.13	14.54	18.05	25.07	56.45	3.17	10.82	13.64	23.89	24.01	19.88
+ **CRAFT (Ours)**	Acc.	60.87	42.49	64.07	77.33	62.23	19.08	86.21	88.24	64.42	29.51	60.38	59.53
	ECE	**3.19**	**10.28**	**10.30**	**6.68**	**11.22**	**10.25**	**2.17**	**4.65**	**7.54**	**9.32**	**3.04**	**7.15**
TDA†	Acc.	61.04	43.79	68.62	77.83	62.40	17.46	85.96	89.53	64.23	40.73	57.34	60.81
	ECE	6.65	10.83	**3.72**	**2.30**	6.77	12.39	**3.96**	4.12	7.08	17.29	**1.60**	6.97
+ **CRAFT (Ours)**	Acc.	60.93	44.15	67.48	77.63	62.20	17.37	86.13	89.21	63.71	40.68	56.95	60.59
	ECE	**3.19**	**4.88**	3.91	2.98	**3.17**	**6.19**	5.85	**2.55**	**2.75**	**4.72**	3.48	**3.97**
SCA†	Acc.	63.74	52.42	68.86	77.73	64.52	20.40	85.94	89.70	65.72	42.04	58.57	62.69
	ECE	**8.32**	9.69	6.35	8.43	4.34	2.37	11.96	2.86	6.20	7.92	9.40	7.08
+ **CRAFT (Ours)**	Acc.	63.98	53.19	70.44	77.70	64.75	20.37	86.51	90.43	66.59	41.31	60.07	63.21
	ECE	8.91	**6.98**	**5.47**	**6.54**	**2.03**	**2.20**	**11.86**	**2.38**	**3.56**	**5.68**	**9.02**	**5.88**
PTA†	Acc.	60.56	44.44	68.49	76.96	63.04	16.98	86.92	88.84	66.01	49.35	59.62	61.93
	ECE	27.75	40.93	19.85	14.57	28.18	39.83	4.48	7.85	23.83	14.71	20.17	22.01
+ **CRAFT (Ours)**	Acc.	60.14	44.92	67.84	76.67	63.03	16.35	86.32	88.92	65.69	49.73	59.37	61.73
	ECE	**18.18**	**10.02**	**7.81**	**6.27**	**18.42**	**18.06**	**3.08**	**3.15**	**10.19**	**13.84**	**8.22**	**10.66**

Table 8: **Cross-dataset image recognition.** We present the results of ViT-L/14.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
Zero Shot	Acc.	75.93	52.90	79.54	90.91	70.51	32.67	93.51	95.50	74.94	60.28	77.90	73.14
	ECE	3.72	3.98	4.26	0.85	3.85	6.11	2.75	2.41	4.14	3.72	3.85	3.60
BoostAdapter	Acc.	76.27	58.22	79.33	90.92	71.37	36.87	93.46	96.02	76.92	59.99	78.95	74.39
	ECE	7.31	17.92	10.66	7.68	21.70	48.93	2.88	3.59	14.16	18.16	15.53	15.32
+ **CRAFT (Ours)**	Acc.	76.23	56.44	79.01	90.81	71.07	36.96	93.43	96.15	76.45	57.88	79.12	73.96
	ECE	**5.51**	**11.74**	**8.47**	**4.18**	**14.12**	**15.86**	**2.10**	**1.91**	**12.17**	**11.22**	**4.68**	**8.36**
TDA	Acc.	76.36	56.26	80.19	90.96	71.79	34.26	93.95	96.15	77.00	66.22	78.36	74.68
	ECE	7.26	9.04	4.73	**0.50**	8.35	15.57	**1.33**	1.95	8.09	10.11	**1.12**	6.19
+ **CRAFT (Ours)**	Acc.	76.24	55.14	80.06	90.94	71.39	33.54	93.68	96.19	75.92	65.06	78.26	74.22
	ECE	**5.43**	**4.31**	**3.11**	0.72	**5.96**	**10.58**	1.91	**1.69**	**6.08**	**5.59**	2.44	**4.35**
TCA	Acc.	71.06	54.61	78.20	82.58	65.72	29.64	90.79	95.25	74.76	59.46	65.30	69.76
	ECE	9.04	9.26	11.58	1.18	10.95	18.30	**1.64**	2.02	7.98	10.15	4.48	7.87
+ **CRAFT (Ours)**	Acc.	70.77	54.08	77.59	82.55	65.26	29.22	90.65	94.65	73.41	58.63	64.86	69.24
	ECE	**6.50**	**5.16**	**5.74**	**0.56**	**5.90**	**12.50**	2.07	**1.63**	**5.42**	**4.48**	**2.35**	**4.76**
PTA	Acc.	77.69	58.16	82.95	91.46	73.96	35.13	94.33	96.51	80.04	70.36	80.35	76.45
	ECE	16.65	29.53	12.00	5.68	19.97	29.95	2.66	2.53	15.12	14.43	9.19	14.34
+ **CRAFT (Ours)**	Acc.	77.43	57.68	82.62	91.53	73.67	35.61	94.25	97.16	80.28	70.05	80.60	76.44
	ECE	**12.39**	**10.96**	**6.93**	**2.61**	**14.95**	**15.50**	**0.73**	**1.15**	**9.69**	**5.84**	**3.13**	**7.63**

E Reliability of Confidence via Selective Prediction

A reliable model should exhibit high accuracy on the retained subset (Acc@$C\%$) alongside a low Area Under the Risk-Coverage curve (AURC).

Table 9 summarizes the ECE, AURC, and Acc@$C\%$ at various coverage levels. The results demonstrate that CRAFT consistently achieves lower risk at equivalent coverage levels compared to the PTA baseline. For instance, on ImageNet-A (Hendrycks et al., 2021b), which is known to induce severe overconfidence, CRAFT improves Acc@10% by $+12.54\%$ and substantially reduces the AURC (-47.14×10^{-3}). These quantitative improvements confirm that CRAFT provides more trustworthy confidence estimates, making it highly suitable for risk-sensitive applications where accurate sample ranking is required.

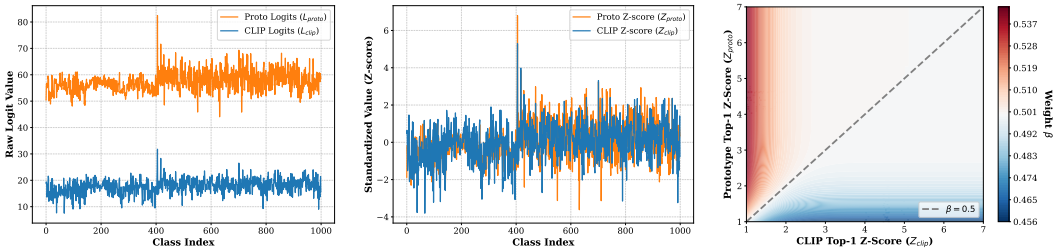
Table 9: Quantitative results for selective prediction. ECE and AURC (scaled by 10^3) evaluate the reliability of confidence (lower is better). Acc@$C\%$ represents the accuracy on the retained subset when only the top $C\%$ most confident predictions are considered (higher is better). The Gap row indicates the difference between PTA+CRAFT and PTA.

Dataset	Method	Calibration & Risk		Accuracy at Coverage (%)					
		ECE (↓)	AURC ($\times 10^3$ ↓)	100% (↑)	90% (↑)	80% (↑)	70% (↑)	50% (↑)	10% (↑)
DTD	PTA	37.39	302.58	**48.35**	**52.30**	**55.28**	58.78	67.26	90.53
	PTA + **CRAFT**	**9.05**	**291.34**	47.22	51.12	55.21	**59.88**	**68.56**	**95.27**
	(Δ)	*-28.34*	*-11.24*	*-1.13*	*-1.18*	*-0.07*	*+1.10*	*+1.30*	*+4.74*
Aircraft	PTA	32.63	565.76	25.65	**27.61**	29.33	31.29	37.03	74.77
	PTA + **CRAFT**	**14.74**	**540.60**	**25.83**	**27.61**	**29.78**	**31.89**	**38.48**	**80.18**
	(Δ)	*-17.89*	*-25.16*	*+0.18*	*0.00*	*+0.45*	*+0.60*	*+1.45*	*+5.41*
ImageNet-A	PTA	32.44	257.80	60.14	63.95	67.41	70.83	76.92	80.32
	PTA + **CRAFT**	**22.63**	**210.66**	**60.39**	**64.50**	**68.14**	**72.02**	**79.34**	**92.86**
	(Δ)	*-9.81*	*-47.14*	*+0.25*	*+0.55*	*+0.73*	*+1.19*	*+2.42*	*+12.54*

F The Rationale Behind Gaussian Probabilistic Routing (GPR)

In this section, we provide a visual intuition for the design of Gaussian Probabilistic Routing (GPR) introduced in Section 4.3.

The primary motivation for GPR is the scale mismatch between different models. As illustrated in Figure 4a, the raw logit outputs from the CLIP (L_{clip}) and Prototype (L_{proto}) models exhibit



(a) Absolute Scale Mismatch (b) Scale-Invariant Z-Score (c) GPR Routing Landscape

Figure 4: **Conceptual visualization of Gaussian Probabilistic Routing (GPR).** (a) Raw logits from CLIP and Proto models show mismatched absolute scales, which complicates the use of static thresholds. (b) GPR standardizes the logits into Z-scores, aligning their scales to fairly compare the relative sharpness of the top-1 predictions. (c) The routing landscape maps the top-1 Z-scores to the interpolation weight β, smoothly adjusting the fusion and naturally approaching 0.5 when both sources are highly confident.

Table 10: Ablation study of GPR variants (w/o Z-score, Entropy-Based, GPR) applied to the TDA and PTA across 11 benchmark datasets. We report Accuracy (Acc.), Expected Calibration Error (ECE). The lowest ECE values for each metric are highlighted in **bold**.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
TDA	Acc.	69.44	44.98	71.30	86.11	67.71	25.32	89.45	94.04	71.11	63.11	67.52	68.19
	ECE	6.48	13.18	4.98	**1.49**	6.86	13.46	**2.86**	2.93	7.37	**6.71**	**1.90**	6.20
+ Entropy-Based	Acc.	69.33	44.92	71.42	86.02	67.32	25.08	89.48	94.28	70.31	62.68	66.98	67.98
	ECE	3.53	4.11	2.85	1.97	3.48	7.87	3.72	**2.01**	3.32	10.91	2.65	4.22
+ w/o Z-score	Acc.	69.35	44.86	71.42	86.02	67.32	25.08	89.51	94.28	70.31	62.31	66.97	67.95
	ECE	3.83	4.26	**2.76**	1.95	3.63	8.62	3.78	2.13	3.53	10.69	2.21	4.31
+ GPR	Acc.	69.33	44.74	71.50	86.03	67.34	25.05	89.48	94.28	70.18	61.30	66.99	67.84
	ECE	**3.43**	**3.86**	2.83	1.99	**3.21**	**7.37**	3.78	2.09	**3.12**	10.80	2.67	**4.10**
PTA	Acc.	70.25	47.87	74.54	86.29	69.30	25.77	91.22	94.89	73.41	61.12	68.56	69.38
	ECE	20.57	37.63	15.19	8.73	22.70	32.46	3.09	3.44	17.56	12.34	14.72	17.13
+ Entropy-Based	Acc.	69.61	46.93	75.03	86.39	69.05	26.13	91.25	94.28	73.91	62.15	68.62	69.40
	ECE	13.44	10.79	4.81	3.51	15.41	15.99	**2.88**	1.84	7.69	**11.74**	5.82	8.54
+ w/o Z-score	Acc.	69.68	47.16	75.07	86.38	69.03	26.16	91.20	94.32	73.67	62.37	68.62	69.42
	ECE	14.07	10.79	5.11	3.83	16.10	16.71	2.98	**1.83**	8.29	11.90	6.46	8.92
+ GPR	Acc.	69.70	47.04	74.87	86.42	69.13	26.19	91.17	94.32	73.86	62.64	68.49	69.44
	ECE	**12.91**	**9.48**	**4.42**	**3.24**	**14.71**	**14.41**	3.04	1.85	**6.96**	13.15	**5.46**	**8.15**

distinct absolute scales. This discrepancy makes it challenging to rely on conventional entropy- or confidence-based gating mechanisms, as a fixed threshold tuned on one domain may not generalize well to another.

To address this scale discrepancy, GPR normalizes the logits into scale-invariant Z-scores, as shown in Figure 4b. By standardizing the distributions, we can fairly evaluate the relative sharpness of the top-1 prediction (highlighted with star markers) regardless of the original logit magnitude.

Finally, Figure 4c depicts the resulting routing landscape based on the standardized Z-scores. The interpolation weight β smoothly transitions to favor the more confident model. Notably, when both models yield highly confident predictions (large Z-scores), β naturally converges toward a balanced fusion ($\beta \approx 0.5$). This behavior allows our method to dynamically adapt to varying confidence levels across diverse datasets without requiring manual hyperparameter tuning.

Furthermore, we present an ablation study in Table 10 to evaluate the effectiveness of GPR against alternative routing strategies. The evaluation is performed on the TDA and PTA baselines. Note that the configurations indicated by '+' (e.g., + Entropy-Based, + w/o Z-score, + GPR) are eval-

uated under the assumption that both Hyperspherical Concentration-Invariant Fusion (HCIF) and Information-Geometric Fisher-Rao Retraction (IGFR) are applied by default.

We compare three routing mechanisms for determining the interpolation weight β:

- **Entropy-Based**: β is computed based on the entropy of the raw logits, without Z-score standardization.
- **w/o Z-score**: β is derived directly from the maximum raw logit values, without Z-score standardization.
- **GPR (Ours)**: Our proposed method standardizes the logits into scale-invariant Z-scores to compare the relative sharpness of the top-1 predictions.

The results in Table 10 demonstrate that GPR consistently achieves the lowest Expected Calibration Error (ECE) across both baselines while maintaining comparable accuracy. For instance, on the PTA baseline, GPR achieves an average ECE of 8.15%, outperforming both the Entropy-Based (8.54%) and w/o Z-score (8.92%) variants. These findings indicate that standardizing the scale mismatch between different models through Z-scores is a crucial component for reliable routing, directly supporting the conceptual intuition discussed above.

G Additional Experiments with the SaLS Method

Table 11: Performance comparison of baseline methods (TDA, TCA, PTA) and the proposed variants across 11 benchmark datasets. We report Accuracy (Acc.), Expected Calibration Error (ECE), and Area Under the Risk-Coverage curve (AURC) to analyze the effect of adding the proposed methods. The lowest ECE and AURC (scaled by 10^3) values for each baseline group are highlighted in **bold**.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
TDA	Acc.	69.42	45.04	71.38	86.11	67.76	25.35	89.56	94.00	71.08	62.95	67.43	68.19
	ECE	6.51	13.26	5.17	**1.45**	6.65	13.39	**3.11**	2.93	7.24	**6.97**	**1.68**	6.21
	AURC	122.82	307.55	98.22	**27.27**	141.76	549.76	15.39	10.78	105.01	**195.89**	124.63	154.46
+ SaLS	Acc.	69.42	45.04	71.38	86.11	67.76	25.35	89.56	94.00	71.08	62.95	67.43	68.19
	ECE	5.05	7.00	4.04	1.73	4.52	10.89	3.38	**1.86**	5.21	8.02	1.93	4.88
	AURC	122.11	**302.62**	97.79	27.35	**141.48**	**547.33**	15.34	10.05	101.13	201.65	**124.42**	153.75
+ CRAFT	Acc.	69.35	44.74	71.58	86.06	67.28	25.29	89.45	94.28	70.10	61.26	66.96	67.85
	ECE	**3.44**	**3.93**	**2.81**	2.08	**3.32**	**7.44**	3.83	2.22	**2.78**	10.99	2.51	**4.12**
	AURC	**120.27**	305.07	**94.04**	27.49	142.57	547.57	**15.07**	**8.93**	**99.45**	202.99	125.12	**153.51**
TCA	Acc.	64.50	43.91	70.40	79.19	61.43	22.29	84.85	90.59	63.15	58.48	56.85	63.24
	ECE	7.86	7.35	15.47	**0.78**	11.00	16.66	**3.62**	**3.11**	11.46	**2.98**	3.76	7.64
	AURC	158.25	316.79	138.08	54.18	202.90	597.10	28.98	23.77	178.84	**227.83**	216.31	194.82
+ SaLS	Acc.	64.50	43.91	70.40	79.19	61.43	22.29	84.85	90.59	63.15	58.48	56.85	63.24
	ECE	5.44	**3.22**	10.56	2.06	6.18	12.62	4.20	3.58	7.22	14.93	3.13	6.65
	AURC	156.21	314.24	119.74	54.26	197.36	592.66	28.77	21.11	156.12	269.98	214.92	193.22
+ CRAFT	Acc.	64.42	43.97	69.87	79.25	61.71	22.47	85.17	91.32	64.08	57.43	56.46	63.29
	ECE	**3.77**	3.44	**5.57**	2.52	**4.32**	**9.31**	4.62	3.88	**3.54**	14.60	**2.05**	**5.24**
	AURC	**152.39**	**313.19**	**106.43**	**54.09**	**192.50**	**586.31**	**27.77**	**18.82**	**140.39**	276.72	**210.39**	**189.00**
PTA	Acc.	70.13	48.05	74.58	86.30	68.93	25.89	91.28	94.85	73.38	60.91	68.55	69.35
	ECE	20.66	37.44	15.19	8.72	23.09	32.38	**2.93**	3.52	17.57	**12.53**	14.74	17.16
	AURC	133.89	305.12	96.75	30.52	146.19	561.25	13.38	6.42	91.15	193.32	123.63	154.69
+ SaLS	Acc.	69.99	46.99	74.99	86.39	69.26	26.04	91.06	94.60	73.12	61.80	68.64	69.35
	ECE	**3.87**	10.46	31.30	**1.55**	**3.39**	**6.40**	4.96	2.12	**3.16**	17.19	**1.54**	**7.81**
	AURC	**120.59**	326.00	102.84	26.94	**128.71**	**541.82**	13.94	**5.03**	83.12	208.24	**118.06**	152.30
+ CRAFT	Acc.	69.72	47.28	74.75	86.43	69.13	26.10	91.11	94.36	73.80	62.53	68.50	69.43
	ECE	12.89	**9.28**	**4.61**	3.22	14.71	14.52	3.11	**1.66**	7.00	13.05	5.44	8.14
	AURC	123.13	**287.92**	**80.70**	**26.45**	134.21	543.59	**12.68**	7.30	**82.81**	**191.90**	120.87	**146.51**

In this section, we provide additional experimental results to demonstrate the effect of incorporating the SaLS method (Murugesan et al., 2024) into our proposed framework. SaLS is a post-hoc

Table 12: Performance comparison of baseline methods (TDA, PTA) and the proposed variants across 11 benchmark datasets. We report Accuracy (Acc.) and Expected Calibration Error (ECE) to analyze the effect of adding the SaLS (Murugesan et al., 2024) method. The lowest ECE values for each baseline group are highlighted in **bold**.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
TDA	Acc.	69.43	44.92	71.17	86.09	67.79	25.23	89.48	93.96	71.16	63.51	67.57	68.21
	ECE	6.56	13.39	5.36	**1.46**	6.68	13.42	**2.98**	2.88	7.09	**7.00**	**1.81**	6.24
+ HCIF + GPR + **IGFR**	Acc.	69.34	44.74	71.54	86.00	67.30	24.99	89.59	94.20	70.13	61.60	66.97	67.85
	ECE	3.44	**3.79**	2.92	1.96	3.28	**7.38**	3.88	1.99	3.08	11.26	2.61	4.14
+ HCIF + GPR + **SaLS**	Acc.	69.34	44.74	71.54	85.99	67.29	25.02	89.59	94.20	70.16	61.59	67.02	67.86
	ECE	**2.99**	4.93	2.91	1.95	2.80	7.47	4.16	**1.79**	2.53	9.28	2.87	3.97
+ HCIF + GPR + **SaLS + IGFR**	Acc.	69.34	44.74	71.54	85.99	67.29	25.02	89.59	94.20	70.16	61.59	67.02	67.86
	ECE	3.00	4.95	**2.90**	1.94	**2.79**	7.47	4.14	**1.79**	**2.49**	9.13	2.87	**3.95**
PTA	Acc.	70.10	47.99	74.54	86.28	68.94	25.68	91.28	94.85	73.46	60.99	68.57	69.33
	ECE	20.69	37.58	15.18	8.74	23.04	32.65	**2.92**	3.45	17.48	12.42	14.71	17.17
+ HCIF + GPR + **IGFR**	Acc.	69.69	47.10	74.79	86.45	69.08	26.61	91.09	94.36	73.70	62.59	68.55	69.45
	ECE	12.92	9.45	**4.33**	3.21	14.75	14.04	3.11	1.58	7.12	13.07	5.38	8.09
+ HCIF + GPR + **SaLS**	Acc.	69.99	47.40	75.03	86.37	69.17	26.10	90.95	94.65	73.09	61.51	68.61	69.35
	ECE	**5.37**	11.95	26.60	0.92	**6.02**	4.62	4.52	1.94	**2.64**	7.66	**1.55**	6.71
+ HCIF + GPR + **SaLS + IGFR**	Acc.	69.99	47.40	75.03	86.37	69.17	26.10	90.95	94.65	73.09	61.51	68.61	69.35
	ECE	5.41	**6.74**	17.66	**0.84**	6.09	**3.53**	4.23	**1.36**	2.90	**6.97**	1.57	**5.21**

calibration technique that rescales the adapted logits so that their range (the difference between the maximum and minimum values) matches that of their paired zero-shot counterparts. While this approach constrains the logit boundaries, it relies solely on the extreme values and does not account for the probability mass distribution across the remaining classes. Because selective prediction and reliability depend heavily on this full distribution, ignoring the remaining probability mass limits the model's ability to maintain a calibrated ordinal ranking of confidence scores. To highlight this limitation, we evaluate the models using the Area Under the Risk-Coverage curve (AURC), which explicitly measures confidence-based ranking performance and risk across all predictions. Here, we explore the efficacy of SaLS both as an independent calibration baseline and as a complementary component within our pipeline.

Table 11 presents a detailed comparison of Accuracy, Expected Calibration Error (ECE), and AURC across 11 benchmark datasets for the TDA (Karmanov et al., 2024), TCA (Wang et al., 2025), and PTA (Huang et al., 2026) baselines, comparing them directly with SaLS and our proposed CRAFT method. The results indicate that applying SaLS to the baselines reduces ECE while maintaining the original classification accuracy. However, when comparing the independent application of these methods, our proposed CRAFT framework demonstrates a clear advantage in preserving the reliability of the predictive distribution. For the TDA and TCA baselines, CRAFT achieves a more substantial average ECE reduction than SaLS. Furthermore, across all three baselines, CRAFT yields lower average AURC values compared to SaLS. For instance, applying CRAFT to the TCA baseline lowers the average ECE to 5.24 and AURC to 189.00, compared to an ECE of 6.65 and AURC of 193.22 with SaLS. For the overconfident PTA, while SaLS provides a competitive average ECE, CRAFT achieves a substantially superior AURC (146.51 compared to 152.30 for SaLS). This performance gap confirms that preserving the hyperspherical geometry via HCIF and bounding confidence via IGFR yields more reliable risk-coverage profiles than simply rescaling extreme logit values.

Table 12 further evaluates the integration of SaLS into our complete pipeline (HCIF + GPR + IGFR). The empirical results demonstrate that SaLS acts as an effective complementary component, generally improving overall model calibration when combined with our geometric constraints. For the TDA baseline, introducing SaLS alongside HCIF, GPR, and IGFR reduces the average ECE to 3.95, while maintaining competitive classification accuracy compared to the baseline.

This improvement is particularly notable in the PTA configurations. The base PTA exhibits a relatively high average ECE of 17.17. While the combination of HCIF, GPR, and IGFR reduces this to 8.09, integrating SaLS into this configuration (HCIF + GPR + SaLS + IGFR) further lowers the average ECE to 5.21.

Although performance varies slightly across individual datasets—for instance, observing minor ECE fluctuations in datasets such as Flower and Pets under the PTA configuration—the overall trend suggests that the SaLS method serves as a beneficial supplementary module. On average, it facilitates better confidence calibration across diverse datasets without causing notable degradation in overall prediction accuracy.

H Additional Experiments with Point Cloud

Table 13: Accuracy (Acc, ↑) and Expected Calibration Error (ECE, ↓) on ModelNet-C (Ren et al., 2022) and ScanObjectNN (Uy et al., 2019) variants (Severity 2). The table compares Zero-Shot ULIP (Xue et al., 2023), PTA (Huang et al., 2026), and PTA+CRAFT. Between PTA and + CRAFT, the better (lower) ECE values are highlighted in **bold**.

Method	Add Global		Add Local		Drop Global		Drop Local		Jitter		Rotate		Scale		Average	
	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
ModelNet-C																
ULIP	33.51	40.70	44.08	31.82	54.70	27.99	50.89	26.40	43.80	28.66	55.39	25.44	51.01	29.12	47.63	30.02
PTA	49.47	43.21	51.50	43.98	61.83	35.58	57.37	38.77	49.88	45.93	61.63	35.79	57.21	39.87	55.56	40.45
+ CRAFT	49.84	**36.25**	51.13	**39.42**	62.28	**31.79**	57.66	**33.38**	49.76	**40.72**	61.63	**31.84**	57.13	**36.06**	55.63	**35.64**
ScanObjectNN (OBJ-ONLY)																
ULIP	32.01	38.28	34.77	36.46	51.12	27.94	38.90	35.74	36.83	32.03	48.71	28.40	44.75	32.05	41.01	32.99
PTA	34.94	56.36	39.76	54.05	53.53	43.52	45.78	50.41	39.59	52.36	51.81	44.47	46.64	50.31	44.58	50.21
+ CRAFT	35.28	**44.88**	39.59	**45.36**	54.04	**37.57**	45.96	**42.35**	40.28	**42.07**	51.64	**38.71**	46.47	**44.11**	44.75	**42.15**
ScanObjectNN (OBJ-BG)																
ULIP	26.51	41.50	27.71	41.64	45.78	32.33	35.11	37.73	31.33	37.09	41.14	34.41	39.59	36.46	35.31	37.31
PTA	32.53	57.54	31.15	61.60	52.15	44.90	42.00	52.00	32.36	59.64	48.36	46.72	46.47	49.45	40.72	53.12
+ CRAFT	32.19	**47.09**	31.67	**51.69**	52.15	**38.27**	41.31	**44.82**	30.98	**50.63**	47.33	**41.05**	45.61	**43.34**	40.18	**45.27**
ScanObjectNN (hardest)																
ULIP	18.98	45.82	18.81	50.50	30.88	43.06	23.77	48.44	21.83	44.81	26.82	45.34	26.30	46.77	23.91	46.39
PTA	24.12	61.98	22.00	68.30	38.38	55.07	32.51	59.40	23.59	66.20	34.25	58.18	30.85	62.20	29.39	61.62
+ CRAFT	23.42	**50.75**	22.03	**58.18**	37.58	**47.75**	32.03	**50.97**	23.63	**55.08**	33.90	**49.77**	30.15	**54.49**	28.96	**52.43**

To evaluate the generalizability of the proposed CRAFT beyond 2D vision-language models, we extend our empirical analysis to the 3D point cloud modality. In this section, we apply CRAFT to PTA (Huang et al., 2026), a recent prototype-based test-time adaptation method built upon the 3D zero-shot model ULIP (Xue et al., 2023).

We evaluate the models on the corrupted 3D benchmarks ModelNet-C (Ren et al., 2022) and three variants of ScanObjectNN (Uy et al., 2019): OBJ-ONLY, OBJ-BG, and the hardest split. Following standard robustness evaluation protocols, we report both Accuracy (Acc) and Expected Calibration Error (ECE) under seven types of severity-2 corruptions (Add Global, Add Local, Drop Global, Drop Local, Jitter, Rotate, Scale). The results are summarized in Table 13.

Consistent with our findings in 2D image classification, the PTA yields considerable improvements in accuracy over the zero-shot ULIP model but incurs a severe penalty in calibration. For instance, on ModelNet-C, PTA improves the average accuracy from 47.63% to 55.56%, yet the average ECE increases from 30.02% to 40.45%. A similar trend is pervasive across all variants of ScanObjectNN. This empirical evidence supports the analysis in Section 4.1: the asymmetric norm inflation caused by Euclidean fusion is not an artifact of 2D vision-language spaces, but a general structural vulnerability of hyperspherical prototype accumulation, leading to uncalibrated, overconfident predictions across modalities.

By integrating our proposed CRAFT framework with PTA, we observe a consistent and notable reduction in ECE across nearly all corruption types and datasets. As highlighted in Table 13, CRAFT reduces the average ECE on the ScanObjectNN (hardest) variant from 61.62% to 52.43%, and from 50.21% to 42.15% on OBJ-ONLY. This improvement validates the efficacy of our Hyperspherical Concentration-Invariant Fusion (HCIF, Section 4.2) in preventing class-wise concentration biases, as well as the utility of the Information-Geometric Fisher-Rao Retraction (IGFR, Section 4.4) in appropriately bounding the fused confidence using the zero-shot prior.

In summary, these supplementary experiments demonstrate that CRAFT is a versatile, modality-agnostic module capable of mitigating the calibration degradation inherent in prototype-based TTA, confirming its utility in 3D point cloud architectures.

I Comparison with L_2-Normalized Euclidean Fusion

We identified asymmetric norm inflation during unnormalized Euclidean fusion as a primary driver of miscalibration. A natural baseline to address this is applying L_2 normalization to the additively fused evidence vector. However, empirical results in Table 14 demonstrate that this straightforward geometric projection is insufficient to fully restore model reliability. For instance, when applied to BoostAdapter (Zhang et al., 2024), simple L_2 normalization yields an average Expected Calibration Error (ECE) of 7.87%. CRAFT improves upon this, lowering the average ECE to 6.98% while maintaining comparable average accuracy (66.80% vs. 66.79%). This trend is consistent across other base methods. For TCA (Wang et al., 2025), simple L_2 normalization results in an average ECE of 5.38% and an accuracy of 63.24%, whereas CRAFT further reduces the ECE to 5.24% while slightly improving accuracy to 63.29%. Similarly, applied to PTA (Huang et al., 2026), CRAFT reduces the average ECE from 10.51% to 8.15% with comparable average accuracy (69.44% vs. 69.42%). This persistent calibration gap across multiple adapters indicates that simply bounding the feature norm is inadequate. Instead, the combination of preserving constant angular transition via SLERP, dynamically routing weights via GPR, and explicitly bounding confidence via IGFR is collectively required to effectively maintain reliability during adaptation.

Table 14: **Effect of Fusion Geometry on Calibration.** We compare our proposed CRAFT framework against a simple L_2-normalized Euclidean fusion baseline, built on top of BoostAdapter, TCA, and PTA. While simple normalization constrains the feature norm, empirical results demonstrate that CRAFT consistently reduces the Expected Calibration Error (ECE) across most datasets without compromising accuracy.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
BoostAdapter (Zhang et al., 2024)													
+ L_2-Norm. Euclidean	Acc.	**69.40**	**44.74**	**70.89**	**86.18**	67.33	27.21	**88.77**	**94.08**	**69.20**	47.35	**69.58**	66.79
	ECE	3.67	14.49	8.99	4.62	12.18	11.51	3.16	2.95	10.56	10.01	4.45	7.87
+ CRAFT (Ours)	Acc.	69.39	**44.74**	**70.89**	**86.18**	**67.35**	**27.27**	**88.77**	**94.08**	69.15	**47.44**	**69.58**	**66.80**
	ECE	**3.47**	**11.34**	**8.49**	**4.57**	**11.80**	**10.79**	**3.06**	**2.49**	**9.98**	**6.40**	**4.35**	**6.98**
TCA (Wang et al., 2025)													
+ L_2-Norm. Euclidean	Acc.	64.41	**43.97**	69.67	**79.25**	**61.74**	**22.50**	**85.20**	**91.32**	64.05	57.17	56.39	63.24
	ECE	4.14	3.70	9.06	**2.20**	4.91	10.20	**4.50**	**3.07**	5.65	**9.77**	**1.99**	5.38
+ CRAFT (Ours)	Acc.	**64.42**	**43.97**	**69.87**	**79.25**	61.71	22.47	85.17	**91.32**	**64.08**	**57.43**	**56.46**	**63.29**
	ECE	**3.77**	**3.44**	**5.57**	2.52	**4.32**	**9.31**	4.62	3.88	**3.54**	14.60	2.05	**5.24**
PTA (Huang et al., 2026)													
+ L_2-Norm. Euclidean	Acc.	**69.86**	**48.52**	73.85	86.21	68.90	25.89	**91.20**	**94.89**	**74.12**	61.79	68.36	69.42
	ECE	14.25	27.28	8.49	4.72	17.34	19.02	**2.03**	2.67	11.54	**2.30**	6.01	10.51
+ CRAFT (Ours)	Acc.	69.70	47.04	**74.87**	**86.42**	**69.13**	**26.19**	91.17	94.32	73.86	**62.64**	**68.49**	**69.44**
	ECE	**12.91**	**9.48**	**4.42**	**3.24**	**14.71**	**14.41**	3.04	**1.85**	**6.96**	13.15	**5.46**	**8.15**

J Standard Deviation Analysis

In this section, we report the standard deviations associated with the performance metrics presented in Table 15 of the main text. The results were obtained across 3 independent random seeds for both the TDA and PTA baselines, along with their respective CRAFT variants, evaluated on 11 benchmark datasets.

As shown in Table 15, the evaluated methods exhibit low variance across all datasets. For the PTA configurations, the average standard deviation for Accuracy (Acc.) across the 11 datasets is ± 0.15 for the baseline and ± 0.17 for PTA + CRAFT. For Expected Calibration Error (ECE), the

Table 15: Standard deviation of Accuracy (Acc.) and Expected Calibration Error (ECE) for the TDA and PTA, and their CRAFT counterparts across 11 benchmark datasets, calculated over 3 random seeds.

Method	Metric	ImageNet	DTD	Flower	Food101	SUN397	Aircraft	Pets	Caltech	UCF101	EuroSAT	Cars	Avg
TDA	Acc.	±0.02	±0.61	±0.11	±0.04	±0.07	±0.17	±0.16	±0.07	±0.43	±0.42	±0.08	±0.20
	ECE	±0.04	±0.63	±0.32	±0.05	±0.05	±0.09	±0.12	±0.20	±0.57	±0.20	±0.14	±0.22
+ CRAFT	Acc.	±0.01	±0.35	±0.18	±0.03	±0.01	±0.18	±0.11	±0.14	±0.31	±0.94	±0.07	±0.21
	ECE	±0.03	±0.60	±0.25	±0.03	±0.09	±0.44	±0.10	±0.17	±0.31	±0.99	±0.16	±0.29
PTA	Acc.	±0.10	±0.39	±0.24	±0.02	±0.04	±0.23	±0.10	±0.07	±0.22	±0.14	±0.12	±0.15
	ECE	±0.10	±0.03	±0.02	±0.03	±0.06	±0.48	±0.15	±0.17	±0.22	±0.30	±0.23	±0.16
+ CRAFT	Acc.	±0.08	±0.53	±0.32	±0.04	±0.05	±0.19	±0.12	±0.17	±0.18	±0.12	±0.03	±0.17
	ECE	±0.06	±0.33	±0.16	±0.04	±0.03	±0.32	±0.11	±0.39	±0.12	±0.16	±0.06	±0.16

average standard deviation is ± 0.16 for both the PTA baseline and the CRAFT variant. Similarly, the TDA baseline and the TDA + CRAFT variant demonstrate comparable stability, yielding minimal standard deviations for both Accuracy and ECE across the evaluations. These consistently low standard deviations indicate that the performance improvements, particularly the ECE reductions achieved by CRAFT, are robust and stable across different random initialization seeds.

AI USE STATEMENT

In this work, we used generative AI tools for drafting sections of the paper, polishing the English text, grammar correction, and improving readability. We have not used generative AI tools for deriving mathematical proofs, synthesizing primary literature, or formulating novel scientific claims, and generating experimental data with AI is not applicable to this work. Additionally, we used generative AI tools for formatting LaTeX tables and generating boilerplate code for data visualization. We have reviewed all AI-assisted work. Specifically, all AI-assisted drafts and text revisions were meticulously cross-checked by the authors to ensure scientific accuracy and preserve the original meaning. Furthermore, any AI-generated code snippets were thoroughly verified, executed, and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies calibrated test-time adaptation for vision-language models using the publicly available ImageNet, its distribution shift variants (ImageNet-A, -V2, -R, -Sketch), 10 fine-grained image classification datasets, as well as 3D point cloud benchmarks including ModelNet-C and ScanObjectNN. The study does not involve human participants or the collection of personal data.

REPRODUCIBILITY STATEMENT

Section 4 describes the proposed CRAFT framework, including Hyperspherical Concentration-Invariant Fusion (HCIF), Gaussian Probabilistic Routing (GPR), and Information-Geometric Fisher-Rao Retraction (IGFR). Section 5.1 specifies the benchmarks, baseline models, and evaluation metrics. Appendices A and B provide the mathematical proofs and theoretical analysis of norm inflation under the von Mises-Fisher model. Appendices C through J provide additional implementation details for the non-i.i.d. streaming scenarios, 3D point cloud experiments, component-ablation settings, and standard deviation analyses.