Quantile-Adaptive Temperature Scaling for Confidence Calibration

Omprakash Chakraborty[1], Leo Fillioux[2], Ismail Ben Ayed[1], and Jose Dolz[1]

[1] ÉTS Montréal, Canada
[2] Université Paris-Saclay, CentraleSupélec, Gustave Roussy, INSERM, CDSU, IHU PRISM, Gif-sur-Yvette

Abstract. Deep neural networks often produce poorly calibrated confidence estimates, overstating their certainty even when predictions are incorrect. Temperature Scaling (TS) remains the most widely used post-hoc calibration method due to its simplicity and effectiveness, yet its global, uniform rescaling of logits fails to correct the highly heterogeneous structure of miscalibration observed across the confidence spectrum. In particular, the largest correctness–confidence discrepancies arise in different quantile regions depending on the setting, which standard TS leaves largely unaddressed. We introduce Quantile-Adaptive Temperature Scaling (QaTS), a simple and efficient post-hoc calibration method that adapts the temperature as a function of a prediction's empirical confidence quantile. By mapping confidences into the quantile space, QaTS normalizes the calibration problem, makes the structure of miscalibration explicit, and enables a monotone temperature function that adapts across quantiles while leaving well-calibrated high-confidence predictions largely unchanged. This quantile-aware formulation aligns naturally with a reparameterized Expected Calibration Error (ECE) objective and yields a sample-wise temperature that is robust across a variety of challenging scenarios, such as class imbalance and distributional shifts. Across a broad range of datasets, architectures, evaluation scenarios and diverse tasks, QaTS consistently, and substantially, outperforms state-of-the-art post-hoc calibration methods, delivering more reliable and trustworthy confidence estimates without modifying model predictions.

Keywords: Uncertainty Calibration · Post-hoc Calibration

1 Introduction

Confidence calibration, the degree to which predicted probabilities reflect actual correctness, plays a vital role in ensuring reliable risk estimation for decision-making systems. Poor calibration can distort perceived failure risk and, in turn, drive suboptimal or misguided decisions. Although deep learning models have substantially improved predictive discrimination, they often suffer from over-confidence, a widely recognized shortcoming that makes calibration a persistent challenge [10, 25]. This issue becomes especially consequential in high-stakes domains such

as healthcare, autonomous driving, and critical infrastructure, motivating a substantial body of work on improving calibration of deep models [20,22,27,29].

While training-based [22,23,27] and post-hoc [10,12,40] confidence calibration approaches exist, the latter offer a more appealing and practical solution in real-world settings, since they avoid costly retraining, do not require access to model weights or gradients, and scale seamlessly to large pretrained models and frozen backbones. In particular, Temperature scaling (TS) [10], a single-parameter variant of Platt Scaling [34], calibrates a model by applying a single scalar temperature to all logits, uniformly shrinking or expanding confidence scores without altering the prediction. Several extensions have been proposed, including class-wise [15], prediction-wise [40] or input-dependent [5] temperatures, often learned through small auxiliary networks. While these variants increase flexibility, they all share the same underlying assumption: miscalibration can be corrected by a smooth, globally consistent rescaling function that applies uniformly across the confidence spectrum.

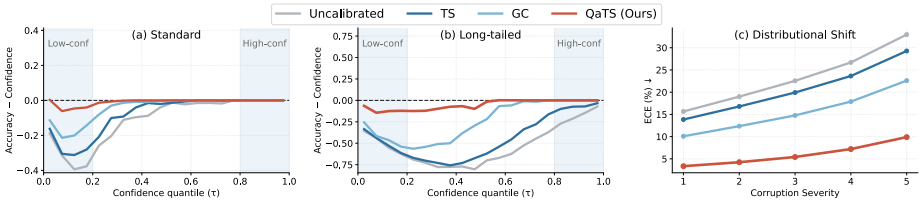

Fig. 1: Calibration behavior across regimes. Calibration gap (Acc − Conf) under (a) *standard* and (b) *long-tailed* setting. (c) ECE under increasing severity of *distributional shift*. Across all regimes, QaTS consistently reduces calibration bias and error, while maintaining robustness under severe distributional shifts.

Nevertheless, our quantile-wise analysis in Fig. 1 reveals that miscalibration (*gray*) is far from uniform across the whole confidence spectrum. In the standard regime (*left plot*), the largest correctness–confidence gaps consistently occur in the lowest confidence quantiles, precisely where uncertainty is highest and reliable confidence matters most, while mid- and high-quantile predictions exhibit substantially smaller discrepancies. In the long-tailed regime (*middle*), this structure changes: the dominant errors shift toward the medium-confidence quantiles, driven by the prevalence of mid-frequency classes that produce many moderately confident yet error-prone predictions. Across these settings, both the shape and location of miscalibration vary substantially, indicating that calibration errors are inherently heterogeneous and regime-dependent. This pronounced heterogeneity further exposes a structural limitation of existing temperature-scaling approaches, i.e., they implicitly assume that miscalibration can be corrected by a single, smooth, globally consistent transformation of the logits. Such uniformity assumptions are fundamentally mismatched with the empirical behavior shown in Fig. 1. Methods like TS [10] and its more advanced versions, such as GC [45],

apply a single functional mapping from logits to temperatures, treating all confidence levels as if they require the same type of correction. Because they do not condition on confidence quantiles, they cannot effectively adapt their corrections to regions where miscalibration is more severe. As a result, these methods often reduce global ECE while leaving the dominant quantile-wise discrepancies substantially uncorrected.

A direct implication of this structural rigidity is the degraded performance of TS-based methods under more challenging scenarios, such as distributional drifts (Fig. 1, *right*). Covariate perturbations often alter the absolute scale of confidence values, while largely preserving their relative ordering. Because the temperature function $T(x)$ in QaTS depends on the empirical confidence quantile τ rather than the raw confidence magnitude p (as in TS-based approaches), it inherits the rank-based stability of quantiles, i.e., they are stable under monotone transformations of the logits, and thus substantially less sensitive to global logit scaling effects that often accompany covariate shift, as long as the relative ordering of confidences is approximately preserved. As a result, QaTS applies its calibration correction consistently across quantiles, even when test samples differ from calibration samples. As long as the ordering of confidence values is approximately preserved (a mild assumption satisfied by most covariate shifts) the quantile-adaptive temperature continues to reduce the dominant contributions to the calibration error in test samples, yielding robust calibration across domains. These observations motivate a calibration strategy that operates directly in quantile space, where the structure of miscalibration becomes explicit and can be corrected in a targeted, distribution-aware manner.

To address these limitations, we propose Quantile-Adaptive Temperature Scaling (QaTS), a calibration method explicitly designed to correct the heterogeneous structure of miscalibration revealed by our quantile-wise analysis. Instead of applying a single global temperature, or predicting temperatures from logits, features, or auxiliary networks, QaTS introduces a sample-wise temperature $T(x)$ that adapts to the empirical confidence quantile of each prediction. By mapping raw confidences into quantile space, QaTS obtains a normalized, distribution-aware measure of how "typical" or "atypical" a prediction is relative to the calibration set. This enables a monotone temperature function that selectively softens low-quantile (overconfident) predictions and sharpens high-quantile ones, improving calibration without altering prediction rankings. Combined with surrogate-based training using strictly proper scoring rules, QaTS provides a simple, efficient, and robust mechanism for correcting miscalibration exactly where it is most severe, yielding confidence estimates that remain reliable even under distribution shift.

We can summarize our contributions as follows:

- We provide a quantile-wise analysis of miscalibration, revealing that calibration errors are highly heterogeneous across the confidence quantile spectrum, and vary substantially across scenarios (e.g., standard vs long-tailed). This analysis reveals structural limitations of existing TS-based methods, which cannot directly target the regions where miscalibration is most pronounced.

- Based on these observations, we introduce a simple yet effective strategy, QaTS, for calibrating neural networks that directly targets the heterogeneous structure of miscalibration, operating directly in quantile space. We formalize a quantile reparameterization of the Expected Calibration Error (ECE), which expresses calibration error as a function of confidence quantiles rather than raw confidence values. This reparametrization exposes where miscalibration is concentrated and enables a temperature function that adapts to the relative rank of each prediction, yielding calibration that is quantile-aware, showing practical benefits in challenging scenarios.
- Comprehensive experiments across multiple benchmarks and distinct tasks, including standard computer vision and medical image classification, text classification and semantic segmentation, as well as diverse scenarios, such as long-tailed distributions and distributional drifts, show that QaTS yields superior uncertainty calibration than state-of-the-art post-hoc methods.

2 Related Work

Uncertainty calibration of deep neural networks has emerged in the last years as an active area of study, with numerous studies analyzing and characterizing the calibration behavior of modern models [10,25]. These methods can be mainly categorized into: *training-time* and *post-hoc* calibration.

Training-time calibration aims at enhancing model calibration during training, typically by integrating additional regularization objectives into the loss function [3,15,22,23,29,32,33]. A popular strategy is to introduce explicit penalties that either penalize overconfident softmax predictions [3,19,32,33] or encourage small logit differences [22,23,29]. Beyond such explicit objectives, several works have shown that widely used classification losses, such as Focal Loss [21] and Label Smoothing [37], implicitly promote higher-entropy predictions, thereby yielding models with lower confidence [27,28]. Additional training-time techniques include data-mixing strategies such as MixUp [39,52] and constraints on the geometry of the logit space, for example by enforcing a fixed logit-norm [43] or logit-range [30]. A common limitation of training-time methods is that they are tightly coupled to the optimization of the underlying classifier, which restricts their use to settings where the model can be modified and retrained.

Post-hoc calibration, in contrast, adjusts the output logits of a pre-trained model without modifying its training procedure [10–12,14,31,40], offering the practical advantage that calibration can be applied independently of the training process and even to fixed or proprietary models. A notable method is temperature scaling (TS) [10], a simplified form of Platt scaling [34], which applies a single global temperature T. While effective at reducing overconfidence on average, global scaling cannot adapt to input-dependent uncertainty, and may under- or over-correct different confidence ranges. More flexible non-parametric techniques, such as Histogram Binning [49] and Isotonic Regression (IR) [50], learn piecewise mappings that can substantially improve calibration. Nevertheless, they lack structural constraints such as smoothness or monotonicity beyond

local segments, which can lead to unstable mappings that generalize poorly to unseen confidence ranges. Recent work explores parameterized temperature scaling, where temperatures depend on the class [7], or on the input via a small auxiliary network [5,40], increasing flexibility while retaining the accuracy-preserving nature of classical scaling. Furthermore, improving calibration has recently been explored from the perspective of the feature space, either by modulating the temperature as a function of learned features [44,45] or by selectively clipping feature magnitudes to increase predictive entropy on high-error samples while preserving discriminative information on well-calibrated ones [38]. AdaTS [12] leverages test-time augmentation by generating multiple transformed versions of each input and averaging their logits to obtain a better-calibrated prediction, but this comes at the cost of substantially increased inference-time computation. Other approaches have extended TS to multiple domains [48], or under distributional drifts [16,41,48], but they require either direct access to samples of each domain [9,16,48], or synthetically generate larger validation sets through image transformations [41], which attempt to simulate domain shift conditions.

In contrast to all existing techniques, the proposed QaTS relies on confidence quantiles, rather than in raw confidence scores, which makes the method inherently robust to the scale, distribution, and calibration quality of the underlying classifier. By operating on rank-based information, QaTS preserves ordering while eliminating dependence on absolute score values, enabling consistent behavior across architectures, datasets, diverse tasks and scenarios.

3 Problem Definition and Motivation

Preliminaries. Let $\mathcal{X} \subset \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{1, \dots, K\}$ the label space. We assume a K-class trained classifier is a mapping function $\phi : \mathcal{X} \to \Delta_K$, where $\Delta_K = \{\mathbf{p} \in \mathbb{R}^K : \sum_{k=1}^K p_k = 1,\ p_k \geq 0\}$ is the probability simplex. Furthermore, the probability predictor ϕ is composed of a feature extractor $g : \mathcal{X} \to \mathbb{R}^K$ that produces the logits vector $\mathbf{l} = (\ell_k)_{1 \leq k \leq K}$, followed by the softmax operator, defined by $\mathtt{sm}$, so that $\phi = \mathtt{sm} \circ g$. Thus, for an input $x \in \mathcal{X}$, we can compute the softmax probability vector $\mathbf{p}(x) = \phi(x) = (p_1(x), \dots, p_K(x))$, as

$$p_k(x) = \frac{\exp(\ell_k(x))}{\sum_{j=1}^K \exp(\ell_j(x))}, \quad k = 1, \dots, K. \tag{1}$$

From this vector, we can now compute the predicted class as $\hat{y}(x) = \arg\max_k p_k(x)$, whose associated confidence is $\mathrm{conf}(x) = \max_k p_k(x)$.

In our post-hoc calibration scenario, we have access to a pre-trained model ϕ, and a calibration set $\mathcal{D}_{cal}$ of $\mathcal{X} \times \mathcal{Y}$, where $X \in \mathcal{X}$ contains *i.i.d.* samples from an unknown distribution with $Y \in \mathcal{Y}$. We can now define a post-hoc calibration function $f : \mathbb{R}^K \to \mathbb{R}^K$, yielding the K-class *calibrated* predictor as $\phi_c = \mathtt{sm} \circ f \circ g$.

Definition 1 (Perfect Calibration). *Let (X, Y) be random variables taking values in $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, \dots, K\}$, respectively, drawn from an unknown distribution π. Let $\phi_c : \mathcal{X} \to \Delta_K$ be a K-class probability predictor, with the*

definitions of $\hat{y}(X) := \arg\max_k p_k(X)$*, and* $\hat{p}(X) := \max_k p_k(X)$*, where* $\mathbf{p}(X) = \phi_c(X)$*, we say that* ϕ_c *is perfectly calibrated if:*

$$\mathbb{P}\big(Y = \hat{y}(X) \mid \hat{p}(X) = p\big) = p, \qquad \forall p \in [0, 1]. \tag{2}$$

Thus, our objective is to train a post-hoc calibration function f that satisfies *Definition 1*, i.e., yields a perfectly calibrated predictor. To this end, we optimize the Negative Log Likelihood (NLL) loss, a standard choice in calibration [10,51].

Expected calibration error as a discrepancy between correctness and confidence. Following Definition 1, let $C(p)$ denote the *correctness* function for a scalar confidence value $p \in [0, 1]$ as

$$C(p) = \mathbb{P}(Y = \hat{y}(X) \mid \text{conf}(X) = p), \tag{3}$$

and let μ denote the distribution of $\text{conf}(X)$ on [0,1], where $\text{conf}(X) := \hat{p}(X)$. The Expected Calibration Error (ECE), in expectation, can be then written as:

$$\text{ECE} = \int_0^1 \big|C(p) - p\big| \, d\mu(p), \tag{4}$$

which measures the $L^1(\mu)$ distance between correctness and confidence.

Quantile reparameterization. The metric in Eq. 4 weights miscalibration in a pointwise manner, based on how often each confidence value occurs. In this scenario, dense regions dominate on the final value, whereas rare confidence values contribute very little. This leads to large errors concentrated in rare confidence regions to be masked by small errors in common regions, yielding a misleading low global ECE. To avoid this, we first define an empirical quantile function. To do so, we first sort the confidences of all calibration samples, $\text{conf}(x^{(1)}) \leq \text{conf}(x^{(2)}) \leq \cdots \leq \text{conf}(x^{(N)})$, and define the empirical quantile function as:

$$F_{\text{conf}}(u) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\text{conf}(x^{(i)}) \leq u\}, \tag{5}$$

which returns the empirical cumulative distribution function (CDF) value of a given input confidence u. Thus, for an image x, we can first compute its empirical quantile as:

$$q(x) = F_{\text{conf}}(\text{conf}(x)). \tag{6}$$

Now, we can reparameterize Eq. 4 by quantiles $\tau = F_{conf}(p)$, so that the same error is integrated uniformly over $\tau \in [0, 1]$. To do this, let $q(x) = F_{\text{conf}}(\text{conf}(x))$ be the empirical quantile of the confidence. Under mild regularity assumptions, $q(X)$ is asymptotically uniform on $[0, 1]$ and the change of variables $p = F^{-1}_{\text{conf}}(\tau)$ yields the equivalent representation

$$\text{ECE} = \int_0^1 \big|C(F^{-1}_{\text{conf}}(\tau)) - F^{-1}_{\text{conf}}(\tau)\big| \, d\tau. \tag{7}$$

This reparameterization leaves ECE unchanged but expresses it on a uniform quantile axis, preventing dense confidence regions from dominating the representation of calibration error. Under this view, ECE becomes the L^1 discrepancy

between correctness and confidence *expressed along the uniform quantile axis*. This representation highlights that miscalibration is often *heterogeneous* across quantiles, and can vary under different conditions, as empirically shown in Fig. 1.

4 Adaptive Quantile Temperature Scaling

Sample-dependent Temperature. To account for individual sample uncertainties when scaling the logits in the softmax function, we propose a sample-wise temperature function, defined as a monotone linear function of $(1 - q(x))$:

$$T(x) = a \cdot (1 - q(x)) + b, \quad q(x) \in [0, 1], \tag{8}$$

where a and b are two positive learnable parameters that modulate the final sample temperature value according to its quantile. This ensures that lower-quantile (i.e., more problematic) predictions receive a different scaling than higher-quantile (i.e., less problematic) ones. Unlike standard TS-based approaches, QaTS conditions the temperature on the sample's confidence quantile, not on its raw logit magnitude. This quantile-based parameterization breaks the global-smoothness assumption and enables corrections that vary systematically across the confidence spectrum, aligning the calibration mechanism with the heterogeneous structure revealed by the quantile-wise analysis (Fig. 1).

Learning the temperature parameters a and b. The parameters a and b are optimized using the calibration (or validation) set so that the scaled probabilities

$$\tilde{p}_k(x) = \frac{\exp(\ell_k(x)/T(x))}{\sum_{j=1}^{K} \exp(\ell_j(x)/T(x))} \tag{9}$$

better reflect empirical correctness frequencies.

Direct optimization of Expected Calibration Error (ECE) is infeasible due to its non-differentiable binned form. Instead, we fit the parameters (a, b) using smooth surrogates such as the negative log-likelihood (NLL), a strictly proper scoring rule. Minimizing this surrogate encourages the calibrated probabilities to align with empirical correctness frequencies, thus reducing calibration error in practice. Specifically, we minimize the NLL over the set of calibration samples:

$$\min_{a,b>0} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\big(\tilde{\mathbf{p}}(x^{(i)}), \mathbf{y}^{(i)}\big) = \min_{a,b>0} -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \log(\tilde{p}_k(x^{(i)}; a, b)). \tag{10}$$

where $\mathbf{y}^{(i)}$ is the ground-truth label of calibration sample $x^{(i)}$, and $\mathcal{L}$ denotes the chosen loss function. We also constraint $a, b > 0$, ensuring a strictly positive, monotonically decreasing temperature over $q \in [0, 1]$. This training procedure yields optimal values of a^* and b^* that adaptively adjust the temperature based on each prediction's confidence quantile, directly targeting the heterogeneity revealed by the quantile reparameterization and thereby improving ECE.

Inference. For each test sample x, we simply scale each predicted logit by the sample-wise temperature scale function in Eq. (8), which dynamically adapts to

the underlying properties of each sample:

$$\tilde{\ell}_k(x) = \frac{\ell_k(x)}{T(x)}, \quad k = 1, \ldots, K. \tag{11}$$

Benefits of the proposed approach. Considering the definition of the empirical quantile in Eq. 6, our method introduces a quantile-dependent temperature value, defined in Eq. 8, which is used to rescale the logits in (11). As $q(x)$ indexes the position of a prediction along the uniform quantile axis, adjusting $T(x)$ as a function of $q(x)$ directly modifies the confidence *precisely in the regions where Eq. (7) shows the largest contribution to ECE.* In particular, *i)* low-quantile predictions (where overconfidence is typically largest) receive higher temperatures, reducing p and decreasing $|C(p) - p|$ in Eq. (7); and *ii)* high-quantile predictions receive smaller adjustments, preserving accuracy and avoiding unnecessary distortion. Since the quantile axis is uniform, any reduction of the discrepancy in the regions with the largest quantile-wise error leads to a monotone decrease of the global ECE. Thus, quantile-adaptive temperature scaling is explicitly aligned with minimizing Eq. (7). Beyond this alignment with the calibration objective, the quantile representation also provides a key advantage under challenging scenarios, such as domain shift. Mapping confidences through the calibration CDF yields a monotone transform, so QaTS depends mainly on the relative ordering of scores rather than their absolute scale. This makes the temperature function more stable under shifts that preserve approximate ranking, even when logits are globally rescaled or shifted. By learning a simple temperature function $T(x) = a \cdot (1 - q(x)) + b$ on these quantile values, the model captures how calibration should vary as a function of relative confidence, not absolute scale. This means that, at deployment, even in the presence of distributional drifts, the ordering of predictions remains far more consistent than their raw magnitudes on these quantile values, the model captures how calibration should vary as a function of *relative* confidence rather than absolute scale.

5 Experiments

Datasets. QaTS is empirically validated across several computer vision classification tasks, including CIFAR-10 and CIFAR-100 [17], ImageNet [35], as well as their long-tailed counterparts [1], with different degrees of data imbalance ratios γ, and ImageNet-LT [24]. Further, we evaluate the robustness of QaTS under distributional shifts by employing corrupted benchmarks such as CIFAR-100-C [13], which introduces 15 corruption types across 5 severity levels. To assess the performance of the proposed method in real-world scenarios, we perform experiments in medical imaging classification tasks, by resorting to HAM10000 [42] and to MedMNIST2D [46,47], which consists of 10 pre-processed 2D sources covering primary data modalities (e.g., X-Ray, OCT, Ultrasound, and CT), and diverse binary and multi-class classification tasks. Last, to show the general applicability of our approach, we further evaluate it on the 20 Newsgroups dataset [26], a

popular Natural Language Processing benchmark for text classification and PascalVOC [6], a well-established image segmentation dataset. Thus, we evaluated QaTS across **162 different settings.**

Baselines. ***Accuracy preserving methods:*** We benchmark QaTS to post-hoc calibration approaches that preserve accuracy, including the raw uncalibrated model (Uncal), temperature scaling (TS) [10], Isotonic Regression (IR) [50], and the more recent AdaTS [12] and Group Calibration (GC) [45], where we report the values of the proposed method combined with temperature scaling, which yields the most competitive performance. ***Non-accuracy preserving methods:*** We also include FeatClip [38], a very recent post-hoc calibration approach that modifies model predictions. Importantly, as highlighted in prior works [45], such approaches often have a noticeable negative impact on model accuracy. Our empirical evaluation confirms this trend, where FeatClip's accuracy degradation becomes substantial in challenging settings such as class-imbalanced scenarios.

Evaluation metrics. To measure the discriminative performance of the different methods, we use classification accuracy (ACC) for classification tasks, and mean Intersection over Union (mIoU) for segmentation. In terms of calibration, we follow the standard literature and resort to the Expected Calibration Error (ECE). In particular, with N samples grouped into M bins $\{b_1, b_2, \ldots, b_M\}$, the ECE is calculated as: $\sum_{m=1}^{M} \frac{|b_m|}{N} |\mathrm{acc}(b_m) - \mathrm{conf}(b_m)|$, where $\mathrm{acc}(\cdot)$ and $\mathrm{conf}(\cdot)$ denote the average accuracy and confidence in bin b_m.

Implementation Details. For image classification, we experiment with multiple convolutional neural networks, such as ResNet-18, ResNet-50 and DenseNet-121, as well as Vision Transformers, including ViT-S/16 and ViT-B/16. We use $M = 15$ bins for ECE computation across all datasets and architectures. Furthermore, we use BERT-B [4] on the NLP recognition task and DeepLabV3 [2] for semantic segmentation on PASCAL VOC2012.

5.1 Results

A note on classification performance. Post-hoc calibration approaches are often designed to operate by reshaping the softmax (or logit) distributions, without altering the predicted class. Nevertheless, the underlying mechanism of several recent methods improve calibration, but at the cost of degrading their discriminative performance. FeatClip [38], a recent state-of-the-art post-hoc calibration approach, is a notable example: *its calibration gains come with measurable accuracy degradation.* To contextualize the calibration results presented in this section, we first report the classification accuracy of FeatClip and compare it to our method, which preserves the predicted class by design. This allows readers to appreciate how certain calibration techniques may degrade accuracy before focusing solely on their calibration behavior. In particular, Fig. 2 presents a pair-wise comparison across 161 dataset–architecture combinations spanning standard, long-tailed, domain-shift, medical and text classification benchmarks. Each point corresponds to the accuracy difference between QaTS and FeatClip (ΔAcc). Overall, **QaTS yields higher accuracy in 94% of the settings.** Importantly, these gains are consistent across diverse regimes, indicating that

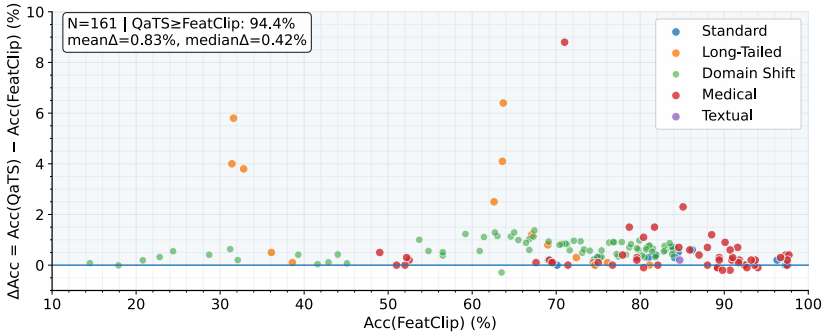

Fig. 2: Accuracy dominance of QaTS over FeatClip across 161 evaluation settings spanning standard, long-tailed, domain-shift, medical and text classification benchmarks. Each point corresponds to the Top-1 accuracy difference for a dataset-architecture pair. Points above zero indicate better accuracy for QaTS.

the calibration improvements of QaTS are not obtained at the expense of discriminative performance. This reinforces the importance of assessing calibration methods alongside classification accuracy rather than in isolation.

Table 1: Calibration performance (ECE↓) in standard image classification benchmarks. Results across datasets and backbones, where ‡ indicates that the method modifies model predictions, which impacts the final accuracy. Best method in **bold**, whereas second best underlined. Performance gap wrt second best method is highlighted as **better** or **worse**. Same convention applies throughout the paper.

Dataset	Backbone	Uncal	TS	IR	AdaTS	GC	FeatClip‡	QaTS
CIFAR-10	RN-50	1.71	1.23	0.64	0.64	0.78	0.55	**0.31**$_{-0.24}$
	RN-18	1.93	0.77	0.59	0.62	0.70	0.50	**0.32**$_{-0.18}$
	DN-121	1.44	1.02	0.67	0.62	0.50	0.58	**0.29**$_{-0.21}$
	ViT-S	1.80	1.01	**0.45**	0.67	1.01	0.62	0.51$_{+0.06}$
	ViT-B	1.97	1.61	0.65	1.18	1.19	0.54	**0.49**$_{-0.05}$
CIFAR-100	RN-50	8.62	5.32	3.08	3.93	3.99	3.01	**2.56**$_{-0.45}$
	RN-18	7.87	3.18	1.71	2.91	2.37	1.70	**1.32**$_{-0.38}$
	DN-121	7.60	5.15	2.91	3.12	3.12	2.71	**2.45**$_{-0.26}$
	ViT-S	7.98	6.74	2.73	3.62	3.37	2.92	**2.54**$_{-0.19}$
	ViT-B	10.79	9.43	4.67	5.34	5.35	2.57	**2.26**$_{-0.31}$
ImageNet	RN-50	3.71	2.26	2.72	2.05	2.21	1.67	**1.58**$_{-0.09}$
	RN-18	2.61	1.86	1.50	1.71	1.84	1.41	**1.34**$_{-0.07}$
	DN-121	2.61	1.89	2.10	1.84	1.79	1.81	**1.71**$_{-0.08}$
	ViT-S	2.11	1.95	2.03	1.28	1.51	1.51	**1.13**$_{-0.15}$
	ViT-B	5.77	3.28	2.80	3.01	3.28	2.93	**2.49**$_{-0.44}$

A. Standard CV benchmarks. Table 1 reports the calibration results on CIFAR-10, CIFAR-100 and ImageNet across five architectures. Overall, **our approach consistently achieves the lowest ECE,** outperforming all competing post-hoc baselines in all but one setting, where it ranks second. On CIFAR-

Table 2: Calibration performance in long-tailed image classification.

Dataset	Backbone	Uncal	TS	IR	AdaTS	GC	FeatClip‡	QaTS
R-10 (mild)	RN-50	19.53	19.23	5.37	16.68	8.05	6.62	**3.76**$_{-2.86}$
	RN-18	19.11	18.20	5.31	15.43	6.78	4.73	**3.54**$_{-1.19}$
	DN-121	19.13	18.24	5.40	15.83	7.38	6.59	**3.60**$_{-2.99}$
	ViT-S	19.14	19.07	5.80	16.75	8.36	**3.19**	3.40$_{+0.21}$
	ViT-B	24.12	24.18	5.24	21.91	12.60	3.66	**3.50**$_{0.16}$
R-100 (extreme)	RN-50	48.64	44.27	27.20	34.74	25.84	**4.78**	5.06$_{+0.28}$
	RN-18	45.80	40.50	25.89	33.74	21.51	2.37	**2.30**$_{-0.07}$
	DN-121	47.27	42.45	26.86	34.12	23.63	4.87	**3.07**$_{-1.80}$
	ViT-S	50.24	46.41	30.71	35.55	29.74	9.73	**8.19**$_{-1.54}$
	ViT-B	52.51	49.15	33.82	37.13	32.89	**8.34**	8.37$_{+0.03}$
ImageNet-LT	RN-50	3.71	4.16	4.50	3.72	4.03	3.35	**2.68**$_{-0.67}$
	RN-18	2.61	2.86	2.56	2.56	4.06	2.55	**2.01**$_{-0.54}$
	DN-121	2.61	2.73	2.75	2.42	2.65	**1.51**	1.62$_{+0.11}$
	ViT-S	2.81	3.51	2.71	2.76	2.97	1.96	**1.93**$_{-0.03}$
	ViT-B	5.77	5.28	5.66	5.61	5.60	5.59	**3.59**$_{-2.00}$

10, QaTS yields ECE of 0.29 (DN-121) and 0.31 (RN-50), yielding clear gains over the second-best method overall, i.e., FeatClip, across CNN and transformer backbones. On CIFAR-100, improvements are more pronounced, where QaTS consistently outperforms the FeatClip by margins up to 0.45 (RN-50) and 0.38 (RN-18). Similarly, on ImageNet, QaTS achieves the best calibration across all architectures, **reducing ECE by up to 0.44** (ViT-B) compared to FeatClip.

B. Long-tailed distributions (Table 2). We now report the calibration results under class-imbalance conditions, including CIFAR-100-LT with *mild* (R-10) and *extreme* (R-100) imbalance. In these scenarios, miscalibration is substantially amplified due to the skewed class-frequency distribution, where dense confidence regions are dominated by head-class predictions. Standard scalar calibration methods (e.g., TS), and recent strategies such as AdaTS and GC struggle to correct this heterogeneous behavior, leading to notably large miscalibration errors. Indeed, in some cases, these approaches further deteriorate the baseline performance (e.g., results in ImageNet-LT), rendering them ineffective in these regimes. In contrast, QaTS yields substantial improvements, often achieving **reductions of 80-90%** (see AdaTS and GC results in R-10 and R-100 settings in CIFAR-100-LT). Moreover, while most competing methods degrade sharply as the imbalance increases from R-10 to R-100, QaTS remains comparatively stable, underscoring its robustness to severe class imbalance. Finally, FeatClip emerges as a competitive alternative in long-tailed scenarios, ranking second in most settings and first in four cases. However, we must highlight two important observations. First, when QaTS ranks second, the ECE gap is marginal. And second, as shown in Fig. 2, FeatClip's accuracy is substantially degraded, sometimes **decreasing by 4%-6% compared to QaTS,** highlighting QaTS's ability to enhance calibration without compromising accuracy.

C. Performance under distributional drifts. When considering real-world deployment, effective calibration methods should demonstrate robustness in handling samples from unknown distributions. To assess this capability in a con-

Table 3: Calibration performance under distributional drifts (CIFAR100 used for calibration, with ViTB-16 as backbone). Note that HFE [16]* needs access to OOD samples of each target domain. Extended results in **Appendix**, Table 1.

OOD-free?	✔	✔	✔	✔	✔	✔	✔	✘
Severity	**Uncal**	**TS**	**IR**	**AdaTS**	**GC**	**FeatClip**‡	**QaTS**	**HFE***
1	15.66	13.85	13.78	13.61	10.08	3.69	**3.35**$_{-0.34}$	5.44
2	19.00	16.79	14.40	16.81	12.36	4.69	**4.24**$_{-0.45}$	7.34
3	22.55	19.92	15.63	19.98	14.77	5.78	**5.42**$_{-0.36}$	9.42
4	26.69	23.64	17.54	23.91	17.88	7.70	**7.20**$_{-0.50}$	12.12
5	32.96	29.27	20.55	29.50	22.60	10.65	**9.89**$_{-0.76}$	15.62

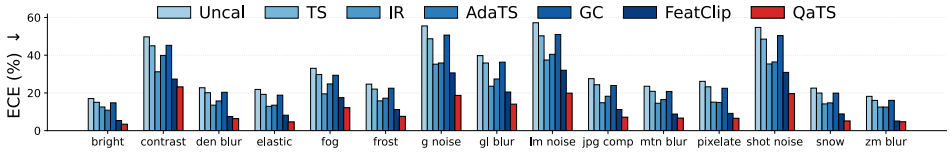

Fig. 3: Calibration under severe domain shift (CIFAR-100-C, severity 5). Per-corruption ECE across all corruption types, highlighting the robustness of QaTS.

trolled and reproducible manner, we now evaluate the performance of all methods on CIFAR-100C, a standard benchmark that introduces a wide range of synthetic corruptions designed to emulate realistic domain shifts such as noise, blur, weather effects, and digital artifacts. To do this, the QaTS parameters (a, b), as well as the parameters for other approaches, are learned on the clean CIFAR-100 validation set and directly applied, without re-estimation, to the corrupted CIFAR-100C samples. We also include HFE [16] as additional baseline, which is specifically designed to operate under distributional drifts. A key limitation of HFE, however, is its reliance on out-of-distribution (OOD) samples for every new target domain, an infeasible requirement in practice because such shifts are typically unknown in advance and cannot be exhaustively anticipated. Table 3 reports average ECE across corruptions for each severity level. Consistent with other settings, QaTS achieves the **lowest ECE among all *OOD-free* methods across all severity levels**. Notably, our approach substantially reduces ECE scores compared to all *accuracy-preserving* methods, **yielding 2×-3× lower ECE** values. In addition, Fig 3 further depicts the ECE by corruption type for severity level 5, demonstrating the superiority of QaTS across corruption types. Concerning *non-accuracy preserving* methods, while performance gap compared to FeatClip is smaller, it widens as corruption severity increases, indicating stronger robustness to severe noise. Finally, even though HFE benefits from access to additional target-domain data, it still exhibits worse calibration than QaTS (Table 3, *last column*), underscoring that our method provides a more effective and practical solution without requiring access to any target samples.

D. Generalization to other tasks and domains. We now assess the performance of QaTS in a broader variety of domains and tasks, detailed below.

Table 4: Calibration performance in medical image classification. Average results across 10 datasets, whose extended results are reported in **Appendix**, Table 2.

Backbone	Uncal	TS	IR	AdaTS	GC	FeatClip‡	**QaTS**
RN-50	7.53	6.87	6.42	6.38	4.98	5.54	**3.56**$_{-1.42}$
RN-18	10.64	9.69	7.53	7.25	5.96	6.67	**4.03**$_{-1.93}$
DN-121	10.09	9.08	7.14	7.17	5.45	5.46	**3.56**$_{-1.89}$
ViT-S	7.78	7.22	7.13	7.12	6.06	4.42	**3.68**$_{-0.74}$
ViT-B	6.58	5.91	5.73	5.97	4.96	3.91	**3.23**$_{-0.68}$

Table 5: Calibration performance (ECE↓) on additional tasks: image semantic segmentation (Pascal VOC-2012) and text classification (NewsGroup-20).

Dataset	**Backbone**	**Uncal**	**TS**	**IR**	**AdaTS**	**GC**	**FeatClip‡**	**QaTS**
VOC-2012	DLV3-R50	4.49	3.69	3.23	3.49	3.26	3.20	**2.94**$_{-0.26}$
NewsGroup-20	BERT-B	5.81	5.42	4.73	4.80	4.53	4.12	**3.80**$_{-0.32}$

• ***Medical Image classification.*** We first include experiments on medical image classification, a domain where calibrated confidence estimates are especially important for reliable clinical decision-making. Average ECE results across 11 datasets for different CNN and ViT backbones are reported in Table 4. Across all backbones, **QaTS consistently achieves the lowest average ECE**, demonstrating robust calibration improvements also in this critical domain.

• ***Image segmentation and Text classification (Table 5).*** Last, to highlight the broad applicability of QaTS, we also assess its performance on a dense predictive task (i.e., semantic segmentation), and a non-vision task (i.e., text classification). It can be observed that the trend is similar to what we observed in image classification tasks, with **QaTS outperforming all the baselines in both tasks**. Furthermore, similar to other scenarios, while FeatClip is competitive in terms calibration, it still degrades the discriminative performance of *accuracy preserving* methods: 69.78→69.51 mIoU in segmentation, and 84.9→84.7 accuracy in text classification. Thus, this comprehensive set of experiments across multiple scenarios, domains and tasks demonstrates that QaTS remains effective across diverse scopes, emerging as a universal calibration strategy.

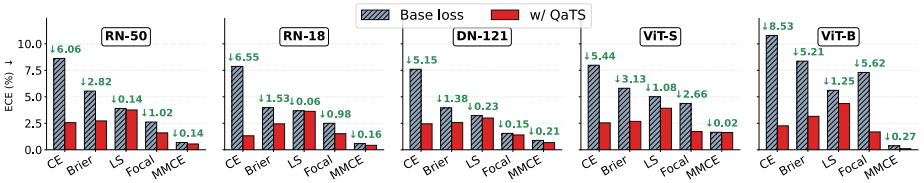

Fig. 4: QaTS complements training-time calibration. ECE (%, ↓) on CIFAR-100 (clean) across backbones and training losses (CE, Brier, LS-0.05, Focal, MMCE), comparing the base loss vs. adding QaTS post-hoc calibration.

E. Compatibility with training-time calibration. We also examine whether QaTS compensates for standard trained models or also complements train-time regularization approaches. We evaluate QaTS on models trained with calibration-aware objectives, including Brier loss [8], Label Smoothing (LS) [37] (with $\alpha = 0.05$), Focal Loss [21], and MMCE [18], which are explicitly designed to reduce overconfidence during optimization. As shown in Fig. 4, while these losses generally reduce ECE compared to standard cross-entropy (CE), substantial miscalibration still remains. Applying QaTS consistently yields further improvements across all backbones. For example, on CIFAR-100 with RN-50, Brier reduces ECE from 8.62 (CE) to 5.54, yet QaTS further lowers it to 2.72. These results demonstrate that QaTS acts as a complementary post-hoc approach that can be seamlessly applied on top of diverse training objectives without modification.

Table 6: Learned QaTS parameters under increasing imbalance severity on CIFAR-100 (i.e., balanced) and CIFAR-100-LT.

Backbone	Balanced		R-10		R-100	
	a	b	a	b	a	b
RN-50	0.0438	1.5960	0.1290	2.0009	1.6389	2.3315
ViT-B/16	0.2814	2.2510	0.3964	2.7824	1.1783	3.0570

F. Analysis on QaTS learned parameters (a, b) (Table 6). Here we analyze the learned parameters (a, b) under varying imbalance severity on CIFAR-100-LT. Across architectures, a and b increase with imbalance, with a growing much more sharply. This indicates that severe imbalance induces highly heterogeneous miscalibration concentrated in the low- and mid-quantile regions (precisely where tail and mid-frequency classes accumulate), promoting QaTS to learn substantially larger temperatures in these problematic regions, where predictions are both less confident and more error-prone. As imbalance increases, QaTS learns a much larger value of a, which increases the temperature selectively in these problematic quantile regions, while b grows only mildly, since high-quantile predictions require little correction. This behavior is consistent with our quantile-wise analysis in Fig.1.

G. Ablation on the temperature parameterization (Table 7). While our default choice for the temperature function $T(q)$ is a simple linear form, more flexible monotone parameterizations could be used. To understand whether this additional flexibility yields measurable benefits, we replace the monotonic linear function in Eq. 8 with a piecewise (PW) linear temperature function defined over K quantile segments (details in **Appendix**, §H). While the piecewise model can slightly reduce ECE on clean datasets, it yields worse results in long-tailed and distributional drifts scenarios. In long-tailed settings, where the calibration set is highly skewed, the larger number of parameters (i.e., $K+1$) tends to overfit the sparsely populated quantile regions, leading to slightly worse generalization. Under distributional drifts, the same higher capacity makes the model more sensitive to confidence patterns that differ from those seen

during calibration, reducing its robustness. This behavior is reflected empirically by the gradual increase of the ECE as K grows. In contrast, the linear form adopted in QaTS, with only two parameters, consistently provides the best robustness–complexity trade-off, capturing the smooth trend of miscalibration across quantiles while avoiding overfitting under complex scenarios. We stress that these observations concern only the choice of the parameterization for $T(q)$ and the underlying quantile-based calibration principle of QaTS remains the same across all monotone function classes.

Table 7: Ablation on temperature function parameterizations (ViTB-16).

Method	C100	C100-LT	C100-C
Uncal	10.79	52.51	32.96
Linear (QaTS)	2.26	**8.37**	**9.89**
PW (K=2)	2.25	8.44	10.14
PW (K=4)	**2.24**	8.68	10.28
PW (K=10)	2.25	8.76	10.64

6 Conclusion

In this work, we show that miscalibration is highly heterogeneous across the confidence spectrum and shifts markedly across standard and more complex regimes, such as long-tailed. This variability exposes a core limitation of existing smooth temperature mappings, which are based on raw-confidence scores and lack an explicit mechanism to target the quantile regions where errors are concentrated. QaTS addresses this by operating directly in quantile space, allowing the temperature to depend on the *relative* position of a prediction within the calibration distribution. This rank-based formulation naturally aligns with a reparameterized ECE objective and remains stable across a wide range of conditions. Evaluated over diverse datasets, architectures, and tasks, this principle yields consistently improved calibration while preserving accuracy, offering a simple and broadly applicable foundation for reliable post-hoc calibration.

References

1. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: NeurIPS (2019), https://arxiv.org/abs/1906.07413
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Cheng, J., Vasconcelos, N.: Calibrating deep neural networks by pairwise constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13709–13718 (2022)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
5. Ding, Z., Han, X., Liu, P., Niethammer, M.: Local temperature scaling for probability calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6889–6899 (2021)

6. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
7. Frenkel, L., Goldberger, J.: Network calibration by class-based temperature scaling. In: 2021 29th European Signal Processing Conference (EUSIPCO). pp. 1486–1490. IEEE (2021)
8. Glenn, W.B., et al.: Verification of Forecasts Expressed in Terms of Probability. Monthly weather review **78**(1), 1–3 (1950)
9. Gong, Y., Lin, X., Yao, Y., Dietterich, T.G., Divakaran, A., Gervasio, M.: Confidence calibration for domain generalization under covariate shift. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8958–8967 (2021)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning (ICML). pp. 1321–1330. PMLR (2017)
11. Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., Hartley, R.: Calibration of neural networks using splines. In: International Conference on Learning Representations (2021)
12. Hekler, A., Brinker, T.J., Buettner, F.: Test time augmentation meets post-hoc calibration: uncertainty quantification under real-world conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 14856–14864 (2023)
13. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
14. Joy, T., Pinto, F., Lim, S.N., Torr, P.H., Dokania, P.K.: Sample-dependent adaptive temperature scaling for improved calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 14919–14926 (2023)
15. Jung, S., Seo, S., Jeong, Y., Choi, J.: Scaling of class-wise training losses for post-hoc calibration. In: International Conference on Machine Learning. pp. 15421–15434. PMLR (2023)
16. Kim, M., Kwon, J.: Uncertainty calibration with energy based instance-wise scaling in the wild dataset. In: European Conference on Computer Vision. pp. 232–248 (2024)
17. Krizhevsky, A., Nair, V., Hinton, G.: Learning Multiple Layers of Features from Tiny Images. Master's thesis, University of Toronto (2009), https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
18. Kumar, A., Sarawagi, S., Jain, U.: Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings. In: International Conference on Machine Learning. pp. 2805–2814. PMLR (2018)
19. Larrazabal, A., Martinez, C., Dolz, J., Ferrante, E.: Maximum Entropy on Erroneous Predictions (MEEP): Improving model calibration for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2023)
20. Larrazabal, A.J., Martínez, C., Dolz, J., Ferrante, E.: Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 273–283. Springer (2023)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 2980–2988 (2017)

22. Liu, B., Ben Ayed, I., Galdran, A., Dolz, J.: The devil is in the margin: Margin-based label smoothing for network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 80–88 (2022)
23. Liu, B., Rony, J., Galdran, A., Dolz, J., Ben Ayed, I.: Class adaptive network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16070–16079 (2023)
24. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-Scale Long-Tailed Recognition in an Open World. In: CVPR (2019)
25. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. Advances in neural information processing systems **34**, 15682–15694 (2021)
26. Mitchell, T.: Twenty newsgroups [dataset]. uci machine learning repository (1997)
27. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems (NeurIPS) **33**, 15288–15299 (2020)
28. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Advances in neural information processing systems (NeurIPS) **32** (2019)
29. Murugesan, B., Adiga Vasudeva, S., Liu, B., Lombaert, H., Ben Ayed, I., Dolz, J.: Trust your neighbours: Penalty-based constraints for model calibration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 572–581 (2023)
30. Murugesan, B., Silva-Rodríguez, J., Ayed, I.B., Dolz, J.: Robust calibration of large vision-language adapters. In: European Conference on Computer Vision. pp. 147–165. Springer (2024)
31. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems (NeurIPS) **32** (2019)
32. Park, H., Noh, J., Oh, Y., Baek, D., Ham, B.: Acls: Adaptive and conditional label smoothing for network calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3936–3945 (2023)
33. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: International Conference on Learning Representations (ICLR) (2017)
34. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet Large Scale Visual Recognition Challenge. International journal of computer vision **115**(3), 211–252 (2015)
36. Sun, X., Leng, X., Wang, Z., Yang, Y., Huang, Z., Zheng, L.: Cifar-10-warehouse: Broad and more realistic testbeds in model generalization analysis. In: International Conference on Learning Representations. vol. 2024, pp. 29238–29265 (2024)
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 2818–2826 (2016)
38. Tao, L., Dong, M., Xu, C.: Feature clipping for uncertainty calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 20841–20849 (2025)

39. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)
40. Tomani, C., Cremers, D., Buettner, F.: Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In: European conference on computer vision. pp. 555–569 (2022)
41. Tomani, C., Gruber, S., Erdem, M.E., Cremers, D., Buettner, F.: Post-hoc uncertainty calibration for domain drift scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10132 (2021)
42. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 180161 (2018)
43. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning (ICML). pp. 23631–23644. PMLR (2022)
44. Xiong, M., Deng, A., Koh, P.W.W., Wu, J., Li, S., Xu, J., Hooi, B.: Proximity-informed calibration for deep neural networks. Advances in Neural Information Processing Systems **36**, 68511–68538 (2023)
45. Yang, J.Q., Zhan, D.C., Gan, L.: Beyond probability partitions: Calibrating neural networks with semantic aware grouping. Advances in Neural Information Processing Systems **36**, 58448–58460 (2023)
46. Yang, J., Shi, R., Ni, B.: MedMNIST classification decathlon: A lightweight autoML benchmark for medical image analysis. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 191–195 (2021)
47. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**(1), 41 (2023)
48. Yu, Y., Bates, S., Ma, Y., Jordan, M.: Robust calibration with multi-domain temperature scaling. Advances in Neural Information Processing Systems **35**, 27510–27523 (2022)
49. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: International Conference on Machine Learning. pp. 625–632 (2001)
50. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 694–699 (2002)
51. Zhang, J., Kailkhura, B., Han, T.Y.J.: Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In: International conference on machine learning. pp. 11117–11128. PMLR (2020)
52. Zhang, L., Deng, Z., Kawaguchi, K., Zou, J.: When and how mixup improves calibration. In: International Conference on Machine Learning. pp. 26135–26160. PMLR (2022)