DIFFUSION SOFT LABELING FOR MODEL CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Overconfidence is a pervasive problem for modern classification models based on neural networks. One key source is the hard sample labels used in cross-entropy loss, which are one-hot probability vectors that encourage high confidence. Existing model calibration approaches have tried introducing softer supervision signals, which however, rely on heuristic rules and may not accurately reflect label uncertainty. In this work, we propose Diffusion Soft Labeling (DiffSL), which assigns soft labels according to a data-driven principle. Specifically, DiffSL constructs initial soft labels from diffusion classifier predictions with top-k class clipping, then corrects wrong labels using ground-truth labels. Extensive experiments show that DiffSL significantly improves calibration performance over existing methods while maintaining competitive classification accuracy.

1 INTRODUCTION

Modern DNNs are often overconfident in their predictions (Guo et al., 2017; Mukhoti et al., 2020), i.e., prediction confidence is mostly higher than actual accuracy in classification tasks, and incorrect predictions may still receive high confidence. This issue is particularly important in applications where model confidence is required to support decision fusion (Ovadia et al., 2019; Begoli et al., 2019), such as autonomous driving and medical diagnosis. Model calibration is the research field dedicated to addressing this issue.

A key source of overconfidence is the supervision signal used during model training. Standard training of classification models adopts data with one-hot labels, which assign all label probability to the ground-truth class regardless of the potential label uncertainty of individual samples (Szegedy et al., 2016; Müller et al., 2019). Moreover, the training objective based on cross-entropy loss keeps pushing the prediction confidence towards 100% even after a sample has been correctly classified (Mukhoti et al., 2020). As a result, the model may become overly confident even for difficult or ambiguous samples. These observations have motivated calibration methods that modify the supervision signal or training data during model training.

There are two typical calibration approaches under such motivation: label smoothing replaces one-hot labels with softer probability vectors by redistributing part of the probability mass from the ground-truth class to the remaining classes (Szegedy et al., 2016; Liu et al., 2022; 2023); data augmentation introduces additional fabricated samples with annotated soft labels from predefined rules (Zhang et al., 2018; Noh et al., 2023; Luo et al., 2025). While achieving better calibration performance, such methods mostly rely on heuristic labeling rules, which may not accurately capture the real uncertainty for sample labels.

In this work, we suggest estimating the label uncertainty following a data-driven principle. Specifically, we propose *Diffusion Soft Labeling (DiffSL)*, a training-time calibration framework that utilizes class-conditional diffusion models to assign soft labels for each sample. DiffSL first constructs a diffusion classifier through Bayesian rules and applies temperature scaling. Predicted logits of top-k classes for each sample are collected, from which the probability vector is calculated as soft label. Finally, soft labels with wrong classification results are corrected according to the ground-truth hard labels. Besides assigning soft labels for the original training data, we further generate inter-class augmented data and apply the same labeling procedure, except that the "ground-truth hard label" for one augmented sample is designated as a set of two reference classes.

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.

Extensive experiments show that DiffSL achieves consistently strong calibration across different datasets and network architectures. Compared with standard cross-entropy training and representative calibration methods, DiffSL obtains the lowest ECE and AECE scores in all six settings, while maintaining competitive classification accuracy across most settings. DiffSL almost completely eliminates the overconfidence issue on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets, achieving a low ECE/AECE score of 0.96%/0.87% on average, which is a 32%/40% relative reduction from the best baseline.

Our main contributions are summarized as follows:

- We introduce Diffusion Soft Labeling (DiffSL), a training-time calibration framework that assigns reliable soft labels for both original training data and augmented data using class-conditional diffusion models.
- We propose two key techniques for successful soft labeling, including *top-k class clipping* that guarantees tractable labeling efficiency, and *label correction* that ensures consistency between soft labels and ground-truth labels.
- Extensive experiments across multiple datasets and network architectures show that DiffSL consistently improves calibration while maintaining competitive classification accuracy.

2 Related Work

Post-hoc calibration. Post-hoc approaches adjust model predictions after training using a validation set. Common approaches include histogram binning (Zadrozny & Elkan, 2001), temperature scaling (Guo et al., 2017), beta calibration (Kull et al., 2017), and Dirichlet calibration (Kull et al., 2019). These approaches are simple to apply and require no model retraining, which allows them to be combined with training-time calibration approaches for further refinement of prediction confidence.

Training-time calibration. Training-time approaches improve calibration by modifying the training objective or supervision. MMCE (Kumar et al., 2018) and soft calibration objectives (Karandikar et al., 2021) directly optimize calibration-related errors, while Focal Loss and FLSD (Mukhoti et al., 2020), CRL (Moon et al., 2020), and MDCA (Hebbalaguppe et al., 2022) regulate model confidence during training. Another line of work modifies the supervision provided by one-hot labels. Label Smoothing (LS) (Szegedy et al., 2016) replaces one-hot labels with a fixed smoothed distribution. The fixed smoothing rule applies the same label adjustment to all samples and may not accurately reflect label uncertainty. MbLS (Liu et al., 2022), CALS (Liu et al., 2023), and ACLS (Park et al., 2023) further introduce adaptive regularization based on margin, class, or prediction information. These approaches are closely related to our method as they improve calibration through training-time supervision, from which we select as main baselines for experimental comparison.

Augmentation-based calibration. Data augmentation has been extensively studied for improving model calibration. Mixup constructs interpolated samples and labels (Zhang et al., 2018), while CutMix and AugMix introduce different forms of input augmentation (Yun et al., 2019; Hendrycks et al., 2020). Later work further studied the effect of Mixup on calibration and proposed calibration-oriented variants (Zhang et al., 2022; Pinto et al., 2022). However, the mixing ratio used for label interpolation may not accurately represent the actual distribution of augmented samples. RankMixup addresses this issue by replacing mixed-label supervision with confidence-ranking constraints (Noh et al., 2023). SKMixup further considers the manifold mismatch caused by linear interpolation, where synthetic samples may fall outside the original class manifolds (Bouniot et al., 2025). CSM uses diffusion-based semantic mixing and calibrated re-annotation to alleviate the mismatch between the mixing ratio and the confidence of generated samples (Luo et al., 2025). The labels of augmented samples are often assigned through heuristic rules, which may also benefit from our soft labeling strategy.

3 Preliminaries

Model Calibration. Model calibration aims to make prediction confidence equal to the accuracy of model predictions for any data subset. Consider an image classification problem with a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ is an image and $y_i \in \{1, \dots, C\}$ is its ground-truth label. A classifier $f_\theta : \mathcal{X} \to \mathbb{R}^C$ outputs class logits, and the predictive distribution is obtained as $\mathbf{p}_\theta(\mathbf{x}_i) =$

$\mathrm{softmax}(f_\theta(\mathbf{x}_i)) = (p_\theta(y=1|\mathbf{x}_i), \cdots, p_\theta(y=C|\mathbf{x}_i))$, where $p_\theta(y|\mathbf{x}_i)$ denotes the predicted probability for class y. The predicted class and its confidence are

$$\hat{y}_i = \arg\max_y p_\theta(y|\mathbf{x}_i), \qquad \hat{p}_i = \max_y p_\theta(y|\mathbf{x}_i). \tag{1}$$

A classifier is well-calibrated if its predicted confidence matches the probability of making a correct prediction (Guo et al., 2017), i.e.,

$$\Pr(\hat{y}_i = y_i | \hat{p}_i = \alpha) = \alpha, \qquad \forall \alpha \in [0,1]. \tag{2}$$

Expected Calibration Error (ECE) measures the discrepancy between prediction confidence and accuracy. In practice, given a finite set of samples, the confidence interval $[0,1]$ is divided into L equally spaced bins $\{B_l\}_{l=1}^L$, and ECE is approximated as

$$\mathrm{ECE} = \sum_{l=1}^{L} \frac{|B_l|}{N} \left|\mathrm{acc}(B_l) - \mathrm{conf}(B_l)\right|, \tag{3}$$

where $\mathrm{acc}(B_l)$ and $\mathrm{conf}(B_l)$ denote the accuracy and average confidence of samples in the l-th bin. An alternative version is Adaptive Expected Calibration Error (AECE) (Ding et al., 2020), which uses adaptive bins containing approximately equal numbers of samples. Lower ECE and AECE indicate better calibration.

Diffusion Models and Diffusion Classifiers. Diffusion models generate samples by progressively denoising Gaussian noise. In a class-conditional diffusion model, the denoising process is conditioned on the class label y. Following the EDM formulation (Karras et al., 2022), a noisy sample at noise level t can be written as $\mathbf{x}_t = \mathbf{x} + t\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The conditional denoiser $D_\phi(\mathbf{x}_t, t, y)$ is parameterized by ϕ and predicts the denoised sample conditioned on class y at noise level t. Unconditional generation can also be achieved through condition dropout by setting $D_\phi(\mathbf{x}_t, t) = D_\phi(\mathbf{x}_t, t, y = \emptyset)$ during both training and inference.

A Diffusion Classifier (DC) uses a class-conditional diffusion model for classification by comparing the diffusion objectives across different classes (Li et al., 2023; Chen et al., 2024). Given an input $\mathbf{x}$, the posterior probability of class y follows Bayes' rule and is given by

$$p_\phi(y|\mathbf{x}) \propto p_\phi(\mathbf{x}|y)p(y). \tag{4}$$

where $p_\phi(\mathbf{x}|y)$ denotes the likelihood of $\mathbf{x}$ conditioned on class y. Assuming a uniform class prior, classification reduces to comparing the class-conditional likelihoods. Since $p_\phi(\mathbf{x}|y)$ is intractable to compute directly, the class-conditional log-likelihood can be approximated using the evidence lower bound (ELBO) of the diffusion model (Li et al., 2023) as

$$\log p_\phi(\mathbf{x}|y) \geq \mathcal{L}_{\mathrm{ELBO}}(\mathbf{x}, y) = -\mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w_t \left\|D_\phi(\mathbf{x}_t, t, y) - \mathbf{x}\right\|_2^2\right] + \mathrm{const.}, \tag{5}$$

where w_t denotes the weighting term at noise level t. Based on this approximation, the class-wise score is defined by

$$d(\mathbf{x}, y) = -\mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w_t \left\|D_\phi(\mathbf{x}_t, t, y) - \mathbf{x}\right\|_2^2\right]. \tag{6}$$

A larger $d(\mathbf{x}, y)$ corresponds to a larger variational lower bound on the class-conditional log-likelihood. The corresponding DC prediction is

$$\hat{y}^{\mathrm{DC}} = \arg\max_y d(\mathbf{x}, y). \tag{7}$$

One drawback of diffusion classifiers is the computation efficiency. The number of neural function evaluations for one prediction is $\mathrm{NFE} = N_\epsilon \cdot C$ instead of normally $\mathrm{NFE} = 1$, where N_ϵ is the number of sampled noise ϵ and C is the number of classes.

4 METHOD

We first provide a rationality analysis for soft labeling and an overview of our *Diffusion Soft Labeling (DiffSL)* algorithm, then introduce the details of our soft labeling and data augmentation method.

4.1 RATIONALITY OF SOFT LABELING.

Soft labels can be theoretically equivalent to and empirically better than hard labels, if properly assigned. Consider the case with infinite training data i.i.d. sampled from joint distribution $p(\mathbf{x}, y)$; the cross-entropy loss essentially converges to the following expectation according to the law of large numbers:

$$\mathcal{L}_{\mathrm{CE}} = -\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(y_i|\mathbf{x}_i) = -\mathop{\mathbb{E}}_{\mathbf{x},y\sim p(\mathbf{x},y)}\log p_\theta(y|\mathbf{x}) = -\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})}\mathop{\mathbb{E}}_{y\sim p(y|\mathbf{x})}\log p_\theta(y|\mathbf{x}), \tag{8}$$

where $p_\theta(y|\mathbf{x})$ is the conditional distribution induced by classifier f_θ. This is equivalent to

$$\mathcal{L}_{\mathrm{CE}} = -\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})}\sum_{y\in\{1,\cdots,C\}} p(y|\mathbf{x})\cdot\log p_\theta(y|\mathbf{x}) = -\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})}\mathbf{p}^{\mathrm{T}}(\mathbf{x})\cdot\log\mathbf{p}_\theta(\mathbf{x}), \tag{9}$$

where $\mathbf{p}^{\mathrm{T}}(\mathbf{x}) = \left(p(y=1|\mathbf{x}),\cdots,p(y=C|\mathbf{x})\right)^{\mathrm{T}}$ can be viewed as soft label for sample $\mathbf{x}$. This means that replacing hard labels with soft probability vectors results in an equivalent optimization objective.

However, the optimal classifiers from the above two paradigms become different under finite empirical data where each $\mathbf{x}$ never appears twice. With hard labels that each $\mathbf{x}$ is assigned one certain label $y = y_{\mathbf{x}}$, Eq. (8) yields $p^*_\theta(y|\mathbf{x}) = \mathbb{I}(y = y_{\mathbf{x}})$. With estimated soft labels $\hat{p}(y|\mathbf{x})$, Eq. (9) yields $p^*_\theta(y|\mathbf{x}) = \hat{p}(y|\mathbf{x})$. Clearly, classifiers trained by hard labels are prone to be overconfident. On the opposite, classifiers trained by soft labels can be better calibrated, as long as the assigned soft labels are close to ground-truth probabilities $p(y|\mathbf{x})$. As a result, we suggest estimating $p(y|\mathbf{x})$ as soft labels through generative models.

4.2 OVERVIEW OF DIFFSL ALGORITHM

The main idea of DiffSL is to estimate $p(y|\mathbf{x})$ using diffusion classifiers. Unlike discriminative classifiers that favor overconfidence under cross-entropy loss, diffusion classifiers are generative classifiers that model the conditional distribution $p(\mathbf{x}|y)$, which does not explicitly encourage overconfidence. Therefore, it can be beneficial for a discriminative classifier to be trained using the probability vectors from a diffusion classifier as soft labels.

However, directly using the above preliminary soft labels may be problematic. First, creating soft labels using diffusion classifiers may be time-consuming when the number of classes C is large, though soft labeling is a one-off procedure during model development. Second, preliminary soft labels may conflict with ground-truth labels, which may harm the overall accuracy of the final classifier.

We solve the above problems by proposing the following two techniques. For the efficiency problem, DiffSL only calculates the top-k possible class logits instead of all C logits, significantly reducing the labeling time while maintaining good calibration ability. In practice, setting $k = 15$ is enough for CIFAR-100 ($C = 100$) and Tiny-ImageNet ($C = 200$), gaining an 85%-92% NFE reduction. For the label conflict problem, if one soft label makes a wrong prediction, DiffSL forces all probabilities for wrong classes to be 0, and re-normalizes the probabilities among all correct classes (an augmented sample may have multiple correct classes).

The overall procedure of DiffSL is shown in Fig. 1. DiffSL utilizes a pretrained reference classifier g for top-k class suggestions, then applies a diffusion classifier to get the probability vector $\mathbf{q}^{\mathrm{DC}}$, after which label correction is performed to get the final soft label $\mathbf{q}$. DiffSL also introduces inter-class augmented data $\mathcal{D}_{\mathrm{aug}}$, which are generated by interpolating two denoising directions given two random reference classes during diffusion sampling. A subset $\widetilde{\mathcal{D}} \subseteq \mathcal{D}_{\mathrm{aug}}$ is selected for final training, whose soft labels $\tilde{\mathbf{q}}$ are acquired through the same procedure.

DiffSL uses KL divergence loss for classifier training with soft labels, which is equivalent to cross-entropy loss in terms of optimization. The overall training objective is the combination of losses on original and augmented data, which is $\mathcal{L} = \mathcal{L}_{\mathrm{orig}} + \mathcal{L}_{\mathrm{aug}}$ where

$$\mathcal{L}_{\mathrm{orig}} = \frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}\left(\mathbf{q}_i \parallel \mathbf{p}_\theta(\mathbf{x}_i)\right), \qquad \mathcal{L}_{\mathrm{aug}} = \frac{1}{|\widetilde{\mathcal{D}}|}\sum_{j=1}^{|\widetilde{\mathcal{D}}|} D_{\mathrm{KL}}\left(\tilde{\mathbf{q}}_j \parallel \mathbf{p}_\theta(\tilde{\mathbf{x}}_j)\right). \tag{10}$$

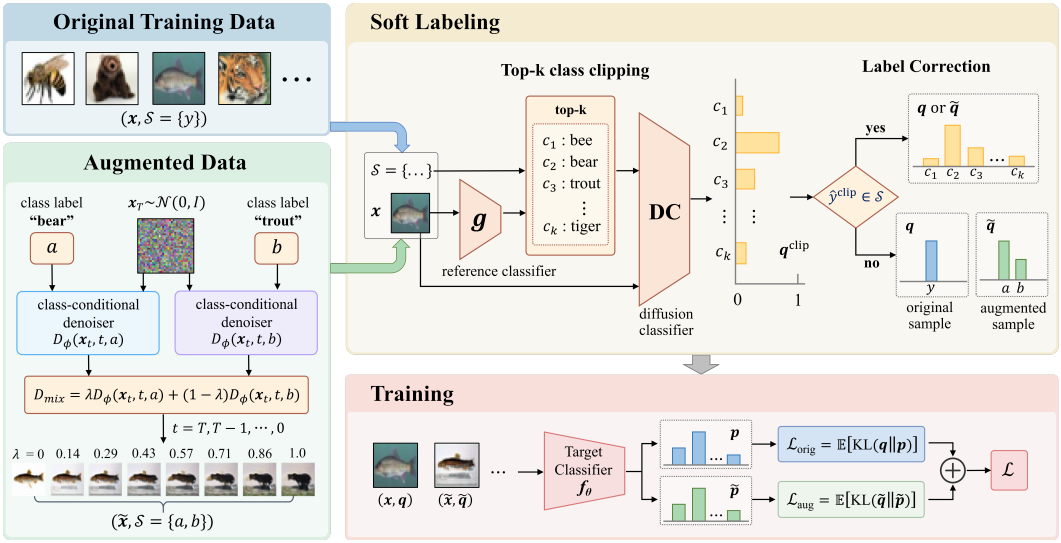


Figure 1: Overview of our DiffSL algorithm. Example images are picked from CIFAR-100 dataset.

4.3 Soft Labeling

Calibrated diffusion classifier. To obtain soft labels of best quality, DiffSL first calibrates the probability distribution of the diffusion classifier (DC) using temperature scaling. Given a sample $\mathbf{x}$, the DC scores are converted into probabilities as

$$q_y^{\text{DC}}(\mathbf{x}; \tau) = \frac{\exp(d(\mathbf{x}, y)/\tau)}{\sum_c \exp(d(\mathbf{x}, c)/\tau)}, \tag{11}$$

where τ denotes the temperature parameter. For each dataset, DiffSL calculate the full-class DC scores on the training data and select τ that achieves the lowest ECE. The selected τ is fixed for all samples in the same dataset. Detailed settings are provided in Appendix A.

Top-k class clipping. Calculating the DC scores for all C classes leads to high computational cost when C is large. To improve efficiency, DiffSL only calculates the DC scores for the top-k candidate classes. DiffSL utilizes a pretrained reference classifier g for top-k class clipping. The reference classifier has the same architecture as the target classifier and is trained on the training data using standard cross-entropy. It is kept fixed during soft labeling.

Given a sample $\mathbf{x}$, we first assign a "ground-truth" class set $\mathcal{S}$ for it. $\mathcal{S}$ contains only one element y for samples from original training dataset, but may contain more elements for augmented sample. Then the top-k candidate set $\mathcal{S}^k$ with $|\mathcal{S}| \le k \le C$ is constructed according to the following rule: first initialize $\mathcal{S}^k$ with $\mathcal{S}$, then select $k - |\mathcal{S}|$ classes from $\{1, \cdots, C\} \backslash \mathcal{S}$ with highest confidence according to reference classifier prediction $g(\mathbf{x})$. DiffSL applies diffusion model evaluation only to classes in $\mathcal{S}^k$, and the preliminary soft label is defined as

$$q_y^{\text{clip}}(\mathbf{x}) = \begin{cases} \frac{\exp(d(\mathbf{x}, y)/\tau)}{\sum_{c \in \mathcal{S}^k} \exp(d(\mathbf{x}, c)/\tau)}, & y \in \mathcal{S}^k, \\ 0, & y \notin \mathcal{S}^k. \end{cases} \tag{12}$$

The resulting probability vector is denoted as $\mathbf{q}^{\text{clip}}(\mathbf{x}) = (q_1^{\text{clip}}(\mathbf{x}), \cdots, q_C^{\text{clip}}(\mathbf{x}))$.

Label correction. The preliminary soft label may conflict with the ground-truth label. The predicted class given $\mathbf{q}^{\text{clip}}(\mathbf{x})$ is

$$\hat{y}^{\text{clip}} = \arg\max_y q_y^{\text{clip}}(\mathbf{x}). \tag{13}$$

If $\hat{y}^{\text{clip}} \in \mathcal{S}$, DiffSL keeps $\mathbf{q}^{\text{clip}}$ unchanged. Otherwise, DiffSL sets the probabilities for all classes not in $\mathcal{S}$ to 0, and re-normalize the probability vector. Specifically, there is

$$q_y(\mathbf{x}) = \begin{cases} q_y^{\text{clip}}(\mathbf{x}), & \hat{y}^{\text{clip}} \in \mathcal{S}, \\ q_y^{\text{clip}}(\mathbf{x}) / \sum_{c \in \mathcal{S}} q_c^{\text{clip}}(\mathbf{x}), & \hat{y}^{\text{clip}} \notin \mathcal{S} \text{ and } y \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The final soft label is denoted as $\mathbf{q}(\mathbf{x}) = (q_1(\mathbf{x}), \cdots, q_C(\mathbf{x}))$.

4.4 Inter-Class Augmentation

DiffSL further introduces inter-class augmentation to generate additional training samples with class uncertainty. The augmented samples are generated by interpolating two class-conditional denoising directions during diffusion sampling, forming transitions between the two generating classes.

Let Λ denote the sequence of mixing coefficients, which are linearly spaced points on $[0, 1]$. Given a class pair (a, b) and $\lambda \in \Lambda$, the denoising predictions conditioned on the two classes are combined as

$$D_{\text{mix}}(\mathbf{x}_t, t; a, b, \lambda) = \lambda D_\phi(\mathbf{x}_t, t, a) + (1 - \lambda) D_\phi(\mathbf{x}_t, t, b), \tag{15}$$

where $D_\phi(\mathbf{x}_t, t, y)$ denotes the denoiser conditioned on class y. To generate one augmented sample $\tilde{\mathbf{x}}_\lambda$, DiffSL follows standard diffusion generation except replacing original condition dropout denoiser $D_\phi(\mathbf{x}_t, t)$ with $D_{\text{mix}}(\mathbf{x}_t, t; a, b, \lambda)$. The class pair (a, b) are randomly selected distinct classes from all C classes. In addition, by using a shared initial noise $\mathbf{x}_T$ and varying λ alone, DiffSL gets a sequence of augmented samples together. We create M sequences of augmented samples to form a whole candidate pool, where M may vary across datasets.

As for soft labeling with augmented samples, the procedure is the almost the same as samples in original dataset, expect that the "ground-truth" class set $\mathcal{S}$ has a different assignment. For a class pair (a, b), $\mathcal{S}$ would be $\{a\}$ or $\{b\}$ if $\lambda = 1$ or $\lambda = 0$, respectively. When $\lambda \in (0, 1)$, we set $\mathcal{S} = \{a, b\}$ to capture all possibilities.

We denote the whole generated dataset as $\mathcal{D}_{\text{aug}}$, but only a subset $\widetilde{\mathcal{D}} \subseteq \mathcal{D}_{\text{aug}}$ is selected for final training. Selection strategy may vary across datasets, we suggest to determine the specific strategy through ECE validation on validation datasets. In this paper, we simply pick N_{aug} samples from each sample sequence to form $\widetilde{\mathcal{D}}$. The whole strategy pool for validation is set to $N_{\text{aug}} \in \{0, 1, 2\}$. We leave elaborated selection strategies as future work.

5 Experiments

5.1 Experimental Setup

We conduct experiments using ResNet-50 and ResNet-101 (He et al., 2016) on CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and Tiny-ImageNet (Le & Yang, 2015). Calibration performance is evaluated using Expected Calibration Error (ECE) (Guo et al., 2017) and Adaptive Expected Calibration Error (AECE) (Ding et al., 2020), together with top-1 accuracy (ACC). Both ECE and AECE are computed with 15 bins. ACC, ECE, and AECE are reported in percentage points.

Baseline methods. We compare with cross-entropy (CE) training and representative training-time calibration methods, including MMCE (Kumar et al., 2018), Focal Loss (FL) and FLSD (Mukhoti et al., 2020), Label Smoothing (LS) (Szegedy et al., 2016), MbLS (Liu et al., 2022), and CALS (Liu et al., 2023). We further compare with augmentation-based calibration methods, including Mixup (Zhang et al., 2018), RankMixup (Noh et al., 2023), SKMixup (Bouniot et al., 2025), and CSM (Luo et al., 2025). We use the official implementations and reported hyperparameter settings when available. For each baseline, we additionally apply post-hoc temperature scaling following Guo et al. (2017), where the temperature is optimized on the validation set and fixed for test evaluation.

Training details. Detailed training protocols and hyperparameter settings for DiffSL and all baselines are provided in Appendix A. All main experiments are repeated over 10 random seeds.

Table 1: Calibration performance on CIFAR-10, CIFAR-100, and Tiny-ImageNet. R50: ResNet-50; R101: ResNet-101. Values in parentheses represent results after temperature scaling.

	CIFAR-10						CIFAR-100						Tiny-ImageNet					
	R50			R101			R50			R101			R50			R101		
Method	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓
CE	95.24	3.09(0.91)	3.06(1.14)	95.20	3.31(1.03)	3.27(1.29)	78.87	11.81(2.88)	11.80(2.89)	78.82	13.14(3.84)	13.13(3.77)	65.00	3.57(1.61)	3.53(1.50)	65.86	4.27(1.74)	4.24(1.67)
MMCE	95.24	3.08(0.92)	3.06(1.13)	95.18	3.42(1.03)	3.41(1.47)	78.61	12.08(2.86)	12.06(2.93)	78.77	13.39(4.18)	13.38(4.13)	65.21	3.94(1.74)	3.86(1.74)	65.81	5.33(2.00)	5.32(1.90)
LS	95.18	3.50(1.26)	4.16(3.20)	95.14	3.37(1.25)	4.25(3.34)	79.11	5.44(4.91)	5.73(4.98)	78.67	8.62(5.80)	9.12(7.64)	65.19	2.49(2.46)	2.52(2.49)	66.01	1.99(1.99)	1.94(1.94)
Mixup	95.27	1.63(1.45)	2.28(2.21)	95.46	2.50(1.47)	2.83(2.31)	79.02	7.35(1.84)	7.33(1.86)	79.79	8.02(2.95)	8.01(2.94)	65.85	2.46(2.17)	2.44(2.11)	66.57	2.00(1.91)	2.07(1.90)
FL	94.92	4.36(1.13)	4.23(1.36)	94.88	3.97(1.24)	3.82(1.41)	78.74	4.71(2.26)	4.70(2.21)	78.96	4.60(3.10)	4.46(3.17)	62.55	5.81(2.09)	5.79(2.10)	63.47	5.03(1.68)	5.00(1.67)
FLSD	94.93	4.37(1.04)	4.29(1.27)	94.91	3.82(1.24)	3.71(1.43)	78.28	4.54(2.19)	4.53(2.12)	78.70	4.32(3.12)	4.23(3.21)	62.46	5.99(2.13)	5.98(2.08)	63.43	5.35(1.87)	5.33(1.86)
MbLS	95.11	1.27(1.27)	3.25(3.25)	95.17	1.32(1.32)	3.32(3.32)	79.02	5.28(5.28)	5.75(5.75)	78.74	6.00(6.00)	8.94(8.94)	64.91	1.98(2.15)	2.03(2.16)	65.83	1.48(1.48)	1.57(1.57)
CALS	95.19	3.41(1.47)	3.38(2.39)	95.19	3.52(1.29)	3.50(2.56)	79.20	11.41(3.86)	11.38(4.14)	78.87	13.23(4.65)	13.22(5.25)	64.90	1.45(1.45)	1.45(1.45)	65.67	1.84(1.84)	1.70(1.70)
RankMixup	94.35	2.93(0.66)	2.90(0.51)	94.28	2.94(0.77)	2.89(0.75)	77.28	10.95(2.13)	10.93(2.14)	77.35	11.45(1.81)	11.44(1.89)	60.41	1.80(1.80)	1.67(1.67)	61.28	2.08(2.04)	1.98(1.97)
SKMixup	95.20	2.65(0.69)	2.58(0.80)	95.36	2.56(0.67)	2.49(0.84)	78.09	3.25(1.80)	3.13(2.00)	79.24	2.96(2.43)	2.78(2.47)	61.41	9.58(1.36)	9.58(1.28)	63.08	9.23(1.42)	9.22(1.40)
CSM	**95.80**	0.57(0.60)	0.39(0.43)	**95.89**	0.54(0.63)	0.44(0.51)	79.39	1.93(1.94)	2.01(2.00)	79.37	2.57(2.25)	2.65(2.31)	65.44	3.41(1.87)	3.33(1.79)	**66.33**	4.09(1.72)	4.04(1.66)
Ours	95.17	**0.44**	**0.31**	95.19	**0.48**	**0.35**	**80.55**	**1.05**	**0.94**	**80.93**	**1.23**	**1.14**	**68.06**	**1.25**	**1.26**	64.78	**1.28**	**1.22**

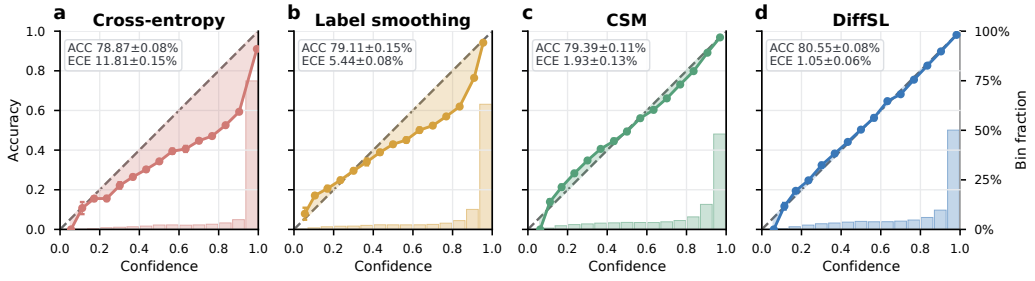


Figure 2: Reliability diagrams on CIFAR-100 with ResNet-50 for CE, Label Smoothing, CSM, and DiffSL. The bars show the fraction of samples in each confidence bin, corresponding to the right y-axis. The dashed diagonal denotes perfect calibration. ACC and ECE are reported as mean ± standard error of the mean (SEM).

For each target architecture, the reference classifier g uses the same backbone and is separately trained with the standard CE objective. The top-k size is set to $k = 10$ on CIFAR-10 and $k = 15$ on CIFAR-100 and Tiny-ImageNet. Since CIFAR-10 contains 10 classes, all classes are retained during soft-label construction. For each dataset and backbone, the number of augmented samples N_{aug} is selected from $\{0, 1, 2\}$ according to validation ECE and fixed before test evaluation. $N_{\text{aug}} = 0$ indicates that no augmented sample is used. For $N_{\text{aug}} = 1$, one edge sample is selected from each sample sequence. For $N_{\text{aug}} = 2$, one intermediate sample is additionally selected from the same sequence. The selected N_{aug} values for ResNet-50/ResNet-101 are $0/0$ on CIFAR-10, $2/2$ on CIFAR-100, and $1/0$ on Tiny-ImageNet. We use the pretrained class-conditional EDM checkpoints released with RDC (Chen et al., 2024), based on the EDM formulation of Karras et al. (2022), to construct soft labels and generate inter-class augmented data. Detailed temperature, generation, and sampling settings are provided in Appendix A.

5.2 Main Results

Table 1 compares DiffSL with representative regularization-based and augmentation-based calibration methods. Three main observations can be made from the results.

(1) DiffSL achieves the lowest ECE and AECE in all six dataset–backbone settings. It reaches an average ECE/AECE of 0.96%/0.87%, corresponding to a 31.5%/40.0% relative reduction from the best baselines (SKMixup for ECE and CSM for AECE). The improvement is particularly clear on CIFAR-100. A reliability diagram is given in Figure 2, showing that the confidence-accuracy curve of DiffSL is near-optimal (close to diagonal), which means DiffSL achieves almost perfect calibration.

(2) DiffSL improves calibration while maintaining competitive classification accuracy. It achieves the highest accuracy in three of the six settings, which are even higher than the CE baseline, we attribute such result to the usage of augmented data in these settings. On average, DiffSL achieves

Table 2: Effect of the soft-label source on CIFAR-10 with ResNet-50.

Soft-label source	ACC↑	ECE↓	AECE↓
Pretrained classifier	**95.30±0.04**	2.94±0.06	2.93±0.05
Conditional DDPM	94.67±0.11	0.70±0.04	0.63±0.10
Conditional EDM (DiffSL)	95.13±0.08	**0.42±0.05**	**0.36±0.07**

Table 3: Effect of label correction across datasets with ResNet-50.

	CIFAR-10			CIFAR-100			Tiny-ImageNet		
Correction	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓
w/o	94.46±0.05	0.49±0.08	**0.29±0.02**	80.28±0.10	1.95±0.12	1.99±0.11	64.64±0.03	2.99±0.09	2.94±0.11
w/	**95.13±0.08**	**0.42±0.05**	0.36±0.07	**80.61±0.16**	**1.08±0.17**	**1.11±0.27**	**67.93±0.03**	**1.12±0.09**	**1.11±0.12**

Table 4: Effect of soft labeling and inter-class augmentation on CIFAR-100 with ResNet-50.

Original label	Augmentation	ACC↑	ECE↓	AECE↓
One-hot	w/o	78.93±0.22	11.93±0.23	11.92±0.23
One-hot	w/	**80.88±0.05**	2.72±0.10	2.59±0.11
Soft label	w/o	78.95±0.14	5.16±0.27	5.12±0.25
Soft label	w/	80.61±0.16	**1.08±0.17**	**1.11±0.27**

the best accuracy, which is 0.51% higher than the second baseline CSM. Overall, DiffSL maintains competitive predictive performance while achieving lower calibration errors.

Remark on post-hoc temperature scaling. Note that our DiffSL does not require post-hoc temperature scaling. In fact, enforcing temperature scaling would lead to ECE/AECE increase in some settings, which is detrimental to DiffSL on average. This is reasonable since DiffSL already achieves near-optimal calibration, applying temperature scaling according to a validation set provides limited benefit and may accidentally cause temperature-overfitting to validation set, especially when validation set is not large enough. Similar phenomena can be observed on other baseline methods, e.g., CSM achieves higher ECE/AECE on CIFAR-10 after temperature scaling.

5.3 Analytical Results

We analyze the effectiveness of main components in DiffSL, including the choice of soft label source, label correction strategy, combination strategy of soft labeling and inter-class augmentation selection of the number of augmented samples, and selection of the top-k candidate-set size. Results in Tables 2–4 are reported as mean ± standard error of the mean (SEM) over 3 seeds (so that the values may have slight difference to the main table with 10 seeds).

Soft label source. We evaluate different sources for constructing soft labels on CIFAR-10 dataset, including a conditional DDPM and a pretrained discriminative classifier. All settings are kept identical except for the source of soft label, so that no augmented data is used in this setting. As shown in Table 2, Soft labels from the pretrained discriminative classifier provides almost no benefit for calibration, achieving nearly the same ACC/ECE/AECE compared with CE baseline. Soft labels from conditional DDPM achieves slightly worse calibration than DiffSL, but is still much better than a pretrained classifier. These results show that soft labels from stronger generative models also behave better in terms of calibration performance.

Label correction. We evaluate label correction by removing the correction step and directly using the preliminary soft labels as training targets. As shown in Table 3, correction consistently improves classification accuracy across all three datasets, and improves calibration on CIFAR-100 and Tiny-ImageNet, while reducing ECE on CIFAR-10 with a slight increase in AECE. These results support correcting preliminary soft labels according to the ground-truth labels when their top-1 predictions are incorrect.

Table 5: Validation and test ECE (%) for different values of N_{aug}.

Dataset	Backbone	Validation ECE↓			Test ECE↓		
		$N_{\text{aug}}=0$	$N_{\text{aug}}=1$	$N_{\text{aug}}=2$	$N_{\text{aug}}=0$	$N_{\text{aug}}=1$	$N_{\text{aug}}=2$
CIFAR-10	ResNet-50	**0.75±0.07**	0.76±0.03	1.08±0.08	**0.44±0.04**	0.67±0.03	0.86±0.06
	ResNet-101	**0.62±0.06**	0.87±0.06	1.09±0.06	**0.48±0.03**	0.81±0.05	0.89±0.06
CIFAR-100	ResNet-50	4.78±0.20	3.00±0.10	**2.11±0.09**	5.13±0.15	4.51±0.10	**1.05±0.06**
	ResNet-101	5.40±0.09	3.81±0.11	**1.55±0.09**	6.13±0.11	5.52±0.08	**1.23±0.06**
Tiny-ImageNet	ResNet-50	1.61±0.13	**1.57±0.10**	1.82±0.11	1.43±0.08	**1.25±0.08**	1.87±0.13
	ResNet-101	**1.27±0.11**	1.66±0.17	2.40±0.14	**1.28±0.07**	1.61±0.10	2.50±0.13

Table 6: Test performance (%) for different Top-k sizes on CIFAR-100 and Tiny-ImageNet.

k	CIFAR-100			Tiny-ImageNet		
	ACC↑	ECE↓	AECE↓	ACC↑	ECE↓	AECE↓
5	80.52±0.06	2.69±0.11	2.67±0.11	68.14±0.05	2.39±0.11	2.29±0.10
10	**80.66±0.06**	1.26±0.03	1.28±0.04	**68.16±0.09**	1.65±0.06	1.49±0.07
15	80.55±0.08	**1.05±0.06**	**0.94±0.10**	68.06±0.13	**1.25±0.08**	**1.26±0.10**
20	80.38±0.09	1.07±0.07	1.03±0.09	68.03±0.09	1.74±0.11	1.65±0.10

Soft labeling and inter-class augmentation. We study the effects of soft labeling for the original training data and inter-class augmentation on CIFAR-100. Table 4 compares one-hot labels and corrected soft labels for the original data, with and without inter-class augmentation. Replacing one-hot labels with soft labels substantially improves calibration even without augmentation, showing the effectiveness of soft labeling for the original training data. Inter-class augmentation further improves calibration and classification accuracy.

Number of augmented samples. We further analyze the configuration of augmented examples, with $N_{\text{aug}} \in \{0, 1, 2\}$. Table 5 shows that the effect of inter-class augmentation varies across datasets and architectures, and validation and test ECE exhibit consistent trends. On CIFAR-10, $N_{\text{aug}} = 0$ performs best for both backbones, adding augmented data leads to slightly worse ECE. CIFAR-100 shows the opposite trend, where $N_{\text{aug}} = 2$ performs the best. Tiny-ImageNet reports different tendencies on two architectures, while ResNet-50 prefer $N_{\text{aug}} = 1$, ResNet-101 prefer $N_{\text{aug}} = 0$. The reason for such divergence over N_{aug} preference may be attributed to the nuanced difference on confidence level of non-augmented classifier. If $N_{\text{aug}} = 0$ already brings a well-calibrated or under-confident classifier, introducing more augmented data with low-confidence soft labels would not be beneficial. Anyway, in all six settings, the value selected by validation ECE also achieves the lowest test ECE, supporting validation-based selection rather than a fixed N_{aug}.

Top-k Size. We analyze the candidate-set size k on CIFAR-100 and Tiny-ImageNet. A larger k includes more competing classes when constructing soft labels from DC predictions, but also increases the number of class-conditional DC evaluations and the corresponding computational cost. As shown in Table 6, increasing k from 5 to 15 substantially improves calibration with only minor accuracy changes, whereas increasing it to 20 provides no further benefit. Thus, $k = 15$ provides a favorable balance between calibration performance and DC evaluation cost.

6 Conclusion

In this work, we introduced Diffusion Soft Labeling (DiffSL) to improve model calibration by assigning soft labels according to a data-driven principle. DiffSL constructs soft labels from diffusion classifier predictions with top-k class clipping and corrects wrong labels using ground-truth labels. The same labeling method is further applied to inter-class augmented data. Extensive experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet show that DiffSL consistently improves calibration performance while maintaining competitive classification accuracy.

AI USE STATEMENT

In this work, generative AI tools were used to assist with parts of the code implementation and for language polishing of the manuscript. All AI-assisted code and edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1:20–23, 2019. doi: 10.1038/s42256-018-0004-1.

Quentin Bouniot, Pavlo Mozharovskyi, and Florence d'Alché-Buc. Tailoring mixup to data for calibration. In *International Conference on Learning Representations*, 2025.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. doi: 10.1109/CVPRW50498.2020.00010.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017.

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

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

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

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*. PMLR, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. CS231n Course Project Report, Stanford University, 2015.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Bingyuan Liu, Jérôme Rony, Adrian Galdran, Jose Dolz, and Ismail Ben Ayed. Class adaptive network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Haoyang Luo, Linwei Tao, Minjing Dong, and Chang Xu. Beyond one-hot labels: Semantic mixing for model calibration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jongyoun Noh, Hyekang Park, Junghyup Lee, and Bumsub Ham. Rankmixup: Ranking-based mixup training for network calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hyekang Park, Jongyoun Noh, Youngmin Oh, Donghyeon Baek, and Bumsub Ham. ACLS: Adaptive and conditional label smoothing for network calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip H. S. Torr, and Puneet K. Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, and James Zou. When and how mixup improves calibration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

A Experiment Details

A.1 Datasets

We evaluate DiffSL on CIFAR-10, CIFAR-100, and Tiny-ImageNet. CIFAR-10 and CIFAR-100 contain 32×32 color images from 10 and 100 classes, respectively. For both datasets, we use 50,000 images for training and reserve 5,000 images each for validation and testing. Tiny-ImageNet contains 64×64 images from 200 classes. We use 100,000 images for training, with 10,000 images used for validation and another 10,000 for testing. The validation set is used for model selection, the selection of N_{aug}, and post-hoc temperature scaling.

A.2 Compared Baselines

We compare DiffSL with representative training-time calibration methods. For regularization-based methods, we use a) MMCE (Kumar et al., 2018) with $\lambda = 1.0$, b) Focal Loss (Mukhoti et al., 2020) with $\gamma = 3$, c) FLSD (Mukhoti et al., 2020) with the FLSD-53 schedule, d) Label Smoothing (Szegedy et al., 2016) with $\alpha = 0.05$, e) MbLS (Liu et al., 2022) with $m = 6$ on CIFAR-10/100 and $m = 10$ on Tiny-ImageNet, and f) CALS (Liu et al., 2023) with $m = 10$ on all three datasets.

We also compare with augmentation-based methods, including a) Mixup (Zhang et al., 2018) with a Beta distribution parameter $\alpha = 0.2$; b) RankMixup (Noh et al., 2023) using the M-NDCG variant with loss weight $w = 0.1$, ranking-list size $Q = 4$, and Beta distribution parameter $\alpha = 2$ on CIFAR-10/100 and $\alpha = 1$ on Tiny-ImageNet; c) SKMixup (Bouniot et al., 2025) with $(\tau_{\max}, \tau_{\mathrm{std}}) = (1.0, 0.25)$ on CIFAR-10/100 and Tiny-ImageNet; and d) CSM (Luo et al., 2025) with scaling parameter $s = 4.0$ on CIFAR-10 and Tiny-ImageNet and $s = 2.3$ on CIFAR-100.

A.3 Training Setup

The training and evaluation setup follows MbLS (Liu et al., 2022). DiffSL is trained for 350 epochs on CIFAR-10/100 and for 100 epochs on Tiny-ImageNet using SGD with a multi-step learning-rate schedule. The initial learning rate is 0.1, with momentum 0.9 and weight decay 5×10^{-4}. The batch size is 128 on CIFAR-10/100 and 64 on Tiny-ImageNet. The learning rate is reduced by a factor of 0.1 at epochs 150 and 250 on CIFAR-10/100 and at epochs 40 and 60 on Tiny-ImageNet. Baselines follow their corresponding training protocols and reported hyperparameter settings. All main experiments are repeated over 10 random seeds.

A.4 DiffSL Implementation Details

Soft-label temperature. For each dataset, we compute the full-class DC scores on the original training set and select a single temperature τ by minimizing the 15-bin ECE of the corresponding DC predictions. The selected temperature is fixed for the dataset and shared across both network architectures. During soft-label construction, τ is applied to the DC scores over the top-k classes defined in Section 4, and the same value is used for the augmented samples. We set $\tau = 0.125$ on CIFAR-10, $\tau = 1.37$ on CIFAR-100, and $\tau = 2.29$ on Tiny-ImageNet.

Inter-class sample generation. For each class pair (a, b), we use multiple random seeds to generate different augmented-sample sequences. Each random seed determines a shared initial Gaussian noise $\mathbf{x}_T$, from which 8 augmented samples are generated using uniformly spaced mixing coefficients $\Lambda = \{0, 1/7, 2/7, \ldots, 1\}$. The samples within each sequence therefore share the same class pair and initial noise and differ only in the mixing coefficient. On CIFAR-10, we use 444

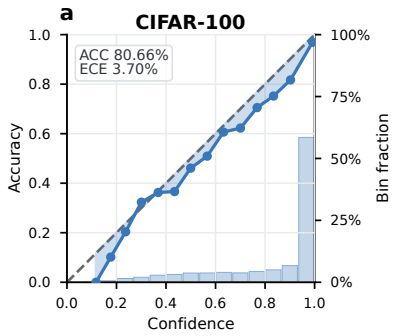


Figure 3: Reliability diagram of the temperature-scaled diffusion classifier on CIFAR-100. The bars show the fraction of test samples in each confidence bin, corresponding to the right y-axis. The dashed diagonal denotes perfect calibration. ACC and ECE are shown in the panel.

random seeds for each of the 45 class pairs, resulting in $M = 19{,}980$ sequences and 159,840 augmented samples. On CIFAR-100, we use 4 random seeds for each of the 4,950 class pairs, resulting in $M = 19{,}800$ sequences and 158,400 augmented samples. On Tiny-ImageNet, we use 2 random seeds for each of the 19,900 class pairs, resulting in $M = 39{,}800$ sequences and 318,400 augmented samples. The EDM sampler uses 20 denoising steps on CIFAR-10 and 25 steps on CIFAR-100 and Tiny-ImageNet.

Augmented-sample selection. After constructing $\mathcal{D}_{\text{aug}}$, the 8 soft-labeled examples within each sequence are reordered in descending order according to the probability assigned to class a. Let $\{\tilde{\mathbf{x}}_{(1)}, \ldots, \tilde{\mathbf{x}}_{(8)}\}$ denote the reordered examples, where $\tilde{q}_{(1),a} \geq \cdots \geq \tilde{q}_{(8),a}$. The first and last positions correspond to the probability edges of classes a and b, while the remaining 6 are treated as intermediate positions. The ordering is determined by the inferred soft labels rather than the mixing coefficients and therefore does not necessarily follow the order of λ.

For each sequence, N_{aug} samples are selected to form the augmented training set $\widetilde{\mathcal{D}}$. When $N_{\text{aug}} = 1$, one of the two edge samples is randomly selected. When $N_{\text{aug}} = 2$, one edge sample is randomly selected together with one sample uniformly drawn from the 6 intermediate positions. $N_{\text{aug}} = 0$ denotes training without augmented samples. The value of N_{aug} is selected using validation ECE for each dataset–backbone pair and is fixed before test evaluation.

B Additional Experimental Analysis

B.1 Comparison with Diffusion Classifier

We compare DiffSL with the diffusion classifier on CIFAR-10 and CIFAR-100. The pretrained diffusion model used by the diffusion classifier is trained on the complete training set. For a fair comparison, we therefore train DiffSL on the complete training set without reserving a validation subset in this experiment. Other experiments use the validation split described in Appendix A. The quantitative results are reported in Table 7, and the corresponding CIFAR-100 reliability diagram is shown in Fig. 3.

B.2 Augmented-sample labeling

Table 8 compares three labeling strategies using the same augmented samples. CSM re-annotation (Luo et al., 2025) uses CLIP to estimate the relative probabilities of the two generating classes. Endpoint interpolation constructs the soft label by linearly interpolating the soft labels of the two endpoints according to the mixing coefficient, $\lambda \mathbf{q}_a + (1 - \lambda)\mathbf{q}_b$. DiffSL assigns soft labels to the augmented samples according to their DC predictions. DiffSL achieves the best performance across all three metrics. The results show that assigning soft labels from DC predictions is more effective than restricting the labels to the generating class pair or inferring them indirectly from the endpoints.

Table 7: Test performance (%) of the diffusion classifier and DiffSL on CIFAR-10 and CIFAR-100 with ResNet-50.

Dataset	Method	ACC↑	ECE↓	AECE↓
CIFAR-10	DC	93.90	1.39	1.36
	Ours	**94.75**	**0.61**	**0.60**
CIFAR-100	DC	80.66	3.70	3.61
	Ours	**80.72**	**1.10**	**0.85**

Table 8: Effect of augmented-sample labeling on CIFAR-100 with ResNet-50. Results are reported as mean ± standard error of the mean (SEM) over three seeds.

Augmented-sample label	ACC↑	ECE↓	AECE↓
CSM re-annotation	78.43±0.16	2.87±0.27	2.79±0.33
DC endpoint interpolation	80.18±0.12	2.03±0.29	2.13±0.23
DiffSL (Ours)	**80.61±0.16**	**1.08±0.17**	**1.11±0.27**