1. Implementation details

Methods for comparison. We provide details of the hyperparameters used for reproducing the results of other methods as follows: (a) ECP [9]: We set the weight of entropy penalty to 0.1 following [9] (b) LS [8]: Following [5, 7], we report the results obtained with $\alpha = 0.05$. (c) FL [7]: We train the models with a fixed regularization parameter ($\gamma$) of 3. (d) Mixup [11]: We use a shape parameter $\alpha$ of 0.2, the best performing one in [11]. (e) FLSD [7]: Following the schedule in [7], we use the parameter $\gamma$ of 5 for the samples whose output probability for the ground-truth class is within [0, 0.2], otherwise we use the parameter $\gamma$ of 3. (f) CRL [6]: We set the balancing parameter of the ranking loss as 1.0. (g) CPC [1]: We set the weights of binary discrimination and binary exclusion losses as 0.1 and 1.0, respectively. (h) MbLS [5]: We train the models with margins of 6 and 10 for the CIFAR10/100 datasets and Tiny-ImageNet dataset, respectively. (i) RegMixup [10]: We obtain the results with a shape parameter $\alpha$ of 10.0 as suggested in [10].

2. More results

$Q$ and $\alpha$. We perform experiments with various combinations of $\alpha$ and $Q$ to further investigate the importance of diverse samples. We adopt the ResNet-50 [2] models trained with M-NDCG for the analyses, and show the calibration performances in terms of ECE and AECE on the validation sets of CIFAR10 [3] and Tiny-ImageNet [4] in Fig. 1 and 2, respectively. From the figures, we can observe three things: (1) Our models using $\alpha$ within the range of [1, 3] achieve the best performances in most cases. As shown in Fig. 2 of the main paper, such distributions enable the model to sample more diverse and large mixup coefficients $\lambda$, suggesting that generating diverse samples with relatively strong interpolations is crucial for our framework. (2) Our models also perform better with larger $\alpha$s than smaller ones, in contrast to vanilla mixup [11, 12]. However, in some cases, the performances worsen with very large $\alpha$. (3) For most $\alpha$ values, both ECE and AECE improve as more augmented samples are used. This indicates that incorporating more diverse ranking relationships based on augmented samples is favorable in our framework. However, using more samples with very small or large $\alpha$ leads to degraded performance. Overall, these observations highlight the importance of using diverse samples and considering a range of $\alpha$ values for effective calibration in our framework.

References

Figure 1: Quantitative results with various combinations of the parameter of Beta distribution ($\alpha$) and the number of augmented samples ($Q$). We plot the variation of both ECE (%) and AECE (%) on the validation split of CIFAR10 [3]. Best viewed in color.

Figure 2: Quantitative results with various combinations of the parameter of Beta distribution ($\alpha$) and the number of augmented samples ($Q$). We plot the variation of both ECE (%) and AECE (%) on the validation split of Tiny-ImageNet [4]. Best viewed in color.