Supplementary Material for Adaptive Calibrator Ensemble: Navigating Test Set Difficulty in Out-of-Distribution Scenarios

We first introduce the experimental setup including training details, dataset split, and computation resources. We also report more metrics (*i.e.,* KSE [7] and BS [1]) in Table A1 and detailed statistical test results of Table 1 in the main paper. Then, we provide more comparative results with Perturbation [16] in Table A5, and we report full results on CIFAR-10-C and ImageNet-C in Table A6 and Table A7, respectively. Lastly, we give more component analysis of the proposed ACE method in Section D.

A. Experimental Setup

A.1. CIFAR-10 Setup

Following the protocol in [6, 12], we use $5,000$ images from the training set of CIFAR-10 as the calibration set. We use ResNet-20 designed for CIFAR-10 and train it using publicly available codes in [12].

A.2. ImageNet Setup

Following the protocol in [6], we divide the validation set of ImageNet into two halves: one for in-distribution test; the other for learning calibration methods. We use ResNet-50, ResNet-152, Vit-Small-Patch32-224 and Deit-Small-Patch16-224. Their weights are publicly provided by PyTorch Image Models (timm-0.5.4) [19].

A.3. Baseline Methods

Our proposed ACE method is used for improving post-hoc methods (*i.e.*, Vector Scaling, Temperature Scaling, and Spline) on OOD test sets. For each baseline, we use the publicly available codes to train the calibration model. We follow the code and use the same training settings (such as regularization, training scheduler, and training hyper-parameters). The codes we used are:
Vector Scaling:
https://github.com/saurabhgarg1996/calibration
Temperature Scaling:
https://github.com/gpleiss/temperature_scaling
Spline:
https://github.com/kartikgupta-at-anu/spline-calibration

A.4. More Metrics for Table 1

We report the ECE (%) result in Table 1. To better prove the effectiveness of our method, we report another two classic metrics: KSE (%) [7] and Brier Score (%) in Table A1. The results in table A1 shows that our method is also effective with these metrics.

A.5. The Statistical Significance Test in Table 1

We adopt the two-sample t-test, which tells whether the performance of the baseline and baseline + ACE has a significant difference. All methods are run for 5 times based on 5 random seeds $(1, 2, 3, 4, 5)$. Given a random seed, we use it to randomly downsample the hard calibration set from the original validation set. For all random seeds, the samples for the baseline are indeed the same. However, when training a calibrator, every mini-batch is randomly sampled and shuffled, thus resulting in randomness. As reported in Table A2, the impact of different random seeds is slight. We also adopt the Welch's t-test in Table A3 to validate this.

A.6. Computation Resource

We use the Pytorch-1.9.1 framework and run all the experiment on one GPU (GeForce RTX 2080 Ti). The CPU is 24 Intel(R) Core(TM) i9-10920X CPU @ $3.50G$Hz.

A.7. Datasets

ImageNet-Validation [2] (https://www.image-net.org);
ImageNet-V2-A/B/C [15]
(https://github.com/modestyachts/ImageNetV2);
ImageNet-Corruption [9]
(https://github.com/hendrycks/robustness);
ImageNet-Sketch [17]
(https://github.com/HaohanWang/ImageNet-Sketch);
ImageNet-Adversarial [10]
(https://github.com/hendrycks/natural-adv-examples);
ImageNet-Rendition [8]
(https://github.com/hendrycks/imagenet-r);
CIFAR-10 [13](https://www.cs.toronto.edu/ kriz/cifar.html);
CIFAR-10-C [9](https://github.com/hendrycks/robustness);

Table A1. We used two other metrics, Brier Score (%), KS-Error (%) [7]. We evaluate two calibrators (Temperature Scaling and Spline). All other settings remain the same with Table 1 of the main paper.

Metric	Methods	ImgNet-V2-A	ImgNet-V2-B	ImgNet-V2-C	ImgNet-S	ImgNet-R	ImgNet-Adv
KSE	UnCal	5.2260	9.5910	4.0399	24.6331	17.8626	50.8544
	Temp.Scaling	4.0937	1.1129	0.8773	15.7880	10.4752	42.6302
	+ACE	3.0661	0.7809	0.8406	**1.0386**	6.7335	38.0691
	Spline	4.4217	1.0765	0.8813	19.6394	13.0808	45.3623
	+ACE	**1.2029**	**0.7239**	**0.3483**	5.8538	**3.5370**	**31.1308**
BS	UnCal	15.7902	13.0527	11.1197	21.6672	18.0285	39.1104
	Temp.Scaling	14.8083	12.6830	10.9561	17.2627	15.2080	30.3974
	+ACE	14.7192	12.6815	10.9532	15.3793	**14.3487**	26.2166
	Spline	14.8779	**12.5798**	10.8702	18.9953	16.1986	32.0494
	+ACE	**14.7086**	12.5804	**10.8640**	**14.9486**	14.6938	**18.8537**

Table A2. The *t-statistic* and p *values* of the two-sample t-test method in Table 1 of main paper. We report the resulting statistics and p values here, which are one-on-one corresponded to the numbers in Table 1. We regard $p < 0.05$ as statistically significant.

Methods		ImgNet-V2-A	ImgNet-V2-B	ImgNet-V2-C	ImgNet-S	ImgNet-R	ImgNet-Adv
Vector Scaling	t-statistic	59.25	37.39	-25.14	355.60	170.03	217.22
	p	$7.31e^{-12}$	$2.87e^{-10}$	$6.70e^{-9}$	$4.37e^{-18}$	$1.60e^{-15}$	$2.25e^{-16}$
Temp. Scaling	t-statistic	615.89	249.42	-195.10	1164.86	800.82	898.46
	p	$5.40e^{-20}$	$7.47e^{-17}$	$5.33e^{-16}$	$3.30e^{-22}$	$6.62e^{-21}$	$2.63e^{-21}$
Spline	t-statistic	120.74	-28.46	60.99	294.01	675.16	109.61
	p	$2.47e^{-14}$	$2.50e^{-9}$	$5.80e^{-12}$	$2.00e^{-17}$	$2.59e^{-20}$	$5.36e^{-14}$

Table A3. The *t-statistic* and p *values* of the Welch's t-test in Table 1 of main paper. We report the resulting statistics and p values here, which are one-on-one corresponded to the numbers in Table 1. We regard $p < 0.05$ as statistically significant.

Methods		ImgNet-V2-A	ImgNet-V2-B	ImgNet-V2-C	ImgNet-S	ImgNet-R	ImgNet-Adv
Vector Scaling	t-statistic	59.25	37.39	-25.14	355.60	170.03	217.22
	p	$4.85e^{-7}$	$3.05e^{-6}$	$1.48e^{-5}$	$3.75e^{-10}$	$7.17e^{-9}$	$2.68e^{-9}$
Temp. Scaling	t-statistic	615.89	249.42	-195.10	1164.86	800.82	898.46
	p	$4.16e^{-11}$	$1.55e^{-9}$	$4.14e^{-9}$	$3.25e^{-12}$	$1.45e^{-11}$	$9.20e^{-12}$
Spline	t-statistic	120.74	-28.46	60.99	294.01	675.16	109.61
	p	$2.82e^{-8}$	$9.06e^{-6}$	$4.32e^{-7}$	$8.02e^{-10}$	$2.88e^{-11}$	$4.15e^{-8}$

B. More Comparison

B.1. Comparison with Perturbation

In Table A4, we compare our method with a recent OOD calibration method Perturbation [16]. In Table A4, we observe that Perturbation improves the baselines on Level 5 of ImageNet- C. In fact, these test sets contain data that are seriously out of distribution. However, for datasets that lean towards being in-distribution, *e.g.*, Level 1 in ImageNet-C, Perturbation worsens the baselines. A probable reason is that the diverse calibration set where Perturbation is trained is closer to heavily OOD data (Level-5). In comparison, our method (ACE) adapts to various test sets through the weighting scheme and yields improvement with statistical significance in most test cases.

B.2. Comparison with TransCal

In Table A5, we compare our method with a recent OOD calibration method TranCal [18]. In Table A5, we observe that TransCal is inferior to our method on the ImageNet-S dataset with ResNet-50.

C. Results on ImageNet-C and CIFAR-10-C

In Table 3 of the main paper, we report the mean ECE (%) across 16 different types of data shift at intensity 5. In addition, we report the complete ECE results on CIFAR-10-C and ImageNet-C at intensity 5 in Table A6 and Table A7. We observe that our method effectively improves the baselines (Spline) and gives state-of-the-art calibration accuracy under 2 out of 3 quartiles and mean value on both CIFAR-10-C and ImageNet-C.

Table A4. Method comparison on **ImageNet-C** datasets [9]. We report ECE (%) for top-1 predictions (in %) of the ResNet-152 model. For each level of corruption (column), we report the average ECE using 25 bins with lowest numbers in **bold** and second lowest <u>underlined</u>. ACE improves calibration performance of two post-hoc calibration methods on all datasets.

	Corruption Intensity				
Method	Level 1	Level 2	Level 3	Level 4	Level 5
Uncalibrated	6.0684	7.8617	9.7938	12.3911	15.5049
Temperature Scaling (TS)	<u>2.4880</u>	**2.7976**	<u>3.7996</u>	5.1836	7.7213
Temperature + Perturbation	9.3084	8.6574	7.6707	5.7594	4.3672
Temperature + ACE	2.9733	<u>3.1130</u>	**3.1306**	**3.1494**	<u>4.3034</u>
Spline	**1.8049**	3.1690	5.2388	7.8672	11.0547
Spline + Perturbation	9.6207	8.1570	6.7643	5.1064	5.2777
Spline + ACE	3.6982	4.2046	4.2944	<u>3.7231</u>	**3.9707**

Table A5. Method comparison on **ImageNet-V2-A, ImageNet-V2-B, ImageNet-V2-C, and ImageNet-S** datasets. Following the protocol in [18], we report ECE (%) for top-1 predictions (in %) of the ResNet-50 model.

Method	ImageNet-V2-A	ImageNet-V2-B	ImageNet-V2-C	ImageNet-S
Uncalibrated	9.50	6.23	4.31	22.32
Temperature Scaling	4.44	2.73	1.68	16.27
TransCal	12.26	4.43	1.86	8.10
Ours	3.56	2.56	1.70	7.53

Table A6. Full results on CIFAR-10-C datasets [9]. We report the lower quartile (25-th percentile), median (50-th percentile), mean and upper quartile (75-th percentile) of ECE computed across 16 different types of data shift at intensity 5 with lowest numbers in **bold** and second lowest <u>underlined</u>.

Metric		Method								
		Vanilla	Temp Scaling	Ensemble	SVI	LL SVI	SVI -AvUTS	SVI -AvUC	Spline	Spline +Ours
ECE	lower quartile	0.2121	0.0997	0.0549	0.0925	0.2027	<u>0.0466</u>	**0.0398**	0.2045	0.0783
	median quartile	0.3022	0.1834	0.1054	0.2146	0.3077	0.1516	<u>0.1107</u>	0.3007	**0.1071**
	mean	0.3151	0.1993	0.1611	0.2389	0.3267	0.1585	<u>0.1374</u>	0.3382	**0.1272**
	upper quartile	0.4148	0.2915	0.2551	0.3636	0.4246	0.2345	<u>0.2303</u>	0.4376	**0.1522**

Table A7. Full results on ImageNet-C datasets [9]. We report the lower quartile(25-th percentile), median (50-th percentile), mean and upper quartile (75-th percentile) of ECE computed across 16 different types of datashift at intensity 5 with lowest numbers in **bold** and second lowest <u>underlined</u>.

Metric		Method								
		Vanilla	Temp Scaling	Ensemble	SVI	LL SVI	SVI -AvUTS	SVI -AvUC	Spline	Spline +Ours
ECE	lower quartile	0.1244	0.0959	0.0503	0.0722	0.1212	0.0420	<u>0.0319</u>	0.0575	**0.0233**
	median quartile	0.1737	0.1392	0.0900	0.1144	0.1684	0.0807	**0.0447**	0.1143	<u>0.0452</u>
	mean	0.1942	0.1600	0.0880	0.1188	0.1868	0.0800	<u>0.0542</u>	0.1147	**0.0477**
	upper quartile	0.2744	0.2364	0.1264	0.1723	0.2676	0.1275	<u>0.0696</u>	0.1363	**0.0606**

Table A8. The adaptive α that we adopt in Table 1 and Table 2 of main paper.

Model	ImgNet-Val	ImgNet-V2-A	ImgNet-V2-B	ImgNet-V2-C	ImgNet-S	ImgNet-R	ImgNet-Adv
ResNet	0.994080	0.918328	0.972311	0.989697	0.63765	0.709984	0.682187
Vit	0.998655	0.896980	0.969018	0.98561	0.538366	0.674307	0.637850
Deit	0.998741	0.912270	0.967555	0.999048	0.612748	0.648445	0.618136

Table A9. Method comparison on CIFAR-10-C and ImageNet-C datasets with ResNet-20 and ResNet-50, respectively. Following the protocol in [14], we report mean ECE (%) across 16 different types of data shift at intensity 5 with lowest numbers in **bold** and second lowest underlined.

Dataset	Vanilla	SVI	SVI -AvUC	Spline	Spline +ACE	Spline +Estimation
CIFAR-10-C	0.1942	0.2389	0.1374	0.3382	**0.1264**	<u>0.1298</u>
ImageNet-C	0.3151	0.1188	<u>0.0542</u>	0.1147	**0.0477**	0.0576

Table A10. Calibration performance of our method integrated with Temperature Scaling on one in-distribution test set and six OOD test sets. ECE (25bins, %) for top-1 predictions. Here we $\mathcal{D}_o$ with the sample size of $\mathcal{D}_h$ $(5,884)$.

Method	ImgNet-Val	ImgNet-V2-A	ImgNet-V2-B	ImgNet-V2-C	ImgNet-S	ImgNet-R	ImgNet-Adv
Temp.Scaling	1.9670	4.3571	2.7234	1.7880	15.6735	10.3832	42.5225
+ACE	1.9623	3.4842	2.5458	1.6764	10.3131	6.6726	37.9957

Table A11. Calibration performance of our method integrated with Temperature Scaling on one in-distribution test set and six OOD test sets. ECE (25 bins, %) for top-1 predictions. We use LCNet-050 and TinyNet-E, which have 60.094% and 59.856% top-1 accuracy, respectively on the validation set of ImageNet dataset. (Note IN is short for ImageNet)

Model	Method	IN-Val	IN-V2-A	IN-V2-B	IN-V2-C	IN-S	IN-R	IN-Adv
LCENet-050	Temp.Scaling	1.8293	6.6047	2.9681	1.6949	20.3415	18.9839	43.1683
	+ACE	1.8238	4.8591	2.2639	1.7516	14.0055	15.3397	39.2584
TinyNet-E	Temp.Scaling	1.3888	6.8949	2.7991	1.7194	22.4438	20.7810	41.3513
	+ACE	1.3857	5.4262	2.4606	1.8311	17.1741	17.7259	38.0800

Table A12. Calibration performance of our method integrated with Temperature Scaling and Spline on the in-distribution and OOD iWildCam-WILDS dataset. ECE (25bins, %) for top-1 predictions and ResNet-50 classifier is used.

Dataset	Uncal.	Temp.Scaling	Temp.Scaling+Ours	Spline	Spline+Ours
iWildCam-WILDS-ID	14.2701	2.6786	2.5833	3.8142	3.6965
iWildCam-WILDS-OOD	13.5552	4.8231	3.9738	4.9902	4.8425

Table A13. Calibration performance of different combination schemes. ECE (25bins, %) for top-1 predictions is reported. Spline baseline and ResNet-152 classifier is used.

Method	ImageNet-V2-A	ImageNet-V2-B	ImageNet-V2-C	ImageNet-S	ImageNet-R	ImageNet-Adv
Uncal.	9.5016	6.2311	4.3117	24.6332	17.8621	50.8544
$\mathbf{z}_o^{\alpha} \otimes \mathbf{z}_h^{1-\alpha}$	5.0091	2.7478	1.3357	6.4506	10.2066	28.4341
$\alpha \cdot \mathbf{z}_o + (1-\alpha) \cdot \mathbf{z}_h$	2.8201	2.0235	1.0550	6.9264	6.8533	31.0926

Table A14. Following the protocol in Gong *et al*. [4], we evaluate proposed ACE under domain generalization setting. We use Spline-based ACE and report ECE (25 bins, %) for top-1 predictions.

Test Set	Uncalibrated	Gong *et al*. [4]	ACE (Spline)
A→C	11.84	12.53	4.82
A→P	6.81	5.56	2.84
A→R	4.31	6.25	3.77

D. More Component Analysis

D.1. An Alternative Method

In L210-216 of the main paper, we mentioned that a possible way to calibrate OOD data is to estimate its difficulty and create a calibration set that has a closer difficulty level with the OOD test dataset. Moreover, according to Sec. 3.5 of the main paper, the average confidence score could serve as an unsupervised indicator to the degree of how out-of-distribution a test set is [5]. Here, we propose another post-hoc calibration method for OOD calibration. Specifically, we first estimate the error rate of a test set [3]:

$$error_{\mathcal{D}_{test}} = (1-\text{Acc}(\mathcal{D}_o))+(\text{avgConf}(\mathcal{D}_o)-\text{avgConf}(\mathcal{D}_{test})). \tag{a-1}$$

Thus, we can compute $d_{\mathcal{D}_{test}}$ as:

$$d_{\mathcal{D}_{test}} = \frac{error_{\mathcal{D}_{test}}}{1-error_{\mathcal{D}_{test}}}. \tag{a-2}$$

According to Table A9, our estimation method is also shown to be effective. Specifically, it has the second lowest ECE on CIFAR-10-C and is only 0.0034 higher than SVI-AvUC on ImageNet-C.

D.2. Easy calibration set and hard calibration set have the same number of samples for tuning the function

The size of $\mathcal{D}_h$ in our submission is $5,884$. We randomly sample the easy calibration set $\mathcal{D}_o$ into the same size ($5,884$), the difficulty of which remains the same due to random sampling. We report performance calibration (ECE, %) of Temperature Scaling and our improved version on all the seven test sets below. The ResNet-152 classifier is used. The results in Table A10 show that our method remains beneficial, i.e., achieving lower ECE when combined with Temperature Scaling, when the easy and the hard calibration sets have the same size. The results show that our method remains beneficial, *i.e.,* achieving lower ECE when combined with Temperature Scaling, when the easy and the hard calibration sets have the same size.

D.3. The original calibration set is not easy

In Sec. 3.5 of main paper, we mentioned that difficulty is a relative concept and depends on the classifier. Note that for a weaker classifier, a certain dataset will be harder. With this in mind, we experimented with two weaker classifiers, (i.e., harder $\mathcal{D}_o$) and observed that our method is still effective. Specifically, we adopt LCNet-050 and TinyNet-E, which have 60.094% and 59.856% top-1 accuracy, respectively on the ImageNet-Val dataset. We apply Temperature Scaling with the proposed method to the two classifiers and report calibration performance (ECE, $\%$) below. These results in Table A11 show that our method consistently improves Temperature Scaling when the "easy calibration set" has high difficulty (*i.e.,* is not easy).

D.4. More types of OOD test sets

We further provide the calibration results (ECE, $\%$) on another challenging and diverse dataset iWildCam-WILDS [11] with the ResNet-50 classifier. iWildCam-WILDS is an animal species classification dataset, where the distribution shift arises due to changes in camera angle, lighting, and background. Table A12 shows that our method can also improve the calibration performance on iWildCam-WILDS, especially, improves temperature scaling by 0.9% decrease in ECE on the OOD test set.

D.5. Combination scheme of adaptive weight α

In the experiment section, we show the effectiveness of the simple linear combination of these two extreme logits. We further test another combination scheme in this section. According to Table A13, it decreases ECE (%) of uncalibration but is slightly worse than current scheme on ImageNet-V2 and ImageNet-R.

D.6. ACE under the domain generalization setting

In L497-L503 of main paper, we discussed the application scenarios where we have access to calibration datasets from multiple domains. Here, we evaluate our ACE with Spline baseline under domain generalization setting, where multiple labeled source domains are given. Table A14 shows ACE achieves lower ECE with Gong *et al*. [4].

References

[1] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

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

[3] S. Garg, S. Balakrishnan, Z. C. Lipton, B. Neyshabur, and H. Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. 2022.

[4] Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8967, 2021.

[5] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1144, 2021.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[7] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021.

[8] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[10] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[11] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[12] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33:18237–18248, 2020.

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[14] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David 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 *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2019.

[15] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[16] Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10132, 2021.

[17] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[18] Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020.

[19] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.