Robust Calibration of Large Vision-Language Adapters Supplementary material

A Supplementary Experiments

A.1 Proof of Proposition 1

If we add a strictly positive constant value $a \in \mathbb{R}_{++}$ to all the elements of a positive logit vector $\boldsymbol{l} \geq \mathbf{0}$, the modified vector is $\boldsymbol{l}' = \boldsymbol{l} + a\mathbf{1}$. Considering $\sigma(\cdot)$ as the softmax function, we can then rewrite the softmax prediction for class k as (we omit the temperature scalar τ for simplicity, as it does not have any impact on this proof):

$$\begin{aligned}\sigma_k(\boldsymbol{l}') = \sigma_k(\boldsymbol{l} + a\mathbf{1}) = \frac{\exp(l_k + a)}{\sum_{j=1}^{K} \exp(l_j + a)} &= \frac{\exp(l_k)\exp(a)}{\sum_{j=1}^{K} \exp(l_j)\exp(a)} \\ &= \frac{\exp(a)}{\exp(a)} \frac{\exp(l_k)}{\sum_{j=1}^{K} \exp(l_j)} \\ &= \frac{\exp(l_k)}{\sum_{j=1}^{K} \exp(l_j)} \end{aligned} \tag{7}$$

This proves the first part of Proposition 1.
Showing $\|\boldsymbol{l}'\| \geq \|\boldsymbol{l}\|$.

Considering $\|\boldsymbol{l}'\|$ as $\|\boldsymbol{l} + a\mathbf{1}\|$, we have:

$$\begin{aligned}\|\boldsymbol{l} + a\mathbf{1}\| - \|\boldsymbol{l}\| &= \sqrt{\sum_{j=1}^{K}(l_j + a)^2} - \sqrt{\sum_{j=1}^{K} l_j^2} \\ &= \sqrt{\sum_{j=1}^{K}(l_j^2 + 2al_j + a^2)} - \sqrt{\sum_{j=1}^{K} l_j^2} \\ &= \sqrt{\sum_{j=1}^{K} l_j^2 + 2a\sum_{j=1}^{K} l_j + Ka^2)} - \sqrt{\sum_{j=1}^{K} l_j^2}. \end{aligned} \tag{8}$$

Since $Ka^2 \geq 0$, and $2a\sum_{j=1}^{K} l_j > 0$ (we assume $a \in \mathbb{R}_{++}$ and $\boldsymbol{l} \geq \mathbf{0}$), the first square root is greater than the second one. This results in a positive value for $\|\boldsymbol{l}+a\mathbf{1}\|-\|\boldsymbol{l}\|$, which demonstrates that $\|\boldsymbol{l}+a\mathbf{1}\| = \|\boldsymbol{l}'\| \geq \|\boldsymbol{l}\|$. Thus, Proposition 1 is proved, *i.e., an increase in the logit norm does not necessarily modify the confidence of the predictions.*

A.2 Proof of Proposition 2

Considering a scalar $a > 1$, $s = a - 1$, and $\sigma(\cdot)$ the softmax function, we can define for the predicted class $k(k = \arg\max_j(l_j))$:

$$\sigma_k(a\boldsymbol{l}) = \frac{e^{al_k}}{\sum_{j=1}^{K} e^{al_j}} = \frac{e^{(s+1)l_k}}{\sum_{j=1}^{K} e^{(s+1)l_j}} = \frac{e^{l_k}}{\sum_{j=1}^{K} e^{l_j + s(l_j - l_k)}} \tag{9}$$

where $l_k = \max_j(l_j)$.

If we consider now that for any $j \in [1, 2, ..., K]$, if $j \neq k$, then $l_j - l_k \leq 0$, we have that:

$$l_j + s(l_j - l_k) < l_j \text{ for } j \neq k \tag{10}$$

Therefore, $e^{l_j + s(l_j - l_k)} < e^{l_j}$ for $j \neq k$ (note that for $j = k$, the exponent remains l_k). Thus, the sum in the denominator for $\sigma_k(a\boldsymbol{l})$ is smaller than the sum in the denominator for $\sigma_k(\boldsymbol{l})$:

$$\sum_{j=1}^{K} e^{l_j + s(l_j - l_k)} < \sum_{j=1}^{K} e^{l_j} \tag{11}$$

Since the numerator e^{l_k} remains the same, we have:

$$\sigma_k(a\boldsymbol{l}) = \frac{e^{l_k}}{\sum_{j=1}^{K} e^{l_j + s(l_j - l_k)}} > \frac{e^{l_k}}{\sum_{j=1}^{K} e^{l_j}} = \sigma_k(\boldsymbol{l}) \tag{12}$$

This proves that *increasing the logit range (by scaling the logits with a factor $a > 1$) increases the confidence of the predicted class*:

$$\sigma_k(a\boldsymbol{l}) > \sigma_k(\boldsymbol{l}) \tag{13}$$

Showing $R(a\boldsymbol{l}) > R(\boldsymbol{l})$.

Let us denote the range $R(\boldsymbol{l})$ as:

$$R(\boldsymbol{l}) = \max(\boldsymbol{l}) - \min(\boldsymbol{l}) \tag{14}$$

For a given scalar $a > 1$, we can scale a logit vector $\mathbf{l}$, whose maximum and minimum values are also scaled:

$$\max(a\boldsymbol{l}) = a \max(\boldsymbol{l}) \quad \text{and} \quad \min(a\boldsymbol{l}) = a \min(\boldsymbol{l}) \tag{15}$$

Following our definition of range $R(\boldsymbol{l})$:

$$R(a\boldsymbol{l}) = a \max(\boldsymbol{l}) - a \min(\boldsymbol{l}) \tag{16}$$

where a can be factored out, leading to:

$$R(a\boldsymbol{l}) = a(\max(\boldsymbol{l}) - \min(\boldsymbol{l})) = aR(\boldsymbol{l}) \tag{17}$$

Last, as $a > 1$, we have that:

$$aR(\boldsymbol{l}) > R(\boldsymbol{l}) \tag{18}$$

which proves that $R(a\boldsymbol{l}) > R(\boldsymbol{l})$. Thus, Proposition 2 is proved.

A.3 Few shot domain generalization in adapters

We supplement the results depicted in the main manuscript for few-shot adapter calibration (Table 1) by providing results for individual datasets. In particular, we considered in this experiment the popular adapter techniques CLIP-Adpater [10], TIP-Adapter [63], and TaskRes(e) [60] and, additionaly, ZS-LP [50]. In Table 5, to each of the above methods, we compare our contributions **ZS-Norm**, **Penalty**, and **SaLS** for ResNet-50 and ViT-B/16 architectures. The adapters are initially trained with source ImageNet dataset, and evaluated under ImageNet distributional shifts, *i.e.*-V2 [46], Sketch [55], Adversarial [17], and Rendition [15]. The classification metric accuracy and calibration metric ECE are reported for the individual datasets. CLIP-Adpater and TIP-Adapter are sensitive to the **ZS-Norm** technique, which is possible due to the method's dependency on specific hyper-parameter settings [50]. Even for these methods, **Penalty** consistently improves calibration while retaining or improving the accuracy. Last, our post-processing technique **SaLS** can retain the accuracy and assist in calibration consistently across all approaches.

A.4 Few shot domain generalization in prompt learning

In the following, we extend the evaluation of few-shot prompt learning generalization with per-dataset metrics, and cross-domain generalization.

ImageNet shifts. In this experiment, prompt learning methods were adapted in 16-shot ImageNet, and evaluated in its corresponding domain drifts (OOD). In this section, we complement the results in Table 2 with detailed per-dataset metrics and additional prompt learning methods. In particular, we evaluate our proposed calibration methods when applied to CoOp [66], CoCoOp [65], ProGrad [67], and MaPLe [22]. We evaluate CoOp and CoCoOp for both ResNet-50 and ViT-B/16 architectures. As MaPLe is specifically designed for transformer architectures, we consider the ViT-B/16 CLIP backbone. Analogously for ResNet-50, we consider the prompt-aligned gradient technique ProGrad. These results are presented in Table 6. Following the earlier reported trends, **Penalty** and **SaLS** consistently provide better calibration and accuracy. In comparison with applying **ZS-Norm** in Adapters, using them in prompt learning provides stable results, and often provides improved calibration compared to the baseline. It is noteworthy to mention that Prompt learning methods such as CoCoOp, ProGrad, and MaPLe are designed for improved generalization, and thus provide better performance than previously evaluated adapters in Section A.3. Despite this, our proposed range re-normalization technique can improve the calibration even for these methods, supporting our observation that the range of the logit indeed plays a key role in calibration.

Cross-domain generalization. This additional experiment evaluates prompt learning methods adapted in ImageNet in 10 fine-grained tasks. These 10 tasks include

Table 5: Detailed results for robust adapters calibration Performance of four ImageNet OOD (V2, S, A, R) datasets in different adapters for the proposed **ZS-Norm**, **Penalty**, and **SaLS**. These results provide per-dataset performance, and thus complement Table 1 in the main manuscript.

Method	V2		S		A		R		OOD Mean	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
Zero-Shot [45]	51.5	3.25	33.33	3.17	21.68	21.32	55.96	0.98	40.62	7.18
ZS-LP [50]	51.22	11.18	27.69	16.97	17.44	34.2	48.95	11.88	36.33	18.56
w/ **ZS-Norm**	50.95	2.78	28.87	7.12	17.48	23.72	51.37	1.81	**37.17**	**8.86**
w/ **Penalty**	51.38	9.31	28.36	14.42	17.43	31.79	50.69	8.21	36.97	15.93
w/ **SaLS**	51.22	5.35	27.69	12.2	17.44	25.85	48.95	7.5	36.33	12.73
CLIP-Adapter [10]	49.20	9.40	25.61	13.28	15.71	31.33	45.75	7.79	34.07	15.45
w/ **ZS-Norm**	45.83	36.47	21.65	17.28	13.69	4.09	39.08	27.23	30.06	21.27
w/ **Penalty**	49.92	5.76	26.16	9.23	17.23	26.96	47.50	2.94	**35.2**	11.22
w/ **SaLS**	49.20	2.64	25.61	5.27	15.71	24.42	45.75	3.45	34.07	**8.95**
TIP-Adapter(f) [63]	54.37	13.46	33.58	14.49	20.79	38.13	57.07	10.08	41.45	19.04
w/ **ZS-Norm**	53.52	15.44	34.13	15.04	21.41	38.63	57.84	10.10	41.73	19.8
w/ **Penalty**	55.36	8.14	35.96	8.09	23.07	29.08	60.53	3.39	**43.73**	12.18
w/ **SaLS**	54.37	3.09	33.58	5.43	20.79	22.51	57.07	1.49	41.45	**8.13**
TaskRes(e) [60]	55.07	4.91	32.66	9.13	20.33	27.52	56.67	3.42	41.18	11.25
w/ **ZS-Norm**	55.18	2.51	32.53	7.45	20.35	25.05	57.13	1.26	**41.3**	9.07
w/ **Penalty**	55.23	4.06	32.83	8.46	19.63	27.77	57.48	2.17	41.29	10.62
w/ **SaLS**	55.07	2.46	32.66	8.24	20.33	23.43	56.67	1.98	41.18	**9.03**

(a) ResNet-50

Method	V2		S		A		R		OOD Mean	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
Zero-Shot [45]	60.71	2.37	46.18	4.76	47.73	8.41	73.98	3.56	57.15	4.78
ZS-LP [50]	60.43	10.54	41.39	17.03	42.15	20.42	70.34	3.72	53.58	12.93
w/ **ZS-Norm**	60.20	4.26	41.81	9.52	42.48	12.90	70.85	1.63	53.84	**7.08**
w/ **Penalty**	60.69	8.65	41.74	14.75	42.64	18.05	70.74	2.11	**53.95**	10.89
w/ **SaLS**	60.43	5.99	41.39	13.19	42.15	14.92	70.34	1.15	53.58	8.81
CLIP-Adapter [10]	59.88	4.77	38.73	8.86	40.13	16.20	63.71	1.43	50.61	7.82
w/ **ZS-Norm**	57.44	16.57	38.71	11.67	38.08	4.19	64.69	17.68	49.73	12.53
w/ **Penalty**	60.41	2.67	40.60	6.39	39.64	13.66	67.16	2.80	**51.95**	6.38
w/ **SaLS**	59.88	2.62	38.73	1.82	40.13	10.53	63.71	2.53	50.61	**4.38**
TIP-Adapter(f) [63]	43.50	56.17	26.64	72.79	27.29	72.24	46.01	53.31	25.86	63.63
w/ **ZS-Norm**	45.68	54.26	29.33	70.57	36.17	63.74	55.38	44.49	41.64	58.27
w/ **Penalty**	51.78	42.77	38.44	48.24	39.15	50.67	67.55	22.22	**49.23**	**40.98**
w/ **SaLS**	43.50	42.55	26.64	50.73	27.29	50.59	46.01	33.62	35.86	44.37
TaskRes(e) [60]	64.01	4.72	45.91	10.14	46.87	14.42	75.26	0.78	58.01	7.52
w/ **ZS-Norm**	64.06	2.39	46.40	7.32	47.61	11.01	75.55	2.16	**58.41**	**5.72**
w/ **Penalty**	64.17	4.03	46.36	8.85	47.39	12.57	75.31	1.16	58.31	6.65
w/ **SaLS**	64.01	2.26	45.91	9.22	46.87	11.77	75.26	1.57	58.01	6.21

(b) ViT-B/16

different target categories, different from the set existing in ImageNet, and evaluate the robustness of the zero-shot capabilities of the learned prompts on new categories. We evaluate the 10 few-shot benchmarks with CoOp based on the prompt learned with the 16-shot setting trained on ImageNet. Results are depicted in Table 7. As reported in the literature [65], the prompt learned by CoOp

Table 6: Detailed results for prompt learning calibration Performance of four ImageNet OOD (V2, S, A, R) datasets in different prompt learning techniques for the proposed **ZS-Norm**, **Penalty**, and **SaLS**. These results provide disentangled results for each task and and serves as a supplement to Table 2 in the main manuscript.

Method	V2		S		A		R		OOD Mean	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
Zero-Shot [45]	51.5	3.25	33.33	3.17	21.68	21.32	55.96	0.98	40.62	7.18
CoOp [66]	55.14	3.94	32.10	6.97	22.35	27.94	53.85	5.01	40.86	10.97
w/ **ZS-Norm**	54.85	3.67	33.01	6.40	22.28	27.44	56.20	3.26	41.59	10.19
w/ **Penalty**	55.02	2.04	34.04	3.60	22.55	24.85	55.88	1.76	**41.87**	8.06
w/ **SaLS**	55.14	1.54	32.10	5.68	22.35	21.95	53.85	2.11	40.86	**7.82**
CoCoOp [65]	55.74	2.19	35.33	3.54	23.69	23.66	58.66	1.37	43.36	7.69
w/ **ZS-Norm**	55.12	3.61	35.74	2.83	24.25	17.64	59.69	4.40	43.70	7.12
w/ **Penalty**	55.10	1.76	35.83	0.65	24.41	19.19	60.10	3.01	**43.86**	**6.15**
w/ **SaLS**	55.74	1.40	35.33	4.09	23.69	21.19	58.66	0.60	43.36	6.82
ProGrad [67]	55.69	2.05	34.31	3.18	22.41	24.46	56.87	0.93	42.32	7.66
w/ **ZS-Norm**	55.89	1.82	33.68	3.83	22.81	24.44	56.46	1.83	42.21	7.98
w/ **Penalty**	55.26	1.43	34.22	2.5	23.4	22.56	57.39	0.85	**42.57**	**6.84**
w/ **SaLS**	55.69	1.42	34.31	3.78	22.41	21.38	56.87	1.01	42.32	6.90

(a) ResNet-50

Method	V2		S		A		R		OOD Mean	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
Zero-Shot [45]	60.76	2.45	46.33	7.65	47.68	8.46	74.01	3.60	57.54	6.29
CoOp [66]	64.25	3.75	46.33	7.65	48.57	14.41	74.48	0.64	58.41	6.61
w/ **ZS-Norm**	64.13	1.64	47.40	2.99	48.71	10.20	74.74	2.57	58.75	**4.35**
w/ **Penalty**	64.64	1.81	47.62	4.60	49.03	11.06	75.44	2.18	**59.18**	4.91
w/ **SaLS**	64.25	1.14	46.33	5.94	48.57	9.82	74.48	2.73	58.41	4.90
CoCoOp [65]	64.24	2.18	48.45	5.00	50.12	10.99	76.16	1.14	59.74	4.83
w/ **ZS-Norm**	64.05	2.63	48.66	0.91	50.53	6.44	76.36	5.78	59.90	3.94
w/ **Penalty**	64.14	1.63	48.96	1.84	50.56	8.31	77.12	3.79	**60.20**	**3.89**
w/ **SaLS**	64.24	1.55	48.45	5.21	50.12	9.69	76.16	2.80	59.74	4.81
MaPLe [22]	64.06	1.73	48.77	3.61	50.69	8.59	76.74	2.6	60.07	4.13
w/ **ZS-Norm**	64.19	2.01	48.87	0.82	50.69	7.26	76.61	4.29	60.09	**3.59**
w/ **Penalty**	64.25	1.99	49.32	2.11	51.35	7.16	77.56	3.88	**60.62**	3.78
w/ **SaLS**	64.06	1.57	48.77	5.05	50.69	7.69	76.74	3.19	60.07	4.38

(b) ViT-B/16

on ImageNet is not sufficient to adapt to the diverse fine-grained tasks, thereby providing lower accuracy and calibration than the original vision-language model (*i.e.* zero-shot). It is worth mentioning that incorporating our logit range normalization techniques, particularly SaLS, provides consistent improvement in calibration compared to the original prompt learning approach CoOp.

A.5 Test time prompt tuning with ImageNet OOD benchmark

Test time prompt tuning methods, such as TPT [49], provide a provision to infer an individual sample directly during the test time. In this supplementary experiment, we analyze our methods with TPT for the ImageNet OOD datasets and

Table 7: Cross-Domain Generalization Performance of CoOp adapted on ImageNet with 16-shots using ResNet-50, and evaluated on 10 fine-grained tasks. Best results (excluding ZS) in bold.

Method	CAL	PET	CAR	FLW	FOO	AIR	SUN	DTD	SAT	UCF	**Avg.**
Zero-shot [45]	85.92	85.72	55.63	65.98	77.31	17.07	58.52	42.32	37.49	61.46	58.74
CoOp	85.11	84.33	52.39	57.33	72.53	13.44	55.96	33.04	24.7	56.30	53.51
w/ **ZS-Norm** ACC	85.76	81.03	54.02	55.66	73.72	15.24	56.43	35.64	24.53	54.03	<u>53.61</u>
w/ **Penalty**	85.88	81.96	50.54	55.95	74.47	12.81	56.36	36.70	30.53	55.11	**54.03**
w/ **SaLS**	85.11	84.33	52.39	57.33	72.53	13.44	55.96	33.04	24.7	56.30	53.51
Zero-shot [45]	4.27	7.04	4.48	4.51	3.09	3.10	3.51	4.87	4.05	5.44	4.44
CoOp	3.20	4.01	4.67	4.49	0.47	7.15	3.41	17.08	17.61	3.87	6.60
w/ **ZS-Norm** ECE	2.60	3.24	7.64	5.74	1.39	5.55	2.56	15.66	10.54	5.18	6.01
w/ **Penalty**	3.15	3.42	7.84	5.62	4.43	6.79	2.57	12.61	8.09	3.11	<u>5.76</u>
w/ **SaLS**	4.92	3.46	2.65	3.68	2.23	7.84	3.57	4.25	15.53	2.48	**5.06**

(a) ResNet-50

Method	CAL	PET	CAR	FLW	FOO	AIR	SUN	DTD	SAT	UCF	**Avg.**
Zero-shot [45]	92.94	89.13	65.34	71.21	86.09	24.75	62.58	44.56	47.86	66.69	65.12
CoOp	92.29	87.60	62.85	60.66	84.50	17.40	61.14	40.37	45.93	66.38	61.91
w/ **ZS-Norm** ACC	89.66	87.76	62.55	65.45	84.76	18.06	61.90	39.42	44.88	64.45	61.89
w/ **Penalty**	93.06	88.58	63.30	68.09	84.72	18.72	62.92	39.60	44.91	65.16	**62.91**
w/ **SaLS**	92.29	87.60	62.85	60.66	84.50	17.40	61.14	40.37	45.93	66.38	<u>61.91</u>
Zero-shot [45]	5.50	4.88	4.08	4.30	2.56	3.31	1.95	3.60	4.53	2.83	3.75
CoOp	1.93	1.54	5.43	8.05	2.56	14.39	4.65	14.07	7.40	4.78	6.48
w/ **ZS-Norm** ECE	2.83	3.68	9.59	4.52	3.42	7.89	1.88	7.60	5.27	3.65	<u>4.83</u>
w/ **Penalty**	2.76	4.10	8.48	2.16	2.16	5.41	2.19	12.97	4.41	5.65	5.03
w/ **SaLS**	4.11	2.92	4.35	6.43	3.95	10.88	1.61	2.73	4.79	2.74	**4.45**

(b) ViT-B/16

complement results on fine-grained datasets depicted in Table 3. The numbers for each ImageNet OOD dataset comparing TPT with our methods are reported in Table 8. In this setting, Zero-Shot inference is better calibrated than the TPT, even when the accuracy increases with adaptation. This drastic degradation in calibration may be largely due to the use of entropy, which favours larger distances between the winner and other logits, thereby increasing the logit range. Through our methods, we have attempted to restrict the logit range from going beyond the Zero-Shot range, providing us the expected improvement in calibration, and even in accuracy in some cases.

A.6 Additional experiments for Test time prompt tuning

In this section, we further study TPT with our methods on 11 few-shot benchmarks. In particular, we complement the results provided in the main manuscript (Table 3) with ImageNet results and the CLIP ViT-B/16 model. In Tab. 9, the overall (Avg.) results show that our calibration methods are better than the baselines, especially in calibration. As expected from a good post-processing technique, **SaLS** retains the accuracy and consistently improves the calibration across tasks. Importantly, even for C-TPT, our method **SaLS** still improves the calibration, proving that even with the best prompt choice for calibration, there

Table 8: Additional tasks for test time prompt tuning calibration Performance of ImageNet OOD datasets (V2, S, A, R) with TPT for the proposed **ZS-Norm**, **Penalty**, and **SaLS**. These results supplement the ones depicted in Table 3 in the main manuscript by integrating four more adaptation tasks.

Method	V2		S		A		R		OOD Mean	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
Zero-Shot [45]	51.50	3.25	33.33	3.17	21.68	21.32	55.96	0.98	40.62	7.18
TPT [49]	54.97	13.77	35.03	15.28	26.61	30.82	59.00	10.45	43.90	17.58
w/ **ZS-Norm**	54.91	13.18	35.02	14.53	26.65	29.49	59.01	9.78	**43.90**	16.75
w/ **Penalty**	54.87	13.65	35.02	15.22	26.17	30.50	58.86	9.99	43.73	17.32
w/ **SaLS**	54.97	12.15	35.03	13.72	26.61	27.98	59.00	7.77	43.89	**16.74**

(a) ResNet-50

Method	V2		S		A		R		OOD Mean	
	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE	ACC	ECE
Zero-Shot [45]	60.83	2.40	46.15	4.80	47.80	8.36	73.99	3.51	57.19	4.77
TPT [49]	63.69	11.61	47.91	16.16	54.84	14.68	77.13	4.77	60.89	11.81
w/ **ZS-Norm**	63.56	11.14	47.76	15.71	54.57	13.60	77.08	4.60	60.74	11.26
w/ **Penalty**	63.53	11.59	47.94	16.00	54.55	13.87	77.07	4.14	60.77	11.40
w/ **SaLS**	63.69	10.48	47.91	15.70	54.84	13.52	77.13	3.75	**60.89**	**10.86**

(b) ViT-B/16

is still scope for improvement by adjusting the predictions logit range. More importantly, our approach **SaLS** can be directly applied to the logit predictions, not requiring pre-training the network, such as C-TPT, making of it an efficient *ready-to-use* solution.

A.7 Additional studies on Logit norm, range, and calibration

In this experiment, we analyze the impact of our contributions in calibration to the logit norm and range. We consider representative methods for few-shot Adapters and Prompt Learning, *i.e.*, TaskRes [60] and CoOp [66], respectively. Fig. 4, and 5 depict the comparison of logit norm/range with ECE for **ZS-Norm**, **Penalty**, and **SaLS** proposed calibration methods. As expected, after applying our method, ECE is reduced in most scenarios. It is worth mentioning that ECE improvements correlate with the decrease in the logit range. This is not the case of the logit norm, which either increases or remains constant. These observations correlate with our hypothesis in the main paper, and demonstrate that logit range plays a key role in calibration.

A.8 Comparison to other calibration methods

We further evaluate the performance of our simplest solution, **SaLS**, compared to several existing unsupervised calibration approaches. Our reasoning behind using these methods, i.e., L-Norm [56] and ECP [43], stems from the fact that they do not require labeled samples, in contrast to most existing methods (for

Table 9: Additional results for Test time prompt tuning calibration for ViT-B/16 backbone Performance of popular 11 few shot datasets with TPT for the proposed **ZS-Norm**, **Penalty**, and **SaLS**. Best result over TPT are highlighted in bold, and second underlined.

		Avg.	INet	CAL	PET	CAR	FLW	FOO	AIR	SUN	DTD	SAT	UCF
ACC	Zero-shot [45]	63.92	66.72	93.31	88.25	65.51	67.40	83.64	23.91	62.56	44.39	42.22	65.24
	TPT [49]	65.18	68.87	94.28	87.41	66.51	68.98	84.64	23.43	65.61	46.69	42.49	68.12
	w/ **ZS-Norm**	65.13	68.83	94.16	87.54	66.61	68.66	84.67	23.31	65.39	46.34	42.98	67.99
	w/ **Penalty**	**65.22**	68.86	94.08	86.64	66.72	68.82	84.43	23.07	65.55	46.28	44.99	67.94
	w/ **SaLS**	65.18	68.87	94.28	87.41	66.51	68.98	84.64	23.43	65.61	46.69	42.49	68.12
	C-TPT [59]	64.59	68.08	93.63	88.20	65.75	69.43	83.07	24.03	64.52	46.16	42.20	65.42
	w/ **ZS-Norm**	64.41	68.09	93.79	88.28	65.87	69.27	83.05	23.91	64.30	45.63	41.28	65.05
	w/ **Penalty**	**64.68**	68.00	93.39	88.06	65.81	69.27	83.16	24.39	64.59	45.69	43.96	65.11
	w/ **SaLS**	64.59	68.08	93.63	88.20	65.75	69.43	83.07	24.03	64.52	46.16	42.20	65.42
ECE	Zero-shot [45]	3.91	1.86	5.08	4.19	4.22	1.87	1.79	5.21	1.96	7.87	6.52	2.50
	TPT [49]	11.36	10.42	4.41	5.45	5.11	13.13	4.24	16.76	11.26	21.23	20.42	11.54
	w/ **ZS-Norm**	10.82	10.21	4.22	5.27	4.86	12.93	3.98	16.35	10.58	20.92	18.28	11.38
	w/ **Penalty**	**9.27**	10.23	4.26	3.44	3.63	12.33	3.25	15.77	10.85	19.62	8.53	10.07
	w/ **SaLS**	9.88	9.56	4.55	5.00	3.86	11.50	4.30	15.54	10.89	18.90	13.88	10.69
	C-TPT [59]	4.81	3.00	4.12	1.46	1.35	5.29	2.66	4.11	5.06	12.41	11.33	2.16
	w/ **ZS-Norm**	5.03	3.01	4.36	1.43	1.51	5.35	2.71	4.25	4.74	12.24	11.13	2.53
	w/ **Penalty**	4.61	3.09	3.76	1.60	1.20	5.46	2.59	4.25	4.87	13.71	6.81	1.86
	w/ **SaLS**	**4.48**	2.22	4.38	3.61	2.53	2.29	1.45	5.60	3.32	9.2	10.46	4.20

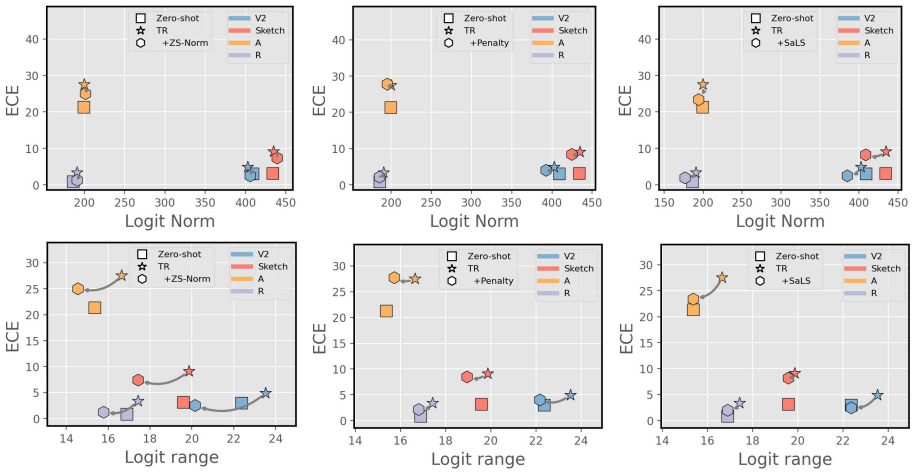

Fig. 4: Additional Logit studies for few-shot Adapters. Analysis of average Logit norm and range after improving the calibration of the Adapter model TaskRes [60] using the proposed logit range regularization methods for improved calibration, i.e., **ZS-Norm** (*left*), **Penalty** (*middle*) and **SaLS** (*right*).

example, Temperature Scaling (TS) needs a large validation set to fix the temperature value). These results, which are reported in Table ??, showcase that the proposed post-processing alternative brings important performance gains, in

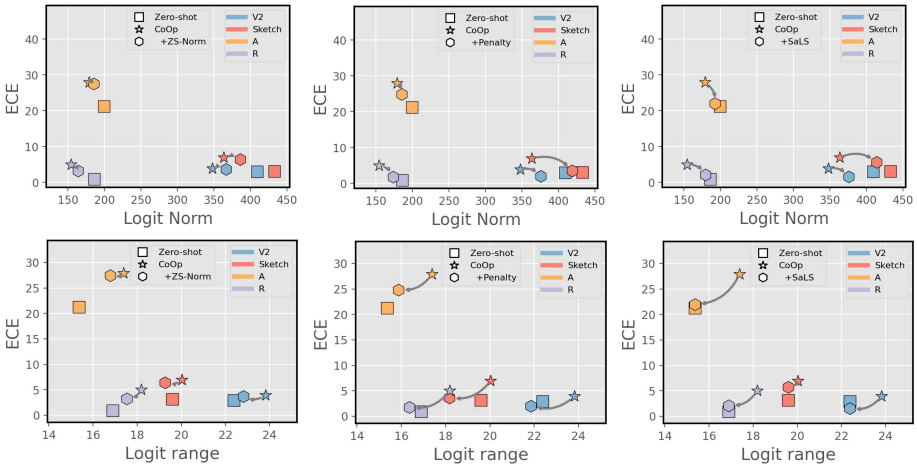

Fig. 5: Additional Logit studies for Prompt Learning. Analysis of average Logit norm and range after improving the calibration of the Prompt learning technique CoOp [66] using the proposed logit range regularization methods for improved calibration, i.e., **ZS-Norm** (*left*), **Penalty** (*middle*) and **SaLS** (*right*).

Table 10: Comparison to unsupervised calibration approaches. Average results are reported on ImageNet OOD.

	+SaLS		+L-Norm [56]		+ECP [43]	
	ACC	**ECE**	**ACC**	**ECE**	**ACC**	**ECE**
TaskRes$_{\text{CVPR'23}}$	41.18	9.03	42.04	14.40	41.37	9.50
CoOp$_{\text{IJCV'22}}$	40.86	7.82	40.82	18.66	41.81	23.76
TPT$_{\text{NeurIPS'22}}$	58.77	9.21	57.46	10.91	57.80	11.32

terms of calibration, without sacrificing discriminative power. This gap is particularly significant in prompt learning, where the proposed **SaLS** improves the ECE on CoOp by 11% and 16%, compared to L-Norm and ECP, respectively.

A.9 Reliability plots

Fig. 6, 7, and 8 depict the reliability plot of ZS, **ZS-Norm**, **Penalty** and **SaLS** for one from each of the setting of Adapters (Clip-Adapter), Prompt learning (CoOp), and Test time prompt tuning (TPT) for few representative cases in ImageNet OOD, and Few shot benchmarks respectively. From these plots, it could be noted that the density of the plots near the expected calibration curve is lower in our methods compared to the baselines, moving closer to ZS without compromising much the accuracy. Furthermore, the difference between the accuracy and the average confidence (in the bottom of the plots) is typically reduced when our approaches are integrated into the original methods, a sign that indicates that a model is better calibrated.

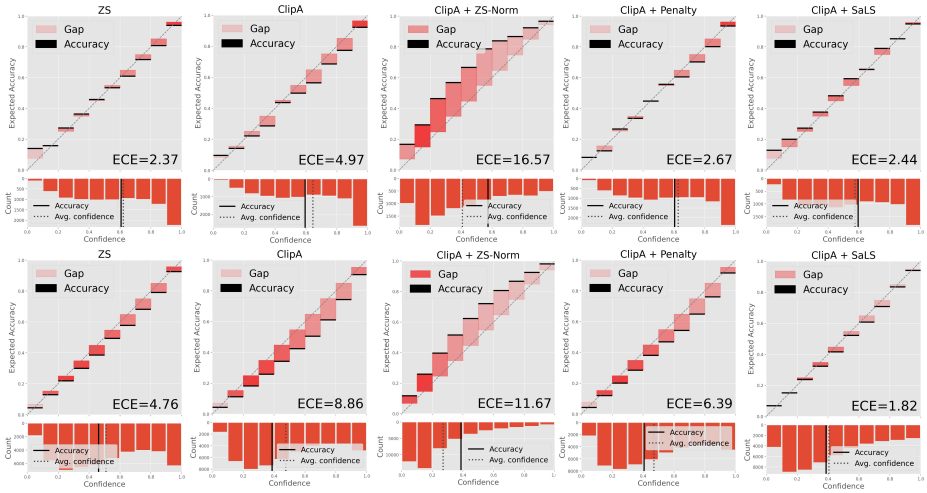

Fig. 6: Reliability plot for Adapter method Clip-Adapter with ImageNet Variants, ImageNetV2 (Top), and ImageNetSketch (Bottom)

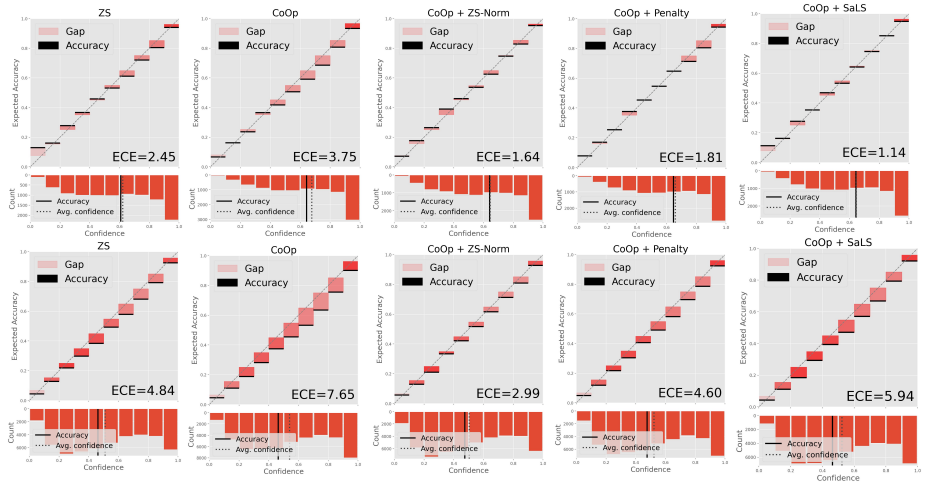

Fig. 7: Reliability plot for Prompt learning method CoOp with ImageNet Variants, ImageNetV2 (Top), and ImageNetSketch (Bottom)

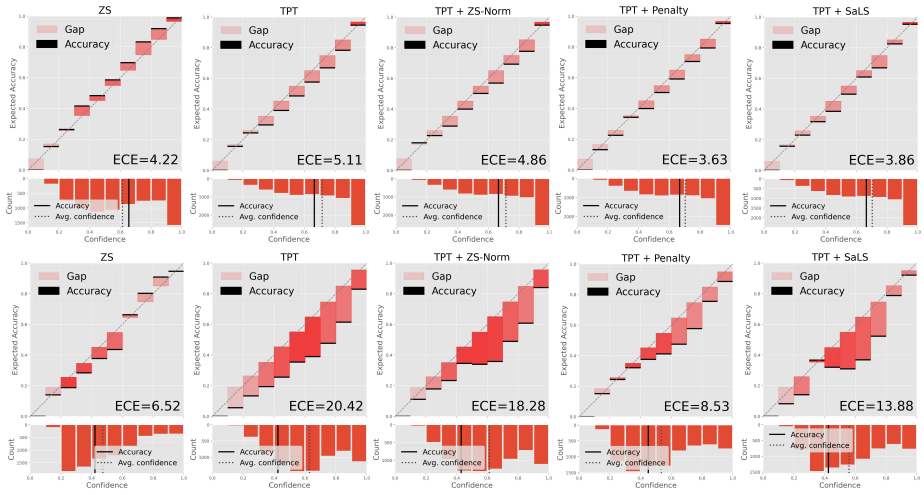

Fig. 8: TPT few shot benchmark reliability plot comparison for ViT-B/16 architecture. From Top to bottom: StanfordCars, EuroSAT