Towards Calibrating Prompt Tuning of Vision- Language Models

Supplementary Material

Method		Cal	Pets	Cars	Flow	Food	Air	SUN	DTD	Euro	UCF	Avg
Zero Shot	Acc.	97.20	91.30	63.60	71.80	90.10	27.70	69.40	53.00	57.00	71.00	69.50
	ECE	6.49	2.25	3.74	3.11	1.57	3.03	1.59	4.53	8.35	3.24	3.58
CoCoOp[24]												
CoCoOp [24]	Acc.	97.77	95.02	70.72	94.62	90.43	35.33	79.19	75.54	85.40	81.71	80.57
	ECE	1.43	3.21	6.89	7.85	0.86	5.42	3.78	3.88	8.09	3.78	4.52
ZS-Norm [13]	Acc.	97.83	95.22	70.65	95.00	90.63	36.03	79.70	76.35	82.53	81.90	80.58
	ECE	2.93	3.20	8.67	7.56	1.49	8.50	7.50	10.82	16.09	4.08	7.08
Penalty [13]	Acc.	97.83	94.98	69.95	92.43	90.72	34.35	79.34	71.49	69.57	80.49	78.12
	ECE	5.00	6.06	10.36	14.90	3.95	6.83	6.69	17.22	20.94	7.17	9.91
Ours	Acc.	97.93	94.69	69.42	94.11	90.51	34.25	78.92	74.77	84.70	82.34	80.16
	ECE	1.13	2.31	7.01	7.98	0.49	5.81	2.63	3.70	6.47	3.23	4.08
ProDA [23]												
ProDA [23]	Acc.	97.61	94.75	69.76	89.96	89.33	33.01	76.17	70.02	81.83	79.99	78.24
	ECE	1.06	1.67	3.86	6.07	0.86	3.52	6.66	10.25	3.73	2.56	4.02
ZS-Norm [13]	Acc.	97.55	94.37	69.77	89.62	89.50	33.03	76.46	71.33	82.00	79.33	78.30
	ECE	1.93	2.22	4.74	7.41	1.31	3.00	6.62	3.73	11.93	2.72	4.56
Penalty [13]	Acc.	97.35	94.61	69.32	89.14	90.36	32.17	76.94	59.80	63.86	78.07	75.16
	ECE	4.00	8.11	7.86	12.89	3.79	4.91	2.09	11.28	18.05	7.83	8.08
Ours	Acc.	97.20	94.31	69.50	87.88	90.02	32.99	76.96	72.91	82.70	80.63	78.51
	ECE	1.75	2.29	6.91	8.21	1.18	3.38	1.18	2.27	5.37	2.50	3.50
ProGrad [26]												
ProGrad [26]	Acc.	97.72	94.67	69.29	81.26	90.33	31.35	76.88	67.13	79.27	78.20	76.61
	ECE	3.53	3.83	6.84	6.82	1.65	2.60	3.70	6.38	12.24	3.92	5.15
ZS-Norm [13]	Acc.	97.87	94.36	69.27	82.91	90.45	32.35	77.92	70.37	75.57	78.70	76.98
	ECE	5.51	4.85	10.02	10.95	2.67	8.50	10.33	23.04	17.46	6.47	9.98
Penalty [13]	Acc.	97.81	94.21	69.02	84.14	90.47	32.65	77.33	57.29	67.21	75.65	74.58
	ECE	5.51	6.39	8.69	13.38	3.49	6.70	5.54	7.70	16.49	5.79	7.97
Ours	Acc.	97.55	94.45	68.95	82.62	90.29	31.57	77.03	68.52	79.78	79.04	76.98
	ECE	3.18	3.46	7.01	6.90	1.36	3.06	3.10	6.41	11.20	3.12	4.88
PromptSRC [12]												
PromptSRC [12]	Acc.	98.08	95.36	78.15	97.95	90.60	40.74	82.63	83.41	93.17	87.09	84.72
	ECE	2.31	2.64	8.65	5.15	1.17	5.26	2.75	2.56	9.27	2.84	4.26
ZS-Norm [13]	Acc.	98.21	95.43	77.55	97.50	90.78	40.78	82.62	81.56	48.78	85.76	79.90
	ECE	4.41	4.58	12.01	8.65	3.21	10.37	5.43	20.69	20.10	6.73	9.62
Penalty [13]	Acc.	98.01	95.69	77.45	97.82	90.36	40.89	82.29	81.84	48.38	85.91	79.86
	ECE	5.41	4.78	10.61	9.65	4.51	12.47	6.43	18.69	19.10	8.98	10.06
Ours	Acc.	98.30	95.57	79.11	98.39	90.67	42.50	82.77	83.83	95.05	86.87	85.31
	ECE	1.03	0.89	8.96	0.97	0.90	6.04	1.34	4.59	4.33	1.39	3.04

Table 1. **Accuracy and calibration performance on base classes across 10 fine-grained classification benchmarks**. We report top-1 accuracy (Acc) and Expected Calibration Error (ECE) for multiple prompt-tuning strategies and diverse calibration baselines. Higher Acc. indicates better classification performance, while lower ECE reflects better calibration.

Contents of Supplementary Material

In this supplementary material, we provide the following:

1. **Calibration analysis for base and novel classes across prompt learning methods** (Sec. 1)
2. **Robustness evaluation under natural distribution shifts** (Sec. 2)
3. **Additional results for ACE and MCE performance metrics** (Sec. 3)
4. **Hyperparameters details** (Sec. 6)
5. **Prompt templates and variations** (Sec. 7)
6. **Variance analysis** (Sec. 8)
7. **Results on Different Backbones**(Sec. 9)
8. **Decision Boundary Visualization**(Sec. 10)

1. Calibration Analysis for Base and Novel Classes Across Prompt Learning Methods

In the main paper, we provide the calibration results for three representative prompt learning methods: CoOp [25], KgCoOp [22], and MaPLe [11]. Here we provide the results for additional prompt learning methods including CoCoOp [24], ProDA [23], ProGrad [26], and PromptSRC [12] for both base and novel classes to demonstrate the broader applicability and effectiveness of our calibration approach.

Calibration Performance on Base Classes. Table 1 presents the accuracy and calibration performance on base classes across 10 fine-grained classification benchmarks. The results consistently demonstrate that our calibration approach achieves superior calibration performance (lower ECE) while maintaining competitive accuracy across all evaluated prompt learning methods. Notably, our method shows the most significant improvements with PromptSRC [12], reducing the average ECE from 4.26 to 3.04 without compromising clean accuracy. The consistent improvements across diverse prompt learning architectures validate the generalizability of our approach. **Calibration Performance on Novel Classess.** We further evaluate the calibration performance on novel classes to assess the generalization capability of our approach when dealing with unseen categories during training. Table 2 presents the accuracy and calibration performance on novel classes across the same 10 benchmarks. Our method demonstrates consistent calibration improvements across all prompt learning methods when evaluated on novel classes. The results show that our approach effectively generalizes to unseen categories, with particularly notable improvements in ECE reduction. For instance, with ProDA [23], our method reduces the average ECE from 9.03 to 3.42 while maintaining comparable clean accuracy (70.22 vs 71.38). Similarly, with CoCoOp [24], we achieve an ECE reduction from 5.55 to 3.86. These results highlight the robustness of our calibration approach in handling the challenging scenario of novel class prediction, where models are more prone to overconfidence due to limited training exposure.

2. Results on Natural Distribution Shifts

Here, we provide results on out-of-distribution datasets for the ImageNet-A [7], ImageNet-V2 [18], ImageNet-R [6], and ImageNet-S [20] datasets demonstrating the robustness of our calibration approach under natural distribution shifts.

Base Classes. Table 3 presents the accuracy and calibration performance on base classes across these 4 natural distribution shift datasets. Our method consistently outper-

Method		Calt	Pets	Cars	Flow	Food	Air	SUN	DTD	Euro	UCF	Avg
Zero Shot	Acc.	94.10	97.10	75.00	77.50	91.10	35.90	75.50	60.60	63.80	78.60	74.30
	ECE	1.60	3.42	3.31	4.91	1.83	6.55	3.48	6.86	9.12	5.52	4.43
CoCoOp[24]												
CoCoOp [24]	Acc.	94.51	97.69	73.26	72.27	91.02	33.47	76.54	57.65	63.14	74.89	73.44
	ECE	1.84	2.64	1.88	9.17	1.64	10.93	2.21	11.26	9.06	4.90	5.55
ZS-Norm [13]	Acc.	94.76	97.24	73.56	70.45	91.43	32.97	76.84	54.59	57.01	71.64	72.05
	ECE	2.63	2.87	2.11	9.39	2.16	7.21	3.99	3.91	9.24	4.36	4.79
Penalty [13]	Acc.	94.29	95.62	75.07	70.52	91.46	33.59	76.85	56.76	54.50	74.08	72.27
	ECE	2.22	5.11	5.69	5.80	4.22	5.30	3.93	11.13	11.44	3.66	5.85
DAC	Acc.	-	-	-	-	-	-	-	-	-	-	-
	ECE	3.65	2.43	2.21	7.74	1.64	9.03	1.09	7.47	13.49	2.70	5.15
Ours	Acc.	94.43	97.45	74.71	71.91	91.62	33.97	76.41	56.40	61.67	76.40	73.50
	ECE	1.51	2.58	3.22	6.36	0.96	8.85	0.87	4.77	6.68	2.76	3.86
ProDA [23]												
ProDA [23]	Acc.	93.99	96.90	73.16	72.51	90.64	31.35	65.02	53.99	51.86	72.76	70.22
	ECE	3.22	1.96	3.18	8.51	0.84	15.03	14.08	16.90	21.85	4.74	9.03
ZS-Norm [13]	Acc.	93.81	97.28	72.53	72.81	90.44	30.09	66.59	52.13	57.77	72.67	71.01
	ECE	2.36	2.42	2.06	8.34	0.94	10.76	12.12	7.65	8.75	4.41	5.98
Penalty [13]	Acc.	93.92	97.20	73.39	73.57	90.70	32.45	67.73	50.48	60.05	72.49	71.20
	ECE	1.53	6.14	3.76	4.36	3.47	7.82	2.51	4.96	14.47	3.86	5.29
DAC	Acc.	-	-	-	-	-	-	-	-	-	-	-
	ECE	4.87	4.72	3.28	6.32	0.70	7.40	1.06	5.68	3.33	4.14	4.15
Ours	Acc.	93.56	97.56	73.81	72.74	91.14	30.57	66.18	53.82	58.58	75.79	71.38
	ECE	1.48	3.25	2.77	5.12	1.01	6.78	1.90	4.60	4.91	2.35	3.42
ProGrad [26]												
ProGrad [26]	Acc.	94.76	97.32	74.85	75.29	91.06	34.43	75.42	56.44	61.98	78.74	74.03
	ECE	1.67	3.52	2.68	7.46	1.76	9.21	2.05	4.48	8.83	3.57	4.52
ZS-Norm [13]	Acc.	94.43	97.37	74.97	75.18	91.18	31.49	74.79	55.80	67.97	77.39	74.06
	ECE	1.80	5.11	5.32	3.73	2.68	3.79	7.10	12.79	12.83	4.83	6.00
Penalty [13]	Acc.	94.87	96.98	75.81	74.54	91.05	34.55	75.03	53.74	66.97	76.80	74.03
	ECE	1.90	5.54	5.07	4.80	3.08	5.31	3.08	4.96	14.86	5.05	5.37
DAC	Acc.	-	-	-	-	-	-	-	-	-	-	-
	ECE	1.97	3.31	2.29	5.04	1.85	10.46	1.32	3.49	6.90	2.42	3.91
Ours	Acc.	94.29	97.48	75.09	74.66	91.21	32.87	74.81	55.60	67.91	78.53	74.25
	ECE	1.03	3.26	1.98	5.14	1.47	9.34	2.32	3.30	6.06	3.06	3.70
PromptSRC [12]												
PromptSRC [12]	Acc.	94.21	97.31	75.58	77.28	91.51	29.73	78.79	61.03	74.72	77.86	75.80
	ECE	1.51	3.26	2.06	5.50	1.77	12.92	1.07	6.68	8.08	2.81	4.57
ZS-Norm [13]	Acc.	94.18	97.61	74.80	76.31	91.60	36.23	78.80	59.02	37.07	77.25	72.29
	ECE	2.06	4.30	3.63	5.86	3.66	4.64	3.57	12.29	12.56	3.95	5.65
Penalty [13]	Acc.	94.17	97.55	74.12	76.34	91.99	36.63	78.14	59.62	37.37	77.76	72.37
	ECE	3.06	5.40	4.63	4.86	4.98	4.94	4.53	11.29	10.66	4.65	5.90
DAC	Acc.	-	-	-	-	-	-	-	-	-	-	-
	ECE	1.58	2.98	2.39	5.03	1.55	8.55	0.79	5.50	7.24	2.46	3.81
Ours	Acc.	94.29	97.28	74.49	75.44	91.68	36.85	78.39	57.53	72.64	77.72	75.63
	ECE	1.14	1.19	2.39	5.44	0.72	9.26	0.77	6.81	8.37	1.89	3.80

Table 2. **Accuracy and calibration performance on novel classes across 10 fine-grained classification benchmarks**. We report top-1 accuracy (Acc) and Expected Calibration Error (ECE) for multiple prompt-tuning strategies and diverse calibration baselines. Higher Acc. indicates better classification performance, while lower ECE reflects better calibration.

Method		INet-V2	INet-S	INet-A	INet-R	Avg
MaPLe [11]						
MaPLe [11]	Acc.	67.35	53.35	68.31	85.22	68.56
	ECE	3.14	3.94	2.52	3.14	3.19
ZS-Norm [13]	Acc.	66.15	53.24	68.41	85.17	68.49
	ECE	3.41	4.12	6.98	7.13	5.41
Penalty [13]	Acc.	66.72	53.04	68.61	85.17	68.39
	ECE	3.61	4.62	7.48	7.33	5.76
Ours	Acc.	67.19	53.15	67.86	85.28	68.37
	ECE	3.09	3.91	2.21	2.05	2.82

Table 3. **Accuracy and calibration performance on base classes across 4 natural distribution shift datasets**. We report top-1 accuracy (Acc) and Expected Calibration Error (ECE) for MaPLe [11]. Higher Acc. indicates better classification performance, while lower ECE reflects better calibration.

forms baseline calibration approaches across all datasets. Notably, our approach achieves superior calibration with an average ECE of 2.82 compared to the vanilla MaPLe baseline (3.19), ZS-Norm (5.41), and Penalty (5.76). The improvements are particularly pronounced on challenging datasets like ImageNet-A and ImageNet-R, where our method reduces ECE from 2.52 to 2.21 and from 3.14 to 2.05, respectively, while maintaining competitive accuracy.

Method		INet-V2	INet-S	INet-A	INet-R	Avg
MaPLe [11]						
MaPLe [11]	Acc.	67.36	53.36	68.31	85.22	68.56
	ECE	3.16	3.73	2.52	3.14	3.14
ZS-Norm [13]	Acc.	66.75	53.34	68.60	85.16	68.46
	ECE	3.54	4.01	6.47	7.21	5.31
Penalty [13]	Acc.	66.75	53.11	68.60	85.16	68.41
	ECE	3.64	4.11	7.47	7.32	5.64
Ours	Acc.	67.20	53.21	67.86	85.28	68.39
	ECE	3.11	3.63	2.20	2.05	2.75

Table 4. **Accuracy and calibration performance on novel classes across 4 natural distribution shift datasets**. We report top-1 accuracy (Acc) and Expected Calibration Error (ECE) for MaPLe [11]. Acc. indicates better classification performance, while lower ECE reflects better calibration.

Novel Classes. Table 4 shows the corresponding results on novel classes under distribution shift. The consistent performance across both base and novel classes demonstrates the generalization capability of our calibration approach. Our method achieves an average ECE of 2.75 on novel classes, significantly outperforming ZS-Norm (5.31) and Penalty (5.64) baselines. The robustness across different types of distribution shifts, including adversarial examples (ImageNet-A), renditions (ImageNet-R), and sketch-like images (ImageNet-S), validates that our approach addresses fundamental calibration issues rather than dataset-specific artifacts.

These results are particularly important for real-world deployment scenarios where models encounter data that differs from the training distribution. The consistent calibration improvements across diverse distribution shifts demonstrate that our method provides reliable confidence estimates even under challenging out-of-distribution conditions.

3. Additional Results: ACE and MCE Performance Metrics

In the main paper, we evaluate classification performance using top-1 accuracy and model calibration using Expected Calibration Error (ECE). Here we provide comprehensive results for additional calibration metrics including Adaptive Calibration Error (ACE) [16] and Maximum Calibration Error (MCE) [14] to further validate the effectiveness of our approach.

Table 5 presents the MCE and ACE results on **base classes** across 10 fine-grained classification benchmarks. Our method demonstrates consistent improvements across both metrics for all evaluated prompt learning methods. For CoOp, our approach reduces the average MCE from 2.40 to

Method		Calt	Pets	Cars	Flow	Food	Air	SUN	DTD	Euro	UCF	Avg
CoOp[25]												
CoOp [25]	MCE	0.24	0.31	0.91	2.46	1.72	4.77	3.99	5.68	0.60	3.34	2.40
	ACE	0.44	0.62	3.65	4.67	3.65	25.70	8.11	12.01	1.95	6.41	6.72
ZS-Norm [13]	MCE	1.40	2.17	2.20	2.76	0.90	4.34	0.82	14.76	11.82	0.87	4.20
	ACE	4.32	7.70	11.26	11.12	3.14	13.05	4.26	49.53	36.04	3.36	14.38
Penalty [13]	MCE	1.62	1.85	2.13	2.50	1.38	2.23	0.94	4.30	11.34	1.63	2.99
	ACE	4.75	6.41	10.01	9.36	5.98	8.61	4.61	21.48	20.86	7.09	9.92
Ours	MCE	0.33	1.01	1.87	1.68	0.12	1.21	0.31	0.56	1.67	0.22	0.90
	ACE	0.98	2.10	7.55	4.94	0.21	2.40	1.30	2.01	5.12	1.15	2.78
MaPLe [11]												
MaPLe [11]	MCE	0.51	0.60	1.68	1.29	0.34	0.91	0.19	1.04	0.74	0.56	0.79
	ACE	2.14	1.19	6.91	3.21	0.73	2.95	1.17	3.71	2.94	1.49	2.64
ZS-Norm [13]	MCE	1.28	1.01	3.79	1.45	2.25	1.95	2.27	2.80	1.24	0.61	1.87
	ACE	5.28	3.21	20.73	7.21	11.17	6.98	8.59	12.37	6.91	3.40	8.59
Penalty [13]	MCE	1.88	1.85	2.30	3.15	1.11	2.27	1.24	3.69	0.61	1.45	1.96
	ACE	5.68	6.29	11.17	11.97	3.71	8.72	6.81	20.23	3.10	7.61	8.53
Ours	MCE	0.62	0.34	1.20	0.87	1.47	0.62	1.11	1.50	0.60	0.94	0.93
	ACE	1.62	0.80	3.98	2.18	3.66	0.94	4.38	7.55	1.54	1.34	2.80
KGCoOp [22]												
KGCoOp [22]	MCE	1.14	1.17	2.23	2.75	0.62	1.17	0.97	1.61	3.31	1.25	1.62
	ACE	2.56	2.95	10.14	11.95	1.59	2.95	4.91	8.39	11.90	4.59	6.19
ZS-Norm [13]	MCE	1.31	1.13	2.28	3.02	0.64	3.12	1.34	3.85	4.06	1.12	2.19
	ACE	2.98	3.06	10.58	13.00	1.69	9.59	6.51	20.40	15.58	5.75	8.91
Penalty [13]	MCE	1.27	1.40	2.10	3.03	0.81	1.67	1.20	2.87	3.30	1.33	1.90
	ACE	4.20	4.61	9.96	12.53	2.44	6.41	5.92	10.66	13.14	6.04	7.59
Ours	MCE	0.86	1.05	1.12	2.01	0.55	1.94	0.81	1.12	3.61	0.86	1.39
	ACE	1.83	2.83	8.1	11.21	1.42	5.11	4.15	7.21	12.65	3.95	5.85

Table 5. **Calibration performance on base classes across 10 fine-grained classification benchmarks**. We report Maximum Calibration Error (MCE) and Adaptive Calibration Error (ACE) for multiple prompt-tuning strategies and diverse calibration baselines. Lower MCE and ACE reflects better calibration.

Method		Calt	Pets	Cars	Flow	Food	Air	SUN	DTD	Euro	UCF	Avg
CoOp[25]												
CoOp [25]	MCE	2.61	0.64	2.38	5.48	1.44	5.01	4.35	7.06	4.41	6.11	3.95
	ACE	3.47	1.67	12.45	18.33	3.84	28.41	13.92	26.88	12.73	19.17	14.09
DAC [21]	MCE	1.50	0.66	1.21	2.08	0.49	3.57	1.03	2.24	3.03	1.70	1.76
	ACE	2.60	1.70	5.17	10.18	1.75	17.27	4.00	10.51	8.58	8.63	7.04
ZS-Norm [13]	MCE	1.46	2.06	0.75	1.08	0.82	2.24	0.55	9.32	11.82	1.14	3.12
	ACE	2.59	7.9	3.01	5.93	3.3	9.86	2.27	21.97	37.04	3.91	9.78
Penalty [13]	MCE	0.92	2.1	0.68	1.16	1.00	2.46	0.76	0.9	7.54	1.36	1.89
	ACE	2.23	7.32	2.69	5.46	4.69	7.44	2.91	4.35	14.94	4.51	5.65
Ours	MCE	1.05	1.26	0.55	0.98	0.25	2.9	0.64	1.97	3.9	1.49	1.42
	ACE	2.3	2.92	2.02	3.67	0.91	10.47	2.99	9.6	11.24	5.04	4.83
MaPLe [11]												
MaPLe [11]	MCE	0.55	1.04	0.56	5.02	0.37	1.48	0.65	3.50	1.66	0.66	1.55
	ACE	1.26	2.44	2.83	12.67	1.12	7.27	2.49	14.90	7.90	2.81	5.57
DAC [21]	MCE	0.39	1.06	0.77	4.42	0.50	2.26	0.40	1.95	2.55	0.66	1.50
	ACE	1.19	2.44	2.57	11.28	1.48	8.90	1.37	8.24	9.12	2.32	4.89
ZS-Norm [13]	MCE	0.84	1.23	2.13	1.76	1.32	1.92	1.07	1.40	0.91	7.51	2.01
	ACE	1.62	4.15	9.46	7.77	3.83	7.85	3.77	6.11	4.84	13.73	6.31
Penalty [13]	MCE	0.84	1.72	1.40	1.32	1.11	1.07	0.91	2.73	7.11	1.76	2.00
	ACE	1.92	7.60	6.41	4.83	4.05	3.47	4.74	10.46	16.73	8.67	6.89
Ours	MCE	0.44	0.52	1.61	1.08	1.70	0.88	2.66	0.55	0.14	0.87	1.05
	ACE	0.32	0.95	6.38	3.13	11.27	2.07	8.49	2.48	0.58	4.74	4.04
KGCoOp [22]												
KGCoOp [22]	MCE	0.53	1.16	0.9	1.06	0.74	1.78	0.39	1.22	3.37	0.69	1.18
	ACE	1.22	3.26	3.36	5.45	2.01	5.86	1.83	5.02	8.69	2.59	3.93
DAC [21]	MCE	0.62	1.21	0.85	1.37	0.64	3.02	0.42	1.33	1.75	0.77	1.20
	ACE	1.56	3.02	3.11	6.61	1.93	11.74	1.82	7.26	6.63	2.70	4.64
ZS-Norm [13]	MCE	0.58	1.17	0.94	1.06	0.75	2.97	0.59	1.90	1.52	0.97	1.25
	ACE	1.33	3.30	3.86	5.31	2.10	8.39	3.30	5.95	6.51	3.81	4.39
Penalty [13]	MCE	0.56	1.43	0.99	1.38	0.87	1.34	0.58	1.68	3.72	1.28	1.38
	ACE	1.19	3.51	3.89	5.14	2.64	5.00	3.35	5.63	13.86	4.23	4.84
Ours	MCE	0.51	1.21	0.85	0.98	0.68	2.78	0.47	1.97	1.35	0.91	1.17
	ACE	1.1	3.43	3.68	4.91	1.92	7.85	2.01	4.11	4.32	3.15	3.65

Table 6. **Calibration performance on novel classes across 10 fine-grained classification benchmarks**. We report Maximum Calibration Error (MCE) and Adaptive Calibration Error (ACE) for multiple prompt-tuning strategies and diverse calibration baselines. Lower MCE and ACE reflects better calibration.

0.90 and ACE from 6.72 to 2.78, representing substantial calibration improvements. Similarly, with KGCoOp, we achieve reductions in MCE from 1.62 to 1.39 and ACE from 6.19 to 5.85. These results are particularly noteworthy as MCE captures the worst-case calibration error, indicating that our method not only improves average calibration but also reduces extreme miscalibration cases.

Table 6 shows the results on **novel classes**. The improvements are consistent with the base class results, demonstrating the generalization capability of our calibration approach. For KGCoOp on novel classes, our method maintains similar MCE performance (1.18 vs 1.17) while slightly improving ACE from 3.93 to 3.65. The robustness across different calibration metrics validates that our approach addresses fundamental calibration issues rather than optimizing for specific metrics.

The consistent improvements across ECE, MCE, and ACE metrics provide strong evidence that our calibration method effectively reduces both average and worst-case calibration errors, making it suitable for deployment in safety-critical applications.

4. Medical Image Analysis

Furthermore, we evaluated our proposed solution on four Med-VLMs: PLIP [8], QuiltNet [9], using three downstream datasets: KatherColon (Kather) [10], PanNuke [5], and DigestPath [3]. We compared the proposed method with other calibration methods, such as CE, MbLs, ZS-Norm, and penalty.

Table 7 shows that the proposed method consistently yields the lowest ECE values when compared with Vanilla Cross Entropy Loss(CE). At the same time, our proposed method gives the lowest reduced Overall Average ECE value compared to other baselines, yielding 7.09 while maintaining the stable accuracy.

5. Loss Component Analyses

Fig.1 shows how $\mathcal{L}_{\text{mom}}$ loss works directly with $\mathcal{L}_{\text{CE}}$ on both Base and Novel classes without $\mathcal{L}_{\text{margin}}$ loss. $\mathcal{L}_{\text{mom}}$ reduces the underconfidence in base classes marginally without any inter-class logit separation. However, it reduces the overconfidence issue alone(without having margin loss) in Novel classes by preserving CLIP's semantic geometry, maintaining relative class structure. As you can see in Fig.2 of the main paper, when incorporating $\mathcal{L}_{\text{margin}}$ with $\mathcal{L}_{\text{mom}}$, it helps to reduce the miscalibration issue further more in both Base and Novel classes by having inter-class separation and maintaining relative class structure geometry.

6. Hyperparameters Details

For all experiments, we use CLIP (ViT-B/16) [17] as the pre-trained vision-language model. Prompt-tuning is conducted in a few-shot setting with 16 samples per class, using a learning rate of 0.005 and a batch size of 8. For each baseline

Model →	PLIP						QuiltNet						Average	
Dataset →	Kather		PanNuke		DigestPath		Kather		PanNuke		DigestPath		All	
Loss ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	Acc ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓
Cross Entropy-based Losses														
Cross Entropy Loss$_{PL}$	83.91	5.92	66.70	17.82	82.87	9.50	87.97	2.49	69.82	19.70	81.59	11.27	78.81	11.12
MbLS$_{PL}$	84.39	3.57	66.70	17.82	82.76	9.53	84.76	3.48	65.54	23.05	82.58	10.90	77.79	11.39
ZS-Norm$_{PL}$	85.63	<u>3.07</u>	71.52	17.22	82.84	7.31	91.91	**0.87**	69.55	19.18	84.78	6.37	81.04	9.00
Penalty$_{PL}$	86.48	3.90	70.49	**3.11**	69.61	**2.40**	89.29	12.28	59.88	**3.41**	79.23	17.53	75.83	7.11
Ours	87.98	**1.31**	65.72	12.40	84.84	4.71	88.70	<u>1.45</u>	68.31	16.17	83.46	6.49	79.83	7.09

Table 7. Comparison of proposed method (with baseline methods using Cross Entropy (CE). Accuracy (ACC, %) and Expected Calibration Error (ECE, %) are shown for PLIP and QuiltNet on histopathology datasets (Kather, PanNuke, DigestPath). Best results are in **bold**, second-best <u>underlined</u>.

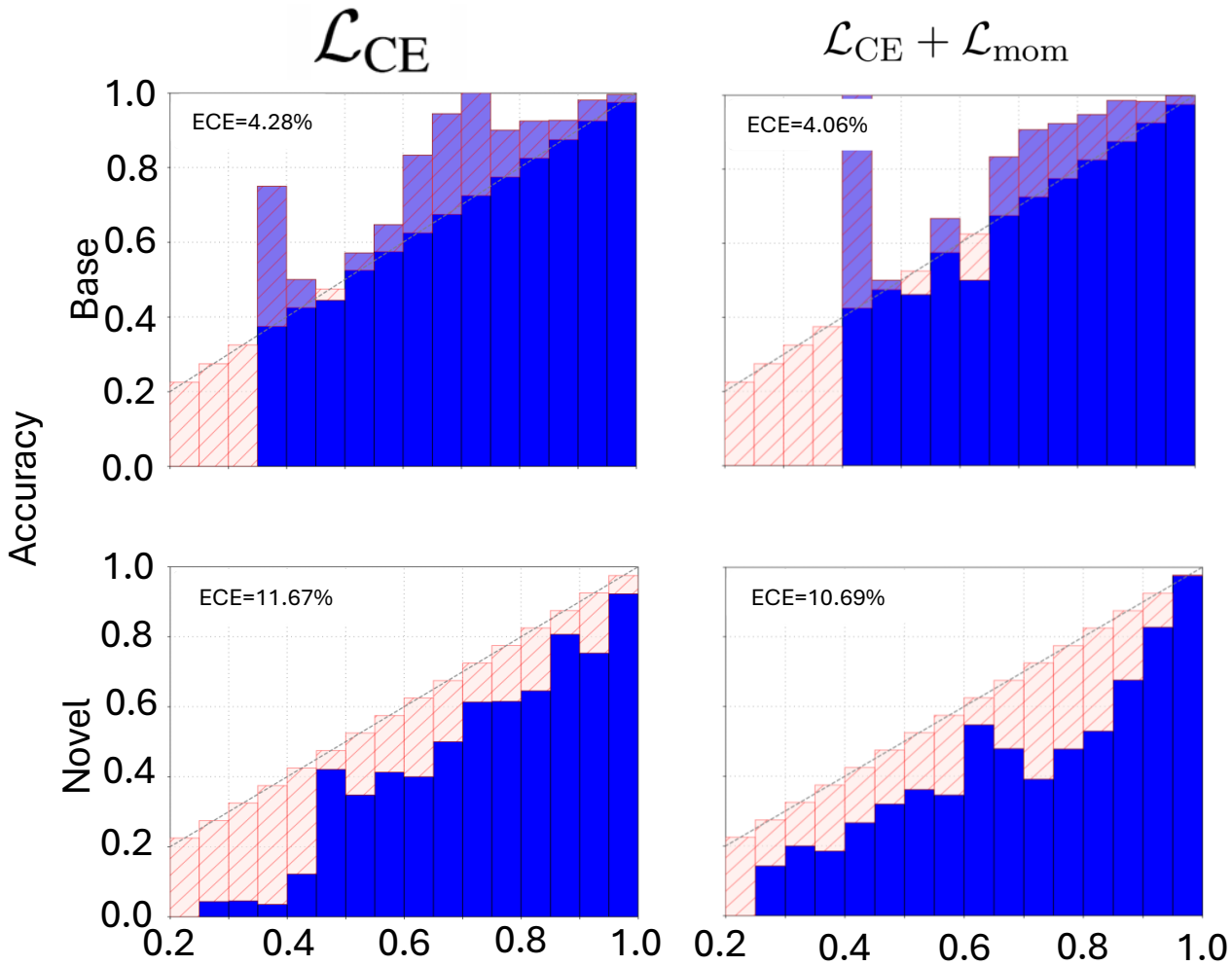


Figure 1. $\mathcal{L}_{\text{mom}}$ loss analyses on both Base and Novel Classes

method, we adopt its official implementation and follow the recommended hyperparameter settings from the original papers. All experiments are performed on an NVIDIA RTX A6000 GPU with 48GB memory.

For our proposed calibration method, we use the following hyperparameters across all experiments: $\lambda_{\text{Margin}} = 1.0$

	ACC			ECE		
(α, β)	Base	Novel	Avg	Base	Novel	Avg
$(0.1, 0.01)$	89.65	82.38	86.02	1.99	5.66	3.83
$(0.2, 0.01)$	89.72	79.66	84.69	3.46	8.53	5.99
$(0.3, 0.01)$	89.28	80.72	85.00	4.46	9.50	6.98
$(0.1, 0.05)$	89.12	79.99	84.56	3.95	6.26	5.10
$(0.2, 0.05)$	83.39	81.49	82.44	4.10	8.20	6.15
$(0.3, 0.05)$	89.75	80.52	85.14	2.01	8.01	5.01

Table 8. Hyperparameter search for α (0.1–0.3) and β (0.01 and 0.05), with results averaged across Caltech [4], Food101 [1], and DTD [2], reported on both base and novel classes.

	ACC			ECE		
λ	Base	Novel	Avg	Base	Novel	Avg
1	89.44	81.51	85.48	3.47	4.54	4.01
3	89.17	82.14	85.65	2.11	3.94	3.03
5	89.71	82.64	86.18	1.80	3.22	2.51
8	89.93	82.44	86.19	2.99	5.24	4.12
10	89.15	82.00	85.58	3.64	5.21	4.43

Table 9. Hyperparameter search for λ_{mom} over values 1–10, with results averaged across Caltech [4], Food101 [1], and DTD [2], reported on both base and novel classes.

controls the strength of the margin-based regularization, $\alpha = 0.1$ balances the average marin, $\beta = 0.01$ is the weight for the variance loss, and $\lambda_{\text{mom}} = 5.0$ controls the local moment matching regularization. Table8 and 9 show how we choose these values. These hyperparameters were fixed across all datasets and prompt learning methods to ensure fair comparison. We conduct 3 random seeds for each experiment and report the average results.

7. Prompt Templates and Variations

In the main paper, Figure 4b presents our method's robustness to different prompt initialization. The following prompt templates were evaluated to assess initialization robustness: "`a nice image of a {}`", "`an example of a {}`", "`a picture of a {}`", and "`a photo of the cool {}`". These templates represent different stylistic and semantic variations commonly used in prompt learning literature. This robustness is particularly valuable in practical deployment scenarios where optimal prompt initialization may not be known in advance.

8. Variance Analysis

To assess the statistical robustness of our approach, we evaluate the variance in performance across 3 random seeds for both accuracy and calibration metrics. Table 10 presents the variance results across 9 novel classes of fine-grained classification benchmarks for CoCoOp and KGCoOp methods. Our approach demonstrates superior stability with consistently lower variance in both accuracy and ECE compared to baseline calibration methods. For CoCoOp, our method achieves significantly lower average variance in accuracy (1.25 vs 1.55) and ECE (0.55 vs 2.53) compared to the vanilla baseline. Similarly, with KGCoOp, we maintain competitive variance performance with average accuracy variance of 0.44 compared to the baseline's 0.57, while substantially reducing ECE variance from 0.58 to 0.21. The reduced variance in calibration error is particularly noteworthy as it indicates that our method provides more consistent and reliable confidence estimates across different experimental runs, which is crucial for deployment in safety-critical applications.

Method		Calt	Pets	Cars	Flow	Food	Air	SUN	DTD	Euro	**Avg**
CoCoOp[25]											
CoCoOp [25]	Var. Acc	0.81	0.08	2.65	1.53	4.41	0.01	3.61	0.19	0.69	1.55
	Var. ECE	0.36	0.03	3.09	8.82	3.88	0.01	6.25	0.10	0.21	2.53
ZS-Norm [13]	Var. Acc	1.02	0.24	4.60	4.62	0.58	0.12	1.29	0.19	0.33	1.44
	Var. ECE	1.69	0.12	0.88	6.66	4.45	0.10	1.04	0.23	5.11	2.25
Penalty [13]	Var. Acc	0.12	0.09	0.11	0.32	0.59	5.81	2.25	0.09	0.08	1.05
	Var. ECE	0.08	0.02	0.20	0.04	0.75	1.00	2.50	0.19	0.02	0.53
Ours	Var. Acc	0.05	0.01	7.72	1.21	1.96	0.07	0.01	0.18	0.01	1.25
	Var. ECE	0.01	0.03	2.43	0.01	1.46	0.00	0.73	0.21	0.06	0.55
KGCoOp [22]											
KGCoOp [22]	Var. Acc	0.03	0.00	2.19	1.21	0.58	0.01	0.85	0.15	0.08	0.57
	Var. ECE	0.04	0.00	4.00	0.25	0.72	0.00	0.01	0.01	0.17	0.58
ZS-Norm [13]	Var. Acc	0.05	0.00	7.29	0.92	0.50	0.01	3.35	0.35	0.00	1.39
	Var. ECE	0.04	0.00	2.31	0.07	0.01	0.01	1.80	0.00	0.04	0.48
Penalty [13]	Var. Acc	0.01	0.00	1.66	0.16	1.23	0.01	1.04	0.11	0.15	0.49
	Var. ECE	0.08	0.06	0.38	0.18	1.19	0.01	2.37	0.07	0.06	0.49
Ours	Var. Acc	0.05	0.00	2.31	0.17	0.96	0.02	0.00	0.40	0.04	0.44
	Var. ECE	0.02	0.00	0.16	0.74	0.44	0.04	0.02	0.10	0.36	0.21

Table 10. **Variance across 3 random seeds for 9 novel classes of fine-grained classification benchmarks**.

9. Results on Different Backbones

To evaluate the adaptability of our method, we conduct experiments on CoOp [25] with different backbones, namely RN-50 and ViT-B/32. The Tables11 and 12results show that our approach consistently outperforms existing methods across both backbones, while also maintaining improvements in accuracy. In base classes, for RN-50, our method achieves an average ECE of 3.46 compared to 4.04 for the vanilla baseline. Similarly, for ViT-B/32, our method attains an ECE of 2.87, outperforming the vanilla baseline at 3.15. For novel classes, our method achieves the second-lowest average ECE of 5.46 on RN-50, with ZS-Norm [13] performing slightly better at 5.23. In contrast, on ViT-B/32, our method achieves the lowest ECE of 5.82, surpassing all other approaches.

10. Decision Boundary Visualization

Figure 2 shows that the Text Momentum-Matching loss better preserves the geometric structure of CLIP's pretrained embedding space by aligning the statistical moments of tuned and frozen text embeddings, compared to ℓ_1 align-

Method		Calt	Pets	Cars	Flow	Food	Air	SUN	DTD	Euro	UCF	Avg
CoOp-RN50[25]												
CoOp-RN50 [25]	Acc.	95.22	90.5	70.22	95.25	82.54	29.95	76.37	74.85	81.32	80.4	76.99
	ECE	1.27	2.32	6.15	4.22	1.15	2.62	2.81	8.65	9.34	1.82	4.04
ZS-Norm [13]	Acc.	95.55	90.96	69.89	95.19	82.71	25.89	76.67	74.27	89.57	79.58	78.03
	ECE	4.04	5.63	10.52	9.92	3.4	17.28	4.35	37.15	35.76	7.18	13.52
Penalty [13]	Acc.	96.31	92.1	68.58	94.65	83.66	26.63	74.57	68.48	47.61	77.9	78.12
	ECE	6.21	8.69	11.85	11.34	5.29	6.4	5.68	18.93	22.61	9.95	10.71
Ours	Acc.	95.44	91.08	69.90	95.22	82.77	29.01	77.07	75.42	89.79	79.96	78.57
	ECE	1.37	2.21	8.65	4.74	0.84	3.32	1.23	5.87	4.12	2.21	3.46
CoOp–ViT-B/32 [25]												
CoOp–ViT-B/32 [25]	Acc.	97.05	92.71	73.13	95.28	84.93	31.87	79.16	77.55	91.10	82.83	79.88
	ECE	1.17	2.45	4.86	4.18	1.06	3.12	2.5	6.15	3.91	2.09	3.15
ZS-Norm [13]	Acc.	97.18	92.08	72.86	94.81	85.12	32.19	79.64	76.23	90.09	82.76	80.25
	ECE	1.98	7.96	9.70	7.91	6.00	12.97	3.48	38.62	39.31	5.25	13.32
Penalty [13]	Acc.	96.73	93.22	73.33	94.75	86.05	29.99	78.95	69.64	53.42	5.94	75.78
	ECE	4.88	6.66	10.54	10.05	4.45	4.82	5.42	22.00	22.15	5.94	9.69
Ours	Acc.	97.46	93.14	73.78	95.09	85.12	31.83	79.58	79.01	91.09	82.87	80.09
	ECE	0.92	1.68	6.77	4.53	0.79	2.91	0.95	4.77	3.52	1.87	2.87

Table 11. **Accuracy and calibration performance on base classes across 10 fine-grained classification benchmarks using RN-50 and ViT-B/32**. We report top-1 accuracy (Acc) and Expected Calibration Error (ECE) for CoOp[25], a prompt-tuning strategy, evaluated with different backbones.

Method		Calt	Pets	Cars	Flow	Food	Air	SUN	DTD	Euro	UCF	Avg
CoOp-RN50[25]												
CoOp-RN50 [25]	Acc.	87.27	92.11	57.84	61.87	82.55	18.72	64.46	41.67	34.15	55.07	11.38
	ECE	3.57	2.10	8.08	10.24	0.86	18.75	9.07	25.47	22.18	13.47	11.38
ZS-Norm [13]	Acc.	88.39	89.97	57.86	59.93	81.52	20.34	65.77	35.79	40.08	57.37	59.70
	ECE	4.30	3.09	2.43	5.99	6.33	4.37	2.64	10.97	9.24	2.98	5.23
Penalty [13]	Acc.	88.75	93.38	61.37	83.63	20.22	67.45	45.21	31.57	58.16	3.01	61.05
	ECE	3.05	7.94	1.7	4.81	8.24	10.02	2.92	9.96	11.67	3.01	6.33
Ours	Acc.	87.52	93.08	59.88	61.47	82.50	21.66	64.94	39.05	42.75	53.94	60.68
	ECE	2.82	2.68	3.43	4.13	1.72	5.36	4.53	8.83	10.93	10.16	5.46
CoOp–ViT-B/32 [25]												
CoOp–ViT-B/32 [25]	Acc.	92.25	94.00	60.04	60.31	85.16	22.12	68.98	47.95	56.47	63.57	64.63
	ECE	3.29	2.30	8.53	13.28	1.33	16.75	8.57	19.74	17.18	10.32	10.13
ZS-Norm [13]	Acc.	91.56	93.46	59.01	54.09	84.50	21.64	70.14	43.44	52.04	65.98	63.59
	ECE	2.62	7.94	4.05	10.94	6.86	5.59	1.34	17.82	14.8	3.13	7.51
Penalty [13]	Acc.	92.39	96.12	59.83	53.78	86.58	23.90	70.52	45.45	44.66	61.77	63.560
	ECE	4.41	6.93	2.66	7.31	5.23	8.63	2.19	8.77	10.71	4.22	6.11
Ours	Acc.	91.41	93.48	61.10	61.65	84.22	23.08	70.59	50.23	56.55	64.02	65.63
	ECE	2.23	2.15	4.33	5.92	0.74	9.90	3.72	14.30	9.14	5.80	5.82

Table 12. **Accuracy and calibration performance on novel classes across 10 fine-grained classification benchmarks using RN-50 and ViT-B/32**. We report top-1 accuracy (Acc) and Expected Calibration Error (ECE) for CoOp[25], a prompt-tuning strategy, evaluated with different backbones.

Method	Total-Time-Per Batch	Peak GPU Memory
MaPLe [11]	15.91(seconds)	1.75 GB
Ours	15.83(seconds)	1.75 GB

Table 13. Training Time taken per epoch (in seconds) and peak GPU memory usage for 16-shot experiment on Flower [15] datasets

ment or Orthogonality-based class-wise dispersion [19] on novel classes.

11. Computational Analyses

To analyse the computational and memory overhead, we measured training time and peak GPU memory on Flower101[15] 16-shot with ViT-B/16, using the original MaPLe[11] configuration as baseline. Compared to MaPLe, our method has less training time and the same GPU usage.

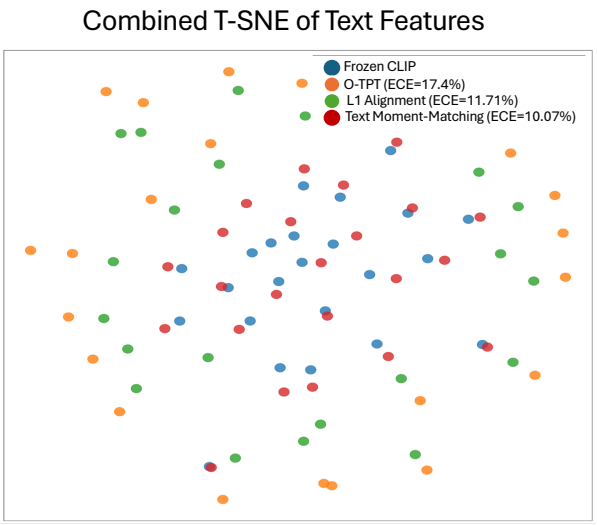
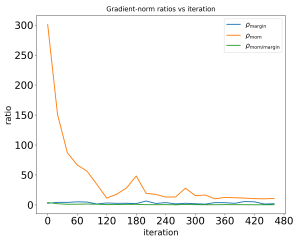


Figure 2. T-SNE visualization of Text Features

Table 14. Grad-norm ratio stats.

metric	median	q10	q90
ρ_{margin}	2.65	1.54	5.07
ρ_{mom}	16.92	10.50	80.84
$\rho_{\text{mom/margin}}$	0.38	0.18	1.30
$\rho_{\text{margin/CE}}$	0.28	0.18	1.42
$\rho_{\text{mom/CE}}$	0.17	0.05	0.51

Figure 3. Grad-norm ratios.



12. Regularizer in scale:

In Fig. 3 and Tab. 14, we log *gradient-norm ratios* to check optimisation-scale dominance. For the margin regularizer $L_{\text{Margin}} = L_{\text{mean}} + L_{\text{var}}$ with $L_{\text{mean}} = -\alpha\,\mathbb{E}[m]$ and $L_{\text{var}} = \beta\,\text{Var}(m)$, and measure $\rho_{\text{margin}} = \frac{\|\nabla_\theta L_{\text{mean}}\|}{\|\nabla_\theta L_{\text{var}}\|+\varepsilon}$, which is bounded (median 2.65; 10–90% [1.54, 5.07]), indicating comparable contributions of the linear mean and quadratic variance terms. For moment matching, with $L_\mu = \|\Delta\mu\|_2^2$ and $L_\Sigma = \|\Delta\Sigma\|_F^2$, we log $\rho_{\text{mom}} = \frac{\|\nabla_\theta(\lambda_{\text{mom}} L_\mu)\|}{\|\nabla_\theta(\lambda_{\text{mom}} L_\Sigma)\|+\varepsilon}$, which after a brief transient stabilises and shows no covariance-term dominance (median 16.92; 10–90% [10.5, 80.84]). Across regularizers, $\rho_{\text{mom/margin}} = \frac{\|\nabla_\theta L_{\text{mom}}\|}{\|\nabla_\theta L_{\text{margin}}\|+\varepsilon}$ stays near a constant scale (median 0.38; 10–90% [0.18, 1.30]), and both remain non-trivial yet typically subdominant to CE: $\rho_{\text{margin/CE}} = \frac{\|\nabla_\theta L_{\text{margin}}\|}{\|\nabla_\theta L_{\text{CE}}\|+\varepsilon}$ and $\rho_{\text{mom/CE}} = \frac{\|\nabla_\theta L_{\text{mom}}\|}{\|\nabla_\theta L_{\text{CE}}\|+\varepsilon}$ have medians 0.28 and 0.17. Thus, our objective is well-conditioned despite mixed-order terms. $\varepsilon = 10^{-12}$ only for numerical stability.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[3] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485, 2022.

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004.

[5] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019.

[6] 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.

[7] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021.

[8] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[9] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023.

[10] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

[11] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

[12] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023.

[13] Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *European Conference on Computer Vision*, pages 147–165. Springer, 2024.

[14] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, 2015.

[15] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[16] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, 2019.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[18] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020.

[19] Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. *arXiv preprint arXiv:2503.12096*, 2025.

[20] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

[21] Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned clip. In *International Conference on Machine Learning (ICML)*, 2024.

[22] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023.

[23] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023.

[24] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

[26] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15659–15669, 2023.