O-TPT: Orthogonality Constraints for Calibrating Test-time Prompt Tuning in Vision-Language Models

Supplementary Material

In this supplementary material, we provide the following:

1. We reveal the relation between calibration performance and angular distances (sec. A1)
2. We compare the SCE performance (sec. A2)
3. Reliabilty plots comparison with C-TPT [5]. (sec. A3)
4. Calibration performance with different hard prompt styles (sec. A4)
5. Calibration with a Combination of C-TPT and O-TPT (sec. A5)
6. O-TPT results on Medical Prompt tuning methods (sec. A6)
7. Pareto Front analysis with varying λ (sec. A7)

A1. Relation Between Calibration Performance and Angular Distances

To further validate our motivation, we conduct an experiment using 80 different hard prompt styles [4] to evaluate their corresponding Expected Calibration Error (ECE) performance and the mean cosine similarity of text features. This evaluation is performed on zero-shot inference using the CLIP-B/16 backbone. Figure 1 illustrates the results for prompts that yield higher accuracies across seven diverse datasets: Flower, Caltech101, SUN397, Cars, Pets, UCF101, and Food101. Specifically, we focus on the top 10 prompt styles that provide the highest accuracies.

The results reveal a clear trend between mean cosine similarity (an angular distance measure) and ECE (calibration performance). A lower mean cosine similarity correlates with a reduced ECE, indicating that greater angular distancing among text features promotes better calibration. This suggests that prompts with text features exhibiting greater angular distances between their representations lead to improved calibration outcomes.

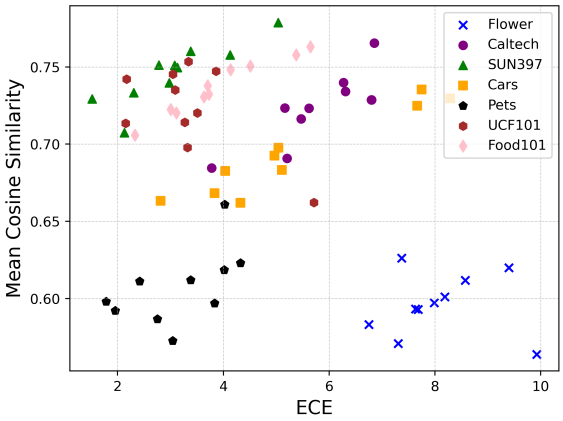
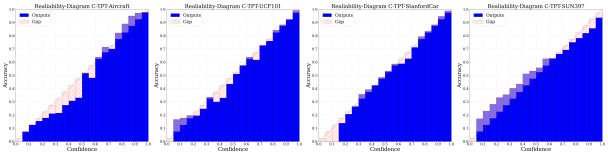
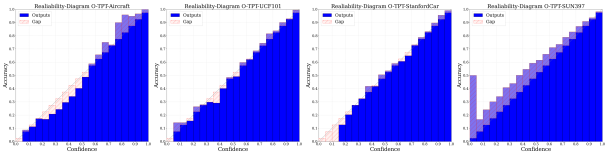


Figure 1. Relation of ECE with cosine similarities (of textual features) on CLIP-B/16 backbone.



(a) C-TPT:Air (b) C-TPT:UCF (c) C-TPT:Car (d) C-TPT:SUN



(e) O-TPT:Air (f) O-TPT:UCF (g) O-TPT:Car (h) O-TPT:SUN

Figure 2. Reliability diagrams for CLIP-B/16.

A2. SCE performance comparison

Tables 1 and 2 present the Static Calibration Error (SCE) results across 10 datasets using CLIP-B/16 and CLIP RN-50 backbones. Our method, O-TPT, outperforms C-TPT on both backbones, achieving an overall average SCE values of **1.07** for CLIP-B/16 and **1.24** for CLIP RN-50, demonstrating improved calibration performance.

A3. Reliability Plots

Figure 2 and Figure 3 illustrate the reliability diagrams for the CLIP-B/16 and CLIP RN-50 backbones, respectively, comparing the performance of C-TPT and O-TPT across the Aircraft, UCF101, Car, and SUN397 datasets. For the CLIP-B/16 backbone (Fig. 2), O-TPT effectively addresses the overconfidence problem and outperforms C-TPT, as evident from the reliability diagrams in the top and bottom rows of Fig. 2. Similarly, with CLIP RN-50 backbone, (Fig. 3) shows that O-TPT provides significantly better calibration compared to C-TPT, particularly in addressing overconfident predictions.

A4. Calibration with different prompts

This section presents the results of O-TPT initialized with different prompts, such as 'a photo of the cool {class}' and

Method	Metric	DTD	FLW	Food	SUN	Air	Pets	Calt	UCF	SAT	Car	Avg
Zero Shot	SCE	1.33	0.59	0.20	0.12	0.52	0.68	0.25	0.52	6.18	0.23	1.06
TPT	SCE	1.44	0.51	0.17	0.15	0.58	0.60	0.16	0.57	7.07	0.25	1.15
C-TPT	SCE	1.31	0.52	0.22	0.14	0.56	0.58	0.22	0.52	6.81	0.22	1.11
O-TPT (Ours)	SCE	1.24	0.53	0.19	0.12	0.56	0.57	0.17	0.51	6.58	1.07	**1.07**

Table 1. Static Calibration Error (SCE) (10^{-2}) performance comparison with CLIP-B/16 backbone.

Method	Metric	DTD	FLW	Food	SUN	Air	Pets	Calt	UCF	SAT	Car	Avg
Zero Shot	SCE	1.31	0.66	0.29	0.12	0.54	0.73	0.35	0.54	7.39	0.23	1.22
TPT	SCE	1.52	0.63	0.25	0.11	0.60	0.54	0.38	0.51	8.23	0.24	1.30
C-TPT	SCE	1.43	0.62	0.26	0.11	0.53	0.67	0.32	0.51	8.07	0.23	1.27
O-TPT (Ours)	SCE	1.34	0.60	0.27	0.12	0.51	0.69	0.3	0.5	7.85	0.22	**1.24**

Table 2. Static Calibration Error (SCE) (10^{-2}) performance comparison with CLIP-RN-50 backbone.

Method	Metric	DTD	FLW	Food	SUN	Air	Pets	Calt	UCF	SAT	Car	Avg
Zero Shot	Acc.	38.4	64.5	81.4	62.4	22.7	86.2	88.1	67.6	34.6	66.5	61.24
	ECE	7.43	4.59	1.10	6.11	2.83	7.43	14.1	2.65	14.1	4.59	7.01
TPT	Acc.	45.5	67.9	84.9	65.9	24.5	87.4	91.5	66.4	43.3	67.2	64.45
	ECE	20.0	14.6	5.74	13.3	19.2	6.34	3.11	14.1	18.2	6.36	12.09
C-TPT	Acc	46.3	69.6	84.1	65.5	24.7	88.8	91.7	67.0	43.0	66.9	64.76
	ECE	18.0	10.6	2.43	10.7	10.5	1.59	1.89	7.42	8.73	1.64	7.35
O-TPT (Ours)	Acc.	44.62	68.29	84.82	63.05	23.16	88.28	91.48	64.74	44.81	66.02	63.92
	ECE	12.85	4.67	1.85	2.67	6.37	3.59	3.0	4.08	8.33	2.71	**5.01**

Table 3. Comparison of calibration performance with CLIP-B/16 backbone with the prompt of 'a photo of the cool {class}'

Method	Metric	DTD	FLW	Food	SUN	Air	Pets	Calt	UCF	SAT	Car	Avg
Zero Shot	Acc.	39.6	57.7	73.0	56.5	16.1	79.8	80.9	56.3	21.9	56.9	60.24
	ECE	6.94	5.14	1.49	3.33	6.42	3.30	4.79	3.76	13.9	4.83	5.39
TPT	Acc.	39.2	61.6	75.8	60.2	17.4	82.6	86.5	59.7	26.3	58.8	56.81
	ECE	24.8	17.0	7.93	11.4	17.5	7.31	6.02	14.4	15.7	4.49	12.65
C-TPT	Acc	39.1	67.0	76.0	60.3	17.4	83.5	87.1	59.6	26.1	57.2	57.33
	ECE	18.0	6.34	3.70	8.28	13.5	1.75	2.85	8.82	11.2	1.65	7.61
O-TPT (Ours)	Acc.	40.54	65.49	75.51	58.98	15.99	83.78	86.98	58.79	26.89	56.77	56.97
	ECE	12.42	3.03	1.32	3.35	8.36	4.47	3.53	3.27	7.21	2.74	**4.97**

Table 4. Comparison of calibration performance with CLIP-RN-50 backbone with the prompt of 'a photo of the cool {class}'

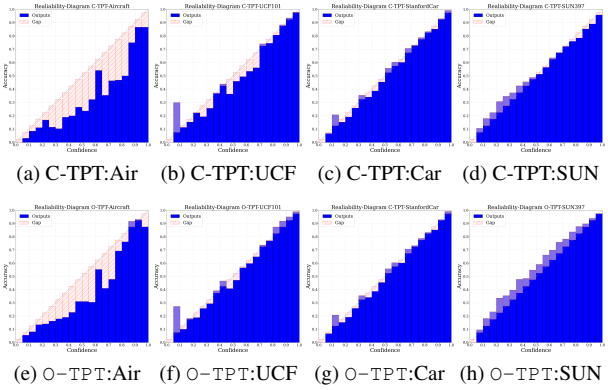
'an example of {class}', across CLIP-B/16 and CLIP RN-50 backbones. Tables 3 and 4 summarize the performance of O-TPT with the prompt 'a photo of the cool {class}' and 5 context token tuning. For CLIP-B/16, O-TPT achieves an overall reduced calibration error of **5.01**, compared to 7.35 for C-TPT, while for RN-50, it achieves **4.97**, com-

Method	Metric	DTD	FLW	Food	SUN	Air	Pets	Calt	UCF	SAT	Car	Avg
Zero Shot	Acc.	42.4	64.7	83.9	61.4	22.3	82.5	90.9	64.8	38.8	64.6	61.63
	ECE	4.94	4.70	2.78	3.33	7.09	2.91	7.51	2.79	13.4	2.49	5.64
TPT	Acc.	45.8	69.4	84.8	65.3	22.9	83.0	93.0	67.1	40.7	67.3	63.93
	ECE	20.5	12.2	5.05	7.94	16.2	7.30	2.91	11.6	20.8	6.26	11.07
C-TPT	Acc	45.4	71.5	84.3	66.0	23.6	86.9	93.8	66.4	51.5	66.6	65.6
	ECE	15.5	4.49	1.36	3.54	9.05	2.89	1.62	3.87	5.18	1.75	4..93
O-TPT (Ours)	Acc.	45.45	70.32	84.79	64.5	22.77	87.76	93.35	65.4	51.011	66.25	65.16
	ECE	11.79	3.22	2.92	4.62	7.92	3.29	3.24	2.63	5.08	1.92	**4.66**

Table 5. Comparison of calibration performance with CLIP-B/16 backbone with the prompt of 'an example of {class}'

Method	Metric	DTD	FLW	Food	SUN	Air	Pets	Calt	UCF	SAT	Car	Avg
Zero Shot	Acc.	41.10	58.10	75.20	56.20	16.10	75.70	80.30	56.30	25.5	55.8	48.45
	ECE	5.20	3.04	3.31	3.68	4.80	2.52	7.91	3.76	9.43	4.80	4.845
TPT	Acc.	41.2	62.7	76.1	60.7	17.9	77.2	87.1	57.7	29.4	57.7	56.77
	ECE	20.2	12.2	4.83	8.19	15.2	6.98	5.12	15.3	11.1	5.52	10.46
C-TPT	Acc	41.2	65.4	75.8	61.4	17.6	78.0	88.4	58.4	30.4	57.1	57.37
	ECE	15.6	2.97	1.90	4.84	7.16	2.72	2.89	6.99	7.69	2.05	5.48
O-TPT (Ours)	Acc.	41.19	65.49	75.62	60.97	16.71	77.79	88.36	57.94	33.32	56.733	57.412
	ECE	13.59	2.49	1.47	3.38	6.6	2.55	2.56	6.2	5.07	2.69	**4.66**

Table 6. Comparison of calibration performance with CLIP- RN-50 backbone with the prompt of 'an example of {class}'



(a) C-TPT:Air (b) C-TPT:UCF (c) C-TPT:Car (d) C-TPT:SUN

(e) O-TPT:Air (f) O-TPT:UCF (g) O-TPT:Car (h) O-TPT:SUN

Figure 3. Reliability diagrams for CLIP RN-50.

pared to 7.61 for C-TPT. Similarly, Tables 5 and 6 present the results for the prompt 'an example of {class}' and 4 context token tuning. Here, O-TPT again outperforms C-TPT, achieving a reduced calibration error of **4.66** (CLIP-B/16) compared to 4.93, and **4.66** (CLIP RN-50) compared to 5.48. These results consistently demonstrate that O-TPT effectively reduces calibration errors across various prompt initializations, showcasing its robustness and adaptability in diverse settings.

A5. Calibration with a Combination of C-TPT and O-TPT

Tab. 7 shows that O-TPT + C-TPT can outcompete O-TPT in calibration performance, thereby revealing the generalizability of O-TPT over a stronger baseline.

Method	Metric	DTD	FLW	UCF
O-TPT	ACC	45.68	70.07	64.16
	ECE	7.88	3.87	2.34
O-TPT + C-TPT	ACC	45.2	70.6	64.1
	ECE	**7.06**	**3.41**	**2.14**

Table 7. O-TPT + C-TPT on DTD,FLW and UCF.

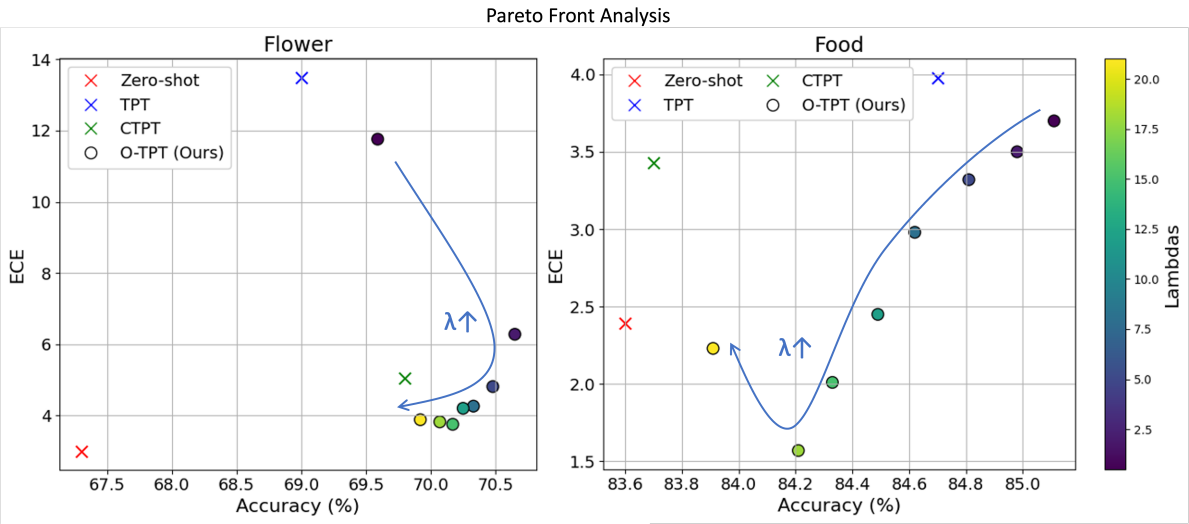


Figure 4. Pareto front analysis

Method	Metric	Covid(BA)	Covid(CA)
BAPLe	ACC	99.90	82.5
	ECE	3.21	15.64
BAPLe+ O-TPT	ACC	99.62	81.36
	ECE	**0.91**	**5.97**

Table 8. MedCLIP: BAPLe + O-TPT on Covid dataset.

Method	Metric	ISIC'18
FPT	ACC	98.43
	ECE	0.2328
FPT+ O-TPT	ACC	98.25
	ECE	**0.1381**

Table 9. FPT+O-TPT on ISIC2018.

Method	Metric	KC
PS	ACC	76.6
	ECE	15.54
PS+ O-TPT	ACC	76.2
	ECE	**12.73**

Table 10. PLIP: Promptsmooth (PS)+O-TPT on KatherColon (KC).

A6. O-TPT results on Medical Prompt tunning methods

In Tab.9, we evaluate FPT[2] and FPT+O-TPT on ISIC2018, showing encouranging ECE reduction while maintaining accuracy. Tab.10 provides comparison using PLIP with Prompt Smooth (PS)[3], where PS+O-TPT improves calibration. Similarly, Tab.8 provides comparison using MedCLIP with BAPLe[1] where BAPLe+O-TPT improves calibration.

A7. Pareto Front analysis with varying lamdas

Fig. 4 shows the Pareto front analysis on Food and Flower datasets, highlighting the accuracy-ECE tradeoff with varying λs. Our method achieves a better balance than TPT and C-TPT across most λ values in two datasets.

References

[1] Asif Hanif, Fahad Shamshad, Muhammad Awais, Muzammal Naseer, Fahad Shahbaz Khan, Karthik Nandakumar, Salman Khan, and Rao Muhammad Anwer. Baple: Backdoor attacks on medical foundational models using prompt learning, 2024. 4

[2] Yijin Huang, Pujin Cheng, Roger Tam, and Xiaoying Tang. Fine-grained prompt tuning: A parameter and memory efficient transfer learning method for high-resolution medical image classification, 2024. 4

[3] Noor Hussein, Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Promptsmooth: Certifying robustness of medical vision-language models via prompt learning, 2024. 4

[4] 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. 1

[5] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *International conference on machine learning*, 2024. 1