Intra-Class Probabilistic Embeddings for Uncertainty Estimation in Vision-Language Models

Zhenxiang Lin Maryam Haghighat Will Browne Dimity Miller
Queensland University of Technology, Brisbane, Australia
{z25.lin, maryam.haghighat, will.browne, d24.miller}@qut.edu.au

We provide additional results and analyses to complement the findings in the main paper. We first report the classification accuracy of different methods under various backbones. We then present supplementary results on PCA-based feature projection to further demonstrate its effect on alleviating ill-conditioned covariance matrices. Finally, we show additional qualitative visualizations of our uncertainty estimates on ImageNet to illustrate the effectiveness of our approach.

A. Accuracy of different methods

Table 1, 2 and 3 show the accuracy of different methods under different backbones.

Table 1. Accuracy of different methods (CLIP ViT-B/32)

Method	ImageNet	Flowers102	Food101	EuroSAT	DTD
MaxCosine	62.02	63.83	81.22	35.99	42.85
MaxSoftmax	62.02	63.83	81.22	35.99	42.85
Entropy	62.02	63.83	81.22	35.99	42.85
TempScaling [9]	62.02	63.83	81.22	35.99	42.85
ProbVLM [10]	61.78	59.29	81.25	35.88	41.84
PCME++ [2]	77.61	82.05	89.05	98.75	75.65
BayesVLM [1]	61.30	64.90	80.61	30.68	30.44
Ours	62.02	63.83	81.22	35.99	42.85

Table 2. Accuracy of different methods (CLIP ViT-B/16)

Method	ImageNet	Flowers102	Food101	EuroSAT	DTD
MaxCosine	66.73	67.69	88.65	42.17	44.33
MaxSoftmax	66.73	67.69	88.65	42.17	44.33
Entropy	66.73	67.69	88.65	42.17	44.33
TempScaling [9]	66.74	67.64	88.65	42.17	44.33
Zero [7]	71.19	68.24	88.36	42.30	45.80
ProLIP++ [3]	73.70	78.84	90.94	44.93	63.36
TrustVLM [6]	65.04	99.01	86.88	81.06	72.28
Ours	66.73	67.69	88.65	42.17	44.33

Table 3. Accuracy of different methods (SigLIP ViT-B/16)

Method	ImageNet	Flowers102	Food101	EuroSAT	DTD
MaxCosine	75.67	83.77	89.63	41.35	62.88
MaxSoftmax	75.67	83.77	89.63	41.35	62.88
Entropy	75.67	83.77	89.63	41.35	62.88
TempScaling [9]	75.67	83.77	89.63	41.33	62.94
Zero [7]	77.71	82.35	88.73	29.00	64.01
BayesVLM [1]	75.68	81.77	89.58	41.30	48.58
TrustVLM [6]	68.68	99.07	88.20	80.62	71.99
Ours	75.67	83.77	89.63	41.35	62.88

B. Additional Results on PCA-based Feature Projection

To complement the results presented in Sec. 4.3, we further report the log condition numbers of class-wise covariance matrices for Flowers102 [8] and DTD [4] datasets. As shown in Fig. 1, applying PCA consistently reduces the condition number, indicating that feature projection mitigates ill-conditioning across datasets of different scales and domains.

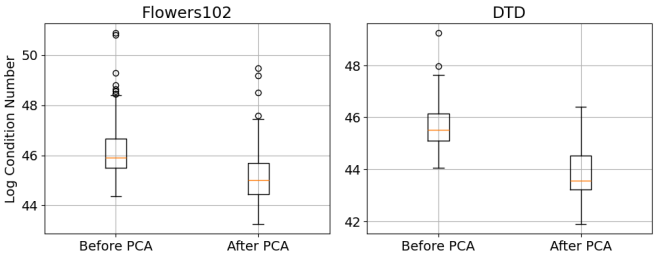

Figure 1. Log condition numbers of class-wise covariance matrices before and after PCA on Flowers102 and DTD. PCA reduces the condition number, indicating our probabilstic embeddings effectively alleviate the ill-conditioned covariance matrices.

C. Additional Quantitative Examples

Figure 2 shows additional quantitative examples on ImageNet [5].

Figure 2. Visualization of our uncertainty on ImageNet [5]. Green means examples where are uncertainty has correctly distinguished between correct and error, and red means when our uncertainty failed to distinguish them. The threshold of uncertainty is 0.5.

References

[1] Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. Post-hoc probabilistic vision-language models. *arXiv preprint arXiv:2412.06014*, 2024. 1

[2] Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*, 2024. 1

[3] Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2025. 1

[4] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[5] 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. 1, 2

[6] Hao Dong, Moru Liu, Jian Liang, Eleni Chatzi, and Olga Fink. To trust or not to trust your vision-language model's prediction. *arXiv preprint arXiv:2505.23745*, 2025. 1

[7] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024. 1

[8] 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. 1

[9] Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *International Conference on Machine Learning*, pages 48791–48808. PMLR, 2024. 1

[10] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1899–1910, 2023. 1