CalibPrompt: Calibration-Aware Prompt Learning for Medical Vision-Language Models

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE
MICHIGAN STATE UNIVERSITY

Abhishek Basu, Fahad Shamshad, Ashshak Sharifdeen, Karthik Nandakumar, and Muhammad Haris Khan

BMVC 2025

Motivation

- Medical Vision-Language Models (Med-VLMs) show high diagnostic accuracy but **generate dangerously overconfident/underconfident predictions**, **risking misdiagnosis and reduced clinical trust.**

- **Limited labeled medical data** necessitates **parameter-efficient calibration approaches** for maintaining diagnostic accuracy and **providing reliable uncertainty estimates**.

DATASET	Med-VLM	ACCURACY(%)	CONFIDENCE(%)	CALIBRATION
Hard Prompt: *A chest X-ray image of [class] patient*				
COVID	BioMedCLIP	84.37	95.0	OVERCONFIDENT
	MedCLIP	78.77	50.0	UNDERCONFIDENT
RSNA18	BioMedCLIP	49.71	79.0	OVERCONFIDENT
	MedCLIP	47.60	34.0	UNDERCONFIDENT
Hard Prompt: *An H&E image of [class]*				
KATHER	PLIP	57.80	74.0	OVERCONFIDENT
	QuiltNet	60.39	58.0	UNDERCONFIDENT
Hard Prompt: *An H&E image patch of [class] skin tissue*				
PanNuke	PLIP	56.42	76.0	OVERCONFIDENT
	QuiltNet	55.59	79.0	OVERCONFIDENT
Hard Prompt: *An H&E image patch of [class] tissue*				
DigestPath	PLIP	80.53	74.0	UNDERCONFIDENT
	QuiltNet	53.39	73.0	OVERCONFIDENT

Contributions

Introduce **CalibPrompt**, a novel approach to calibrate Med-VLMs during prompt learning with minimal parameter tuning eliminating the need for expensive full-model retraining.

- **Smoothed Accuracy and Confidence Matching Loss (SMAC):** Aligns predicted confidence scores with smoothed empirical accuracy distributions, **addressing diagnostic ambiguities** and **reducing overconfident predictions in medical tasks**.

- **Angular Seperation Loss (AS):** Encourages **greater angular distance** between class-specific textual embeddings, **preventing artificial compression of the feature space** and improving confidence calibration.

- Validate the effectiveness of **CalibPrompt** through experiments across diverse medical datasets(Kather, COVIDx, RSNA18, PanNuke, DigestPath) and Med-VLMs (**BioMedCLIP**, **MedCLIP**, **PLIP**, **QuiltNet**), and show that it achieves superior **calibration performance.**

Proposed Method

How does it perform in limited data?

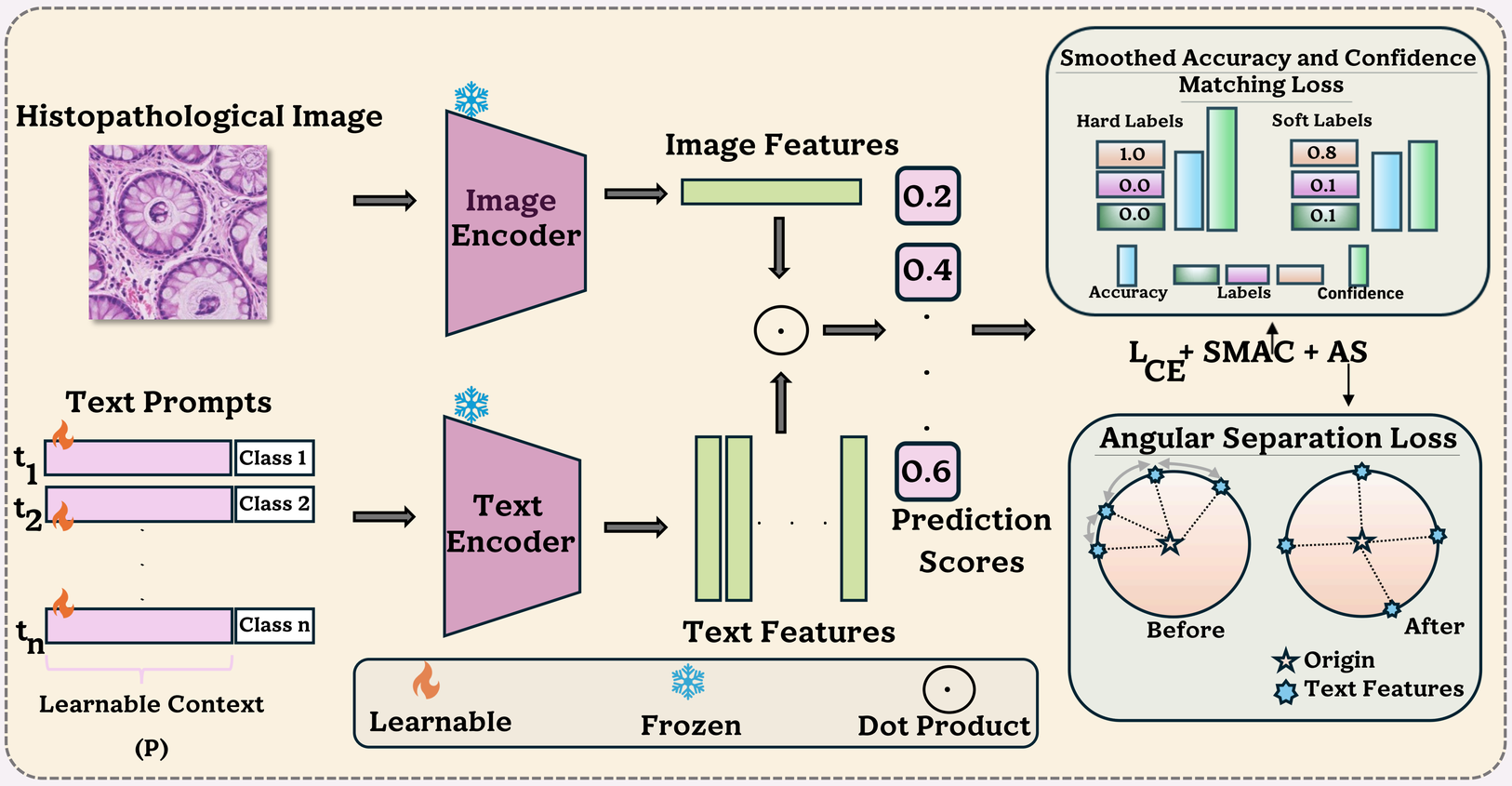
CalibPrompt utilizes **few-shot prompt learning**, making it highly effective and practical for medical scenarios with very limited labeled data

How to ensure parameter efficiency?

Using **Prompt Tuning - update only ~0.1% of parameters** while keeping the pre-trained model frozen.

How does CalibPrompt provide reliable predictions?

SMAC loss calibrates confidence by aligning it with true accuracy, while **AS loss** enhances discrimination by separating class embeddings. **Together, they balance confidence reliability and class distinctness**, yielding well-calibrated predictions.



Results

Med-VLM →	PLIP				BioMedCLIP			
Loss ↓	ECE ↓	ACE ↓	MCE ↓	KDE ↓	ECE ↓	ACE ↓	MCE ↓	KDE ↓
Cross Entropy-based Losses								
Cross Entropy	13.66	13.66	8.53	13.25	6.82	6.78	2.65	6.54
CE + DCA (BMVC'20)	12.7	12.9	7.4	12.76	4.77	7.03	2.26	4.43
CE + MMCE (ICML'18)	10.92	11.06	6.33	10.77	8.55	8.55	3.57	8.49
CE + MDCA (CVPR'22)	6.89	6.88	4.09	6.6	4.88	4.87	2.08	5.1
MbLS (CVPR'22)	13.68	13.68	8.53	13.26	6.79	6.77	2.65	6.56
LogitNorm (ICML'22)	20.38	20.38	14.46	19.73	22.38	22.38	10.47	22.18
ZS-Norm (ECCV'24)	12.27	12.27	7.99	11.82	6.13	6.11	2.11	5.91
Penalty (ECCV'24)	2.76	**3.05**	**1.16**	3.27	5.88	6.06	2.26	6.08
CE + SMAC [Ours]	6.89	6.88	4.09	6.6	4.79	4.78	2.09	5.02
CE + (SMAC + AS) [Ours]	5.89	5.77	2.9	5.82	**4.14**	**4.28**	**1.7**	**4.6**
Label Smoothing-based Losses								
Label Smoothing	10.76	10.76	6.32	10.71	14.79	14.79	7.08	14.64
LS + MDCA (CVPR'22)	6.44	6.43	2.87	6.37	16.09	16.09	9.19	15.87
LS + SMAC [Ours]	2.75	4.05	1.9	**2.09**	12.82	12.82	5.49	12.69
LS + (SMAC +AS) [Ours]	**2.64**	3.51	1.73	2.25	5.22	5.14	2.15	5.27