Calibration-Aware Prompt Learning for Medical Vision-Language Models

Abhishek Basu[1]
abhishek.basu@mbzuai.ac.ae
Fahad Shamshad[1]
fahad.shamshad@mbzuai.ac.ae
Ashshak Sharifdeen[1]
ashshak.sharifdeen@mbzuai.ac.ae
Karthik Nandakumar[1,2]
nandakum@msu.edu
Muhammad Haris Khan[1]
muhammad.haris@mbzuai.ac.ae

[1] Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

[2] Department of Computer Science and Engineering, Michigan State University (MSU), East Lansing, USA

Abstract

Medical Vision-Language Models (Med-VLMs) have demonstrated remarkable performance across diverse medical imaging tasks by leveraging large-scale image-text pretraining. However, their confidence calibration is largely unexplored, and so remains a significant challenge. As such, miscalibrated predictions can lead to overconfident errors, undermining clinical trust and decision-making reliability. To address this, we introduce `CalibPrompt`, the first framework to calibrate Med-VLMs during prompt tuning. `CalibPrompt` optimizes a small set of learnable prompts with carefully designed calibration objectives under scarce labeled data regime. First, we study a regularizer that attempts to align the smoothed accuracy with the predicted model confidences. Second, we introduce an angular separation loss to maximize textual feature proximity toward improving the reliability in confidence estimates of multimodal Med-VLMs. Extensive experiments on four publicly available Med-VLMs and five diverse medical imaging datasets reveal that `CalibPrompt` consistently improves calibration without drastically affecting clean accuracy. Our code is available at https://github.com/iabh1shekbasu/CalibPrompt.

1 Introduction

Medical Vision-Language models (Med-VLMs) have emerged as powerful tools for medical image analysis, leveraging large-scale image-text pretraining to enable zero-shot classification across diverse medical imaging tasks [1, 19, 36]. These models align medical images with textual descriptions, facilitating interpretation and diagnosis without requiring task-specific fine-tuning. However, despite their strong performance in recognizing medical concepts, Med-VLMs often suffer from poor calibration, where their confidence scores fail

to reliably indicate actual correctness [8, 24], which is particularly concerning in medical imaging, as miscalibrated model can lead to misdiagnoses and undermine clinical trust [16].

Model calibration techniques generally fall into two categories: post-hoc calibration and training-time calibration. Post-hoc methods, such as Platt scaling [26] and temperature scaling [4], adjust confidence scores after training via a transformation function. While computationally inexpensive, they have two key limitations: (1) reliance on a small validation set, which may not reflect real-world medical distributions [22, 31], and (2) failure to improve the model's internal representations, leaving calibration issues unresolved at the decision-making level [25]. In contrast, training-time calibration jointly optimizes accuracy and calibration, leading to more robust and generalizable confidence estimates [14]. By integrating calibration objectives into training, it ensures well-calibrated Med-VLM predictions across medical tasks, enhancing clinical trust. However, fine-tuning large-scale Med-VLMs with calibration objectives is often impractical due to high computational cost and requirement of massive labeled medical datasets.

Meanwhile, prompt tuning has emerged as an efficient alternative to full-model fine-tuning for adapting Med-VLMs to downstream tasks with limited data [37]. Unlike conventional fine-tuning, which updates the entire model, prompt tuning modifies only a small set of learnable parameters, significantly reducing computational costs while preserving generalization [5, 10]. This efficiency is particularly valuable in medical imaging, where labeled data is scarce and full fine-tuning is often impractical. Despite its strong performance in data-limited settings, prompt tuning primarily optimizes classification and does not inherently improve model calibration. This raises a key question: *Can the efficiency of prompt tuning, with its low data requirements and minimal parameter updates, be leveraged to enhance calibration without compromising adaptability?* Addressing this is crucial to ensure Med-VLMs produce both accurate and well-calibrated predictions for reliable clinical decision-making.

In this paper, we introduce `CalibPrompt`, the first approach to calibrate Medical Vision-Language Models during prompt learning. Specifically, we make following technical contributions.

— We investigate a simple regularizer that aligns softened accuracy with model confidences to effectively calibrate under class ambiguities inherent in medical imaging.

— We propose a novel angular separation loss that promotes angular gap between *textual features* during prompt tuning, specifically tailored to the multimodal architecture of Med-VLMs.

— We demonstrate the effectiveness of `CalibPrompt` through comprehensive experiments across four publicly available Med-VLMs and five downstream datasets spanning different imaging modalities, achieving superior calibration performance while tuning only 0.1% of the model parameters.

2 Related Work

Medical Vision-Language Models (Med-VLMs). Med-VLMs inspired by Contrastive Language-Image Pretraining (CLIP), have advanced medical imaging by aligning image-text pairs across modalities such as X-ray, histopathology, and retinal imaging. These models enable zero-shot and few-shot classification, making them particularly valuable in data-scarce medical applications [1, 19, 36]. To adapt Med-VLMs efficiently, prompt learning (PL) has

emerged as a lightweight alternative to full-model fine-tuning [6, 10, 37]. By introducing learnable prompt tokens without modifying the backbone, PL enhances task performance while maintaining computational efficiency, making it well-suited for medical imaging with limited data. In this work, we investigate whether PL can simultaneously improve calibration and task adaptation, positioning it as a parameter-efficient alternative to computationally intensive calibration methods for accurate and trustworthy predictions in medical AI.

Confidence Calibration. Confidence calibration assesses how well a model's predicted confidence aligns with its actual accuracy, a critical requirement in high-stakes domains like medical imaging. Well-calibrated models yield reliable uncertainty estimates, crucial for clinical decision-making, as overconfident misclassifications can lead to severe consequences [4, 16]. Post-hoc calibration methods, such as Temperature Scaling, adjust model logits via a learned temperature parameter optimized on a held-out validation set [4]. While computationally efficient, these methods heavily depend on labeled datasets closely matching the target distribution [22, 31].To overcome such limitations, train-time calibration methods incorporate calibration objectives directly into model training, typically through auxiliary loss functions alongside primary task-specific objectives, resulting in more robust and generalizable confidence estimates [7, 14, 25]. For instance, MACSO [13] aligns predicted confidences with softened target distributions derived from the model's internal knowledge, utilizing correlation-based distance measures. Other effective train-time approaches include Margin-based Label Smoothing (MbLS) [18], which imposes inequality constraints on logit distances to prevent overly confident predictions, and Logit Normalization (LogitNorm) [33], which enforces a constant norm on logits during training to mitigate overconfidence. Moreover, Murugesan et al. [20] identified expanded logit distributions in prompt-tuned models as a significant calibration issue, introducing Zero-Shot Normalization to restore alignment with pretrained distributions. Test-time calibration methods like C-TPT [34] addresses calibration via test-time prompt tuning by optimizing text feature dispersion using prototypes. Concurrent with our research, O-TPT [29] addresses calibration in vision-language models (VLMs) through test-time prompt tuning, enforcing strict orthogonality constraints on textual features without relying on labeled data. Similarly, Wang et al. [31] proposed DAC, which adjusts softmax temperatures based on semantic distances between embeddings, primarily targeting novel-class calibration in domains with numerous classes. However, medical imaging datasets typically involve fewer classes, limiting DAC's applicability in specialized medical contexts. Their findings underscore that post-hoc calibration methods alone cannot fully recover the pretrained calibration behaviour of VLMs. In contrast, we introduce a novel prompt-based calibration framework that operates effectively in few-shot scenarios by jointly optimizing regularization objectives in both probability and feature spaces. Unlike O-TPT, our method encourages (rather than strictly enforces) angular separation, providing greater flexibility to capture nuanced class relationships within specialized medical imaging domains.

3 Method

Our goal is to calibrate Med-VLMs in data-limited settings to ensure that their confidence scores accurately reflect prediction correctness. In medical imaging, miscalibrated models can lead to overconfident errors with serious clinical implications. To this end, we introduce a new approach `CalibPrompt` under prompt-learning setup which is built on two novel regularizers. The first regularizer matches the softened accuracy with predicted confidences,

while the second is an angular separation loss that explicitly maximizes the proximity between textual features. Below, we first describe zero-shot inference with Med-VLMs, then explain the prompt learning basics, and finally introduce our calibration-aware prompt tuning approach `CalibPrompt`.

3.1 Preliminaries:

Zero-Shot Inference for Med-VLMs: Med-VLMs learn a joint representation of images and text through contrastive pretraining, enabling zero-shot classification. These models consist of an image encoder $\mathbf{E}_{\text{img}} : \mathcal{I} \to \mathbb{R}^d$ and a text encoder $\mathbf{E}_{\text{txt}} : \mathcal{T} \to \mathbb{R}^d$, where $\mathcal{I}$ and $\mathcal{T}$ denote the image and text spaces, respectively. Given an input image $\mathbf{I} \in \mathcal{I} \subseteq \mathbb{R}^{H\times W\times C}$, the image encoder extracts a d-dimensional feature vector $\mathbf{v} = \mathbf{E}_{\text{img}}(\mathbf{I})$. Similarly, the text encoder maps a textual prompt $\mathbf{t}(y) \in \mathcal{T}$ associated with class label $y \in \mathcal{Y}$ into a text feature vector $\mathbf{u} = \mathbf{E}_{\text{txt}}(\mathbf{t}(y))$. During zero-shot inference, class labels $\{y_1, \ldots, y_K\}$ are converted into text prompts using a predefined template, such as $\mathbf{t}(y_i) =$ "`A H&E image of [CLASS` y_i`]`", and are processed by the text encoder to obtain $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$, where $\mathbf{u}_i = \mathbf{E}_{\text{txt}}(\mathbf{t}(y_i))$. For a test image $\mathbf{I}_t$, the similarity between the image and text features is computed as $s_i = \text{sim}(\mathbf{v}_t, \mathbf{u}_i)$, where $\mathbf{v}_t = \mathbf{E}_{\text{img}}(\mathbf{I}_t)$. The final classification probabilities are obtained using a softmax function as $\mathbb{P}(y_i|\mathbf{I}_t) = \frac{\exp(\tau s_i)}{\sum_{j=1}^{K}\exp(\tau s_j)}$, where τ is the softmax temperature parameter. The predicted label is then given by $\hat{y}_t = \arg\max_{y\in\mathcal{Y}} \mathbb{P}(y|\mathbf{I}_t)$. The corresponding confidence is given by $\hat{p}_t = \max_{y\in\mathcal{Y}} \mathbb{P}(y|\mathbf{I}_t)$. While this zero-shot framework enables flexible classification, Med-VLMs often produce overconfident predictions as shown in Table 1. A naive solution is to fine-tune Med-VLMs with explicit calibration objectives (i.e. train-time auxiliary losses); however, this is computationally expensive and requires extensive labeled data. Thus, an efficient alternative is needed to enhance calibration without full model retraining.

Prompt Learning: Prompt learning (PL) has emerged as an efficient alternative, enabling adaptation to new tasks without modifying the model backbone. Instead of updating the entire network, PL optimizes a small set of learnable prompt tokens, making it particularly useful for data-scarce medical applications. When a text prompt $\mathbf{t}(y_i) \in \mathcal{T}$ is passed to the text encoder, it is tokenized into a sequence of word embeddings. Typically, a class-specific text prompt is represented as $[\mathbf{w}]_1\ [\mathbf{w}]_2\ \cdots\ [\texttt{CLASS}\ y_i]$, where each $[*]$ denotes a word embedding. In PL, all fixed embeddings (except for the class token) are replaced with M learnable embeddings, transforming the prompt into $\mathbf{p}_{i1}\ \mathbf{p}_{i2}\ \cdots \mathbf{p}_{iM}\ [\texttt{CLASS}\ y_i]$, where each prompt embedding $\mathbf{p}$ has the same dimensionality as $[\mathbf{w}]$. Let $\mathcal{P} = \mathbf{p}_{im}$, where $i \in [1,K]$ and $m \in [1,M]$, represent the set of all learnable prompts. The output text feature vector, incorporating these learned prompts, is denoted as $\mathbf{u}_i(\mathcal{P})$, and the modified zero-shot classifier is $f_{\mathcal{P}}$.

Limitation and Motivation: While prompt learning effectively adapts Med-VLMs to downstream tasks with limited data, our empirical analysis reveals a critical limitation for Med-VLMs: *it increases calibration error despite improving classification accuracy*. We observe that prompt-tuned models consistently exhibit high Expected Calibration Error [21], indicating a mismatch between confidence scores and actual correctness. To understand this behavior, we analyze the geometric properties of learned textual prompts and find that prompt tuning significantly increases intra-class cosine similarity, causing class representations to become highly aligned (see Fig. 1 left). While this enhances classification separability, it also amplifies confidence scores, leading to overconfident predictions and calibration error, with approximately 22% of misclassifications occurring at high confidence levels of 0.9-1.0 (see Fig.1 middle). Further, our results reveal a strong correlation between high cosine sim-

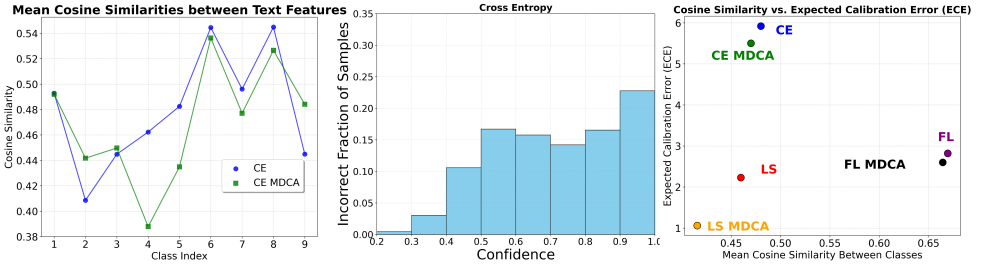

Figure 1: **Analysis of Prompt Learning Effects on Model Calibration**. *Left*: Cross entropy (CE) shows higher text feature similarity between classes than CE MDCA. *Middle*: Cross entropy histogram demonstrating overconfident misclassifications with higher confidence levels. *Right*: Greater feature similarity (CE, CE MDCA) directly correlates with increased calibration error compared to regularized approaches (LS, FL MDCA).

ilarity and increased miscalibration (see Fig. 1 right), underscoring the need for explicit text feature space regularization.

3.2 `CalibPrompt`: Calibration-Aware Prompt Learning

We introduce `CalibPrompt` as shown in Fig. 2, a new approach to improve confidence calibration in zero-shot classifiers based on Med-VLMs. Motivated by our observations, CalibPrompt incorporates learnable prompts into the text encoder and optimizes them with our proposed calibration-aware auxiliary losses to enforce appropriate confidence calibration while keeping the model backbone frozen. Specifically, given a zero-shot classifier f based on a pre-trained Med-VLM $(\mathbf{E}_{\text{image}}, \mathbf{E}_{\text{text}})$ and a few labeled samples $\{(\mathbf{I}_n, y_n)\}_{n=1}^{N}$ from a downstream dataset $\mathcal{D}$, where $\mathbf{I}_n \in \mathcal{I}$ and $y_n \in \mathcal{Y}$, `CalibPrompt` optimizes the learnable prompts $\mathcal{P}$ to jointly minimize classification loss and calibration error:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \frac{1}{N} \sum_{n=1}^{N} \Big[\mathcal{L}_{\text{CE}}(f_{\mathcal{P}}(\mathbf{I}_n), y_n) + \lambda \mathcal{L}_{\text{calib}}(f_{\mathcal{P}}(\mathbf{I}_n), y_n) \Big], \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, $\mathcal{L}_{\text{calib}}$ is our overall calibration objective, and λ balances accuracy and calibration objectives. The prompts are updated via backpropagation while keeping the Med-VLM frozen, preserving its pre-trained knowledge while optimizing for calibrated predictions. To address miscalibration, we introduce two complementary objectives: conforming softened accuracy with predicted confidences ($\mathcal{L}_{\text{SMAC}}$) and the Angular Separation Loss ($\mathcal{L}_{\text{AS}}$), forming our calibration objective $\mathcal{L}_{\text{calib}} = \alpha \mathcal{L}_{\text{SMAC}} + \beta \mathcal{L}_{\text{AS}}$.

Smoothed Accuracy and Confidence Matching (`SMAC`): Medical image classification often involves inherent class ambiguities, where diagnostic categories exhibit overlapping visual features. Traditional hard-label-based calibration methods [14] enforce overly rigid decision boundaries, leading to miscalibrated overconfidence. To address this, we propose aligning predicted confidences with smoothed empirical class frequencies in a class-wise manner, termed `SMAC`. This provides a nuanced training signal that captures inherent ambiguities in medical imaging. Let $\mathbf{p}_n = f_{\mathcal{P}}(\mathbf{I}_n)$ denote the predicted probability distribution for image $\mathbf{I}_n$ across K classes, and let $y_n \in \{1, 2, ..., K\}$ be the ground truth label. The `SMAC` loss

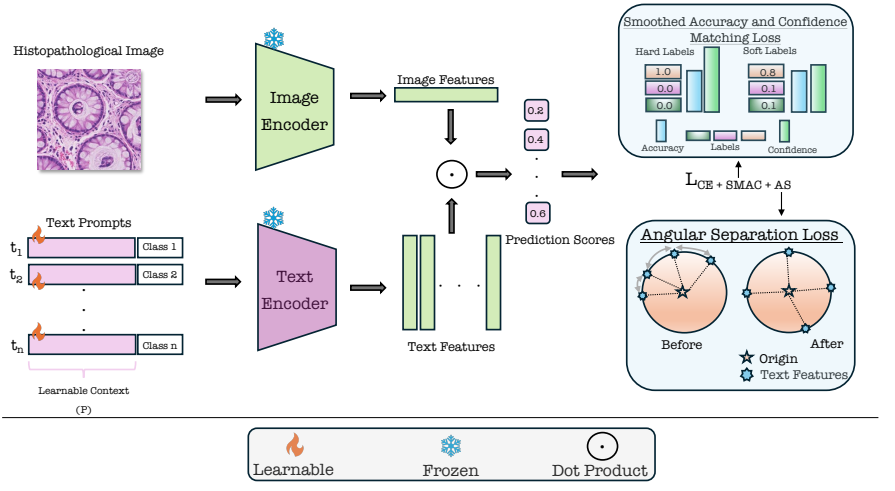

Figure 2: Overview of CalibPrompt. Learnable prompts are optimized using classification and calibration losses—SMAC and AS—while keeping the image and text encoders frozen. The SMAC loss aligns confidence with smoothed accuracy, while AS improves feature separation in the text embedding space.

is formulated as:

$$\mathcal{L}_{\text{SMAC}} = \frac{1}{K}\sum_{c=1}^{K}\Big|\underbrace{\frac{1}{N}\sum_{n=1}^{N} p_n^{(c)}}_{\substack{\text{avg. predicted}\\ \text{confidence}}} - \underbrace{\left[(1-\alpha)f_c + \frac{\alpha(1-f_c)}{K-1}\right]}_{\substack{\text{smoothed class}\\ \text{frequency}}}\Big|, \tag{2}$$

where $f_c = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}[y_n = c]$, and $\alpha \in [0,1)$ controls smoothing intensity. Using smoothed frequencies for confidence estimation, SMAC allows relaxed matching between predicted and empirical class distributions, thus reducing the likelihood of overconfident predictions in ambiguous scenarios.

Angular Separation (AS) Loss : Building on our observation that prompt tuning increases text embedding similarity (see Fig. 1), we address a key challenge in medical image classification where high inter-class feature similarity leads to overconfident predictions and degraded calibration. We propose an **Angular Separation Loss** for the textual embeddings, which discourages excessive similarity between class embeddings by minimizing their average pairwise cosine similarity. This ensures well-separated textual feature representations, improving confidence calibration while preserving classification accuracy. Mathematically, let $\mathbf{Z} \in \mathbb{R}^{K\times D}$ be the text feature matrix, where each row $\mathbf{z}_i$ represents the feature embedding of class i in a D-dimensional space. We compute the cosine similarity matrix between all pairs of feature vectors as $\mathbf{S} = \mathbf{Z}\mathbf{Z}^T$ where S_{ij} measures the cosine similarity between class embeddings $\mathbf{z}_i$ and $\mathbf{z}_j$. To focus only on inter-class relationships, we mask the diagonal elements (self-similarities) as $\mathbf{S}_{\text{off-diag}} = \mathbf{S} - \text{diag}(\mathbf{S})$. The Angular Separation Loss is then defined as the mean similarity across all class pairs:

$$\mathcal{L}_{\text{AS}} = \frac{1}{K(K-1)}\sum_{i=1}^{K}\sum_{j\neq i} S_{ij}. \tag{3}$$

Table 1: Zero-shot accuracy (%), confidence (%), and ECE (%) of Med-VLMs on X-ray and histopathology datasets. Final column shows over/underconfidence. Prompts used are shown above rows.

Dataset	Model	Accuracy (%) ↑	Confidence (%)	ECE (%) ↓	Calibration
Hard Prompt: *A chest X-ray image of [class] patient*					
COVID	BioMedCLIP	84.37	95.0	10.70	Overconfident
	MedCLIP	78.77	50.0	28.67	Underconfident
RSNA18	BioMedCLIP	49.71	79.0	29.49	Overconfident
	MedCLIP	47.60	34.0	14.05	Underconfident
Hard Prompt: *An H&E image of [class]*					
Kather	PLIP	57.80	74.0	16.32	Overconfident
	QuiltNet	60.39	58.0	4.20	Underconfident
Hard Prompt: *An H&E image patch of [class] skin tissue*					
PanNuke	PLIP	56.42	76.0	19.33	Overconfident
	QuiltNet	55.59	79.0	23.89	Overconfident
Hard Prompt: *An H&E image patch of [class] tissue*					
DigestPath	PLIP	80.53	74.0	6.14	Underconfident
	QuiltNet	53.39	73.0	19.90	Overconfident

By minimizing this loss, class embeddings become more distinct, reducing confidence overestimation and improving calibration. While SMAC loss refines probability-space confidence calibration, $\mathcal{L}_{\text{AS}}$ explicitly regularizes the text embedding space, ensuring both feature separability and confidence reliability.

4 Experiments

Datasets, baselines and implementation details: We hypothesize that the full fine-tuning approach can lead to overfitting in large networks when training data is limited, resulting in suboptimal feature representations. We evaluate our method with two regularizers on four Med-VLMs: PLIP [9], QuiltNet [11], MedCLIP [32], and BioMedCLIP [35], using five downstream datasets: COVIDX [28], RSNA18 [30], KatherColon [12], PanNuke [3], and DigestPath [2]. All experiments are conducted on an NVIDIA RTX A6000 GPU with 48GB memory. Our baselines include cross-entropy (CE), focal loss (FL) (given in supplementary), and label smoothing (LS), along with their combinations with established calibration regularization techniques such as DCA [17], MMCE [15] , MDCA [6], ZS-Norm[20], Penalty[20], MbLS [18], and LogitNorm [33]. We evaluated these against our proposed methods SMAC and SMAC with AS across both full model fine-tuning and prompt learning approaches. We used an 8-shot setting (8 images per class) for both variations, we used a learning rate of 2×10^{-7} for full model fine-tuning, while for prompt learning we used a learning rate of 0.002. Detailed hyperparameters are discussed in the supplementary.

Performance of the Med-VLMs is measured using accuracy (ACC). Similarly, for calibration, Expectation Calibration Error [21] (ECE), Adaptive Calibration Error [23] (ACE), Maximum Calibration Error [21] (MCE), and Expectation Calibration Error - Kernel Density Estimates[27] ECE^{KDE}.

Table 2: Comparison of proposed regularizers (SMAC, SMAC+AS) with baseline methods using Cross Entropy (CE), and Label Smoothing (LS). Accuracy (ACC, %) and Expected Calibration Error (ECE, %) are shown for PLIP and QuiltNet on histopathology datasets (Kather, PanNuke, DigestPath). Subscripts FT and PL denote Few-shot Fine-Tuning and Prompt Learning. Best results are in **bold**, second-best underlined.

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$_{FT}$	81.17	8.46	60.08	15.79	82.68	5.50	79.19	17.20	72.80	7.67	75.93	4.62	75.31	9.87
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
CE + DCA$_{FT}$	82.63	8.20	59.37	17.57	83.76	5.53	80.11	17.90	51.19	13.07	31.49	33.14	64.76	15.90
CE + DCA$_{PL}$	85.74	3.60	67.94	11.48	74.71	13.91	88.16	1.50	74.47	4.29	84.24	1.71	79.21	6.08
CE + MMCE$_{FT}$	81.16	8.46	60.15	15.68	82.86	5.52	79.07	17.06	73.05	7.26	80.50	7.60	76.13	10.26
CE + MMCE$_{PL}$	83.52	3.07	67.05	12.16	78.81	9.68	90.97	2.11	68.08	17.95	85.13	4.49	78.93	8.24
CE + MDCA$_{FT}$	81.14	8.41	59.90	16.08	82.54	5.48	79.14	17.17	72.80	7.67	31.07	32.53	67.77	14.56
CE + MDCA$_{PL}$	83.91	5.79	70.67	9.69	88.92	4.08	89.99	1.61	69.29	13.46	84.89	6.55	81.28	6.86
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
LogitNorm$_{PL}$	86.52	5.12	57.10	31.72	84.80	9.04	88.22	3.42	72.91	13.88	87.17	5.26	79.45	11.41
ZS-Norm$_{PL}$	85.63	3.07	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
CE + SMAC$_{FT}$	81.39	8.52	64.37	9.20	82.94	5.75	79.75	17.85	72.78	5.87	80.96	8.64	77.03	9.31
CE + SMAC$_{PL}$	84.11	5.42	70.67	9.69	88.92	4.08	89.99	1.57	69.29	13.46	84.89	6.55	81.31	6.80
CE + (SMAC + AS)$_{FT}$	81.45	8.55	64.16	9.56	83.04	5.78	78.38	13.13	72.91	6.02	81.26	8.81	76.87	8.64
CE + (SMAC + AS)$_{PL}$	84.65	5.05	65.81	9.30	89.75	**2.47**	89.53	2.82	69.02	16.51	86.52	6.03	80.88	7.03
Label Smoothing-based Losses														
Label Smoothing$_{FT}$	80.64	9.32	59.83	15.66	82.24	5.81	77.99	17.34	72.73	7.73	76.58	7.11	75.00	10.50
Label Smoothing$_{PL}$	85.28	2.23	67.85	16.01	80.87	5.50	89.43	3.26	67.55	16.93	76.69	7.03	77.95	8.49
LS + MDCA$_{FT}$	80.72	9.44	60.24	14.92	82.55	5.88	78.04	17.36	72.73	7.73	77.45	12.38	75.29	11.29
LS + MDCA$_{PL}$	83.58	**1.06**	68.93	9.39	83.29	3.49	91.53	4.37	75.32	3.91	88.05	0.82	81.78	3.84
LS + SMAC$_{FT}$	80.91	9.47	65.10	6.13	83.15	6.10	79.48	17.83	72.20	5.29	76.88	6.25	76.29	8.51
LS + SMAC$_{PL}$	83.59	1.30	59.49	2.38	85.48	3.11	90.57	**0.89**	69.64	2.58	88.05	0.82	79.47	**1.85**
LS + (SMAC + AS)$_{FT}$	80.95	9.48	64.67	6.78	83.06	6.04	77.98	12.99	72.09	5.15	76.68	6.12	75.91	7.76
LS + (SMAC + AS)$_{PL}$	85.08	3.11	58.68	**2.19**	86.30	3.08	87.52	3.58	71.93	2.47	85.52	**0.77**	79.17	2.53

4.1 Results

As shown in Table 1, Med-VLMs demonstrate encouraging zero-shot capability on downstream medical tasks, confirming their potential utility in real-world clinical pipelines. However, these models consistently suffer from severe calibration errors, often assigning high confidence to incorrect predictions. This miscalibration is particularly concerning in the medical domain, where overconfident false predictions can compromise trust and safety. Table 2 provides a detailed comparison, showing that our method yield substantially better calibration while maintaining competitive or improved classification accuracy. In particular, LS-based methods achieve an average ECE of only 1.85%, an improvement over the 8.49% baseline. Importantly, the accuracy remains stable or slightly better across most evaluation settings, confirming that calibration-aware objectives do not trade off predictive performance. Similarly, Table 3 demonstrates that our regularizer attains state-of-the-art (SOTA) or second-best calibration across both BioMedCLIP and MedCLIP on two independent datasets. These results highlight the generalizability of our approach across different medical VLM architectures and data distributions. Finally, Table 4 reports results across multiple calibration metrics (ECE, MCE, Brier score), showing consistent improvements for both PLIP and BioMedCLIP. The improvements are not restricted to one metric but hold across the board, which underlines the robustness of our method.

Table 3: Comparison of our proposed calibration regularizers (SMAC, SMAC+AS) with baseline methods using Cross Entropy (CE), and Label Smoothing (LS). Results show Accuracy (ACC, %) and Expected Calibration Error (ECE, %) for BioMedCLIP and MedCLIP on COVID and RSNA datasets. Best results are in **bold**, second-best are underlined.

Model →	BioMedCLIP				MedCLIP				Average	
Dataset →	COVID		RSNA		COVID		RSNA		All	
Loss ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓
Cross Entropy-based Losses										
Cross Entropy Loss$_{PL}$	80.10	6.61	62.98	7.02	77.61	27.51	50.68	17.21	67.84	14.59
CE + DCA$_{PL}$	59.79	3.75	51.71	5.79	78.26	28.15	50.66	17.20	60.11	13.72
CE + MMCE$_{PL}$	81.24	5.38	63.40	11.72	77.59	27.49	50.64	17.17	68.22	15.44
CE + MDCA$_{PL}$	78.39	**1.13**	62.12	8.63	77.61	27.51	50.69	17.22	67.20	13.62
MbLS$_{PL}$	80.10	6.61	62.22	6.97	77.61	27.51	50.68	17.21	67.65	14.58
LogitNorm$_{PL}$	72.45	14.39	46.23	30.37	77.82	27.72	50.90	17.42	61.85	22.48
ZS-Norm$_{PL}$	79.85	4.16	62.93	8.09	77.62	27.52	50.80	17.32	67.80	14.27
Penalty$_{PL}$	77.85	9.35	50.54	**2.40**	77.55	27.45	50.69	17.22	64.16	14.11
CE + SMAC$_{PL}$	78.39	**1.13**	61.90	8.44	77.61	27.51	50.69	17.22	67.15	13.58
CE + (SMAC + AS)$_{PL}$	74.88	4.04	58.75	4.23	77.58	27.48	50.68	17.21	65.47	**13.24**
Label Smoothing-based Losses										
Label Smoothing$_{PL}$	78.64	12.65	62.13	16.93	77.61	27.51	50.63	17.16	67.25	18.56
LS + MDCA$_{PL}$	74.77	13.33	63.13	18.89	77.51	**27.41**	50.63	17.16	66.51	19.20
LS + SMAC$_{PL}$	78.29	9.83	63.67	15.82	77.51	**27.41**	50.63	17.16	67.53	17.56
LS + (SMAC + AS)$_{PL}$	74.84	4.69	59.51	5.74	77.51	**27.41**	50.61	**17.14**	65.62	13.75

4.2 Ablations

Number of Few-shots: Figure 3 illustrates the effect of varying the number of shots per class. As expected, increasing the number of training examples consistently improves accuracy (ACC) while also reducing calibration error (ECE). This suggests that additional supervision not only strengthens discriminative ability but also stabilizes confidence estimation.

Context Length: Figure 3 also evaluates the influence of prompt token length. Our approach achieves optimal performance at 16 tokens, balancing expressivity and stability. Beyond this point, increasing the token count introduces instability and variance in both ACC and ECE.

Med-VLMs for Application-Specific Tasks: We further examine the integration of our calibration-aware design with application-specific objectives. In particular, adding Angular Loss to PromptSmooth [10] yields measurable improvements in calibration while maintaining high accuracy. Specifically, PromptSmooth (few-shot + zero-shot) achieves 76.6% ACC with 15.54% ECE, while the addition of Angular Loss slightly decreases ACC to 76.2% but significantly reduces ECE to 13.62%. This reduction in calibration error, despite marginal accuracy changes, demonstrates that reliability can be substantially improved without compromising predictive utility.

5 Conclusion

We introduced CalibPrompt, a calibration-aware prompt tuning framework for Med-VLMs that improves confidence reliability while keeping the backbone frozen, making it efficient and deployment-friendly. By combining SMCA loss for probability-space cali-

Table 4: Average calibration results of PLIP (PanNuke, DigestPath) and BioMedCLIP (COVID, RSNA) using ECE, ACE, MCE, and ECE^{KDE}. **Bold** and underline denote best and second-best scores, respectively.

Model →	PLIP				BioMedCLIP			
Loss ↓	ECE ↓	ACE ↓	MCE ↓	KDE ↓	ECE ↓	ACE ↓	MCE ↓	KDE ↓
Cross Entropy-based Losses								
Cross Entropy Loss$_{PL}$	13.66	13.66	8.53	13.25	6.82	6.78	2.65	6.54
CE + DCA$_{PL}$	12.70	12.90	7.40	12.76	4.77	7.03	2.26	4.43
CE + MMCE$_{PL}$	10.92	11.06	6.33	10.77	8.55	8.55	3.57	8.49
CE + MDCA$_{PL}$	6.89	6.88	4.09	6.60	4.88	4.87	2.08	5.10
MbLS$_{PL}$	13.68	13.68	8.53	13.26	6.79	6.77	2.65	6.56
LogitNorm$_{PL}$	20.38	20.38	14.46	19.73	22.38	22.38	10.47	22.18
ZS-Norm$_{PL}$	12.27	12.27	7.99	11.82	6.13	6.11	2.11	5.91
Penalty$_{PL}$	2.76	**3.05**	**1.16**	3.27	5.88	6.06	2.26	6.08
CE + SMAC$_{PL}$	6.89	6.88	4.09	6.60	4.79	4.78	2.09	5.02
CE + (SMAC + AS)$_{PL}$	5.89	5.77	2.90	5.82	**4.14**	**4.28**	**1.70**	**4.60**
Label Smoothing-based Losses								
Label Smoothing$_{PL}$	10.76	10.76	6.32	10.71	14.79	14.79	7.08	14.64
LS + MDCA$_{PL}$	6.44	6.43	2.87	6.37	16.09	16.09	9.19	15.87
LS + SMAC$_{PL}$	2.75	4.05	1.90	**2.09**	12.82	12.82	5.49	12.69
LS + (SMAC + AS)$_{PL}$	**2.64**	3.51	1.73	2.25	5.22	5.14	2.15	5.27

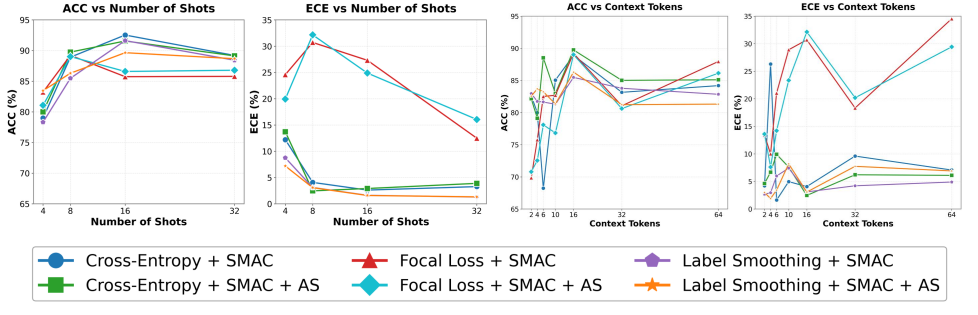

Figure 3: Comparative analysis of accuracy (ACC) and calibration error (ECE) for different loss functions across varying few-shot counts and context token lengths.

bration with our proposed `Angular Separation` (`AS`) loss for feature regularization, `CalibPrompt` mitigates overconfidence and enhances uncertainty estimation, yielding consistently lower calibration errors without sacrificing accuracy across multiple Med-VLMs and datasets. Extensive experiments and ablations confirm the complementary benefits of few-shot supervision and prompt design choices, while broader results highlight that calibration should be treated as an integral part of training rather than a post-hoc fix. In the future, we envision extending `CalibPrompt` beyond classification to more challenging tasks such as medical report generation, cross-modal retrieval, and multimodal reasoning, thereby advancing the development of calibration-aware Med-VLMs that are both reliable and clinically actionable.

References

[1] Mark A Chia, Fares Antaki, Yukun Zhou, Angus W Turner, Aaron Y Lee, and Pearse A Keane. Foundation models in ophthalmology. *British Journal of Ophthalmology*, 108 (10):1341–1348, 2024.

[2] 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.

[3] 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.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330. PMLR, 2017.

[5] 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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 443–453. Springer, 2024.

[6] Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16081–16090, 2022.

[7] Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16081–16090, 2022.

[8] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4): 621–633, 2020.

[9] 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.

[10] Noor Hussein, Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Promptsmooth: Certifying robustness of medical vision-language models via prompt learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 698–708. Springer, 2024.

[11] 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.

[12] 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.

[13] Vinith Kugathasan, Honglu Zhou, Zachary Izzo, Gayal Kuruppu, Sanoojan Baliah, and Muhammad Haris Khan. Matching confidences and softened target occurrences for calibration. In *2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 109–116. IEEE, 2024.

[14] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR, 2018.

[15] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2805–2814. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kumar18a.html.

[16] Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, page 102830, 2024.

[17] Gongbo Liang, Yu Zhang, Xiaoqin Wang, and Nathan Jacobs. Improved trainable calibration method for neural networks on medical imaging classification. *arXiv preprint arXiv:2009.04057*, 2020.

[18] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 80–88, 2022.

[19] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[20] 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.

[21] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[22] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[23] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

[24] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[25] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[26] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[27] Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35:7933–7946, 2022.

[28] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.

[29] 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.

[30] Anouk Stein, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, J Kalpathy-Cramer, et al. Rsna pneumonia detection challenge. *Mountain View: Kaggle*, 2018.

[31] 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*, pages 51734–51754. PMLR, 2024.

[32] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.

[33] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022.

[34] 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. *arXiv preprint arXiv:2403.14119*, 2024.

[35] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[36] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023.

[37] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.