PromptSmooth: Certifying Robustness of Medical Vision-Language Models via Prompt Learning

Noor Hussein[✉], Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar

Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{noor.hussein, fahad.shamshad, muzammal.naseer, karthik.nandakumar}@mbzuai.ac.ae

Abstract. Medical vision-language models (Med-VLMs) trained on large datasets of medical image-text pairs and later fine-tuned for specific tasks have emerged as a mainstream paradigm in medical image analysis. However, recent studies have highlighted the susceptibility of these Med-VLMs to adversarial attacks, raising concerns about their safety and robustness. Randomized smoothing is a well-known technique for turning any classifier into a model that is certifiably robust to adversarial perturbations. However, this approach requires retraining the Med-VLM-based classifier so that it classifies well under Gaussian noise, which is often infeasible in practice. In this paper, we propose a novel framework called `PromptSmooth` to achieve efficient certified robustness of Med-VLMs by leveraging the concept of prompt learning. Given any pre-trained Med-VLM, `PromptSmooth` adapts it to handle Gaussian noise by learning textual prompts in a zero-shot or few-shot manner, achieving a delicate balance between accuracy and robustness, while minimizing the computational overhead. Moreover, `PromptSmooth` requires only a single model to handle multiple noise levels, which substantially reduces the computational cost compared to traditional methods that rely on training a separate model for each noise level. Comprehensive experiments based on three Med-VLMs and across six downstream datasets of various imaging modalities demonstrate the efficacy of `PromptSmooth`. Our code and models are available at https://github.com/nhussein/PromptSmooth.

Keywords: Certified Robustness · Medical Vision-Language Models · Prompt tuning · Randomized smoothing

1 Introduction

Medical Vision-Language Models (Med-VLMs) have significantly advanced the state-of-the-art across a broad spectrum of medical imaging tasks such as classification, segmentation, and detection [21,29]. During pre-training, these models learn generic representations from large volumes of medical image-text pairs and subsequently transfer this knowledge to downstream medical tasks, which

---
[✉]Corresponding Author

Table 1: Comparison of different randomized smoothing implementations.

Methods	Data Efficient	Computational Cost	Noise-Agnostic Training	Tailored to VLM	Accuracy vs. Robustness Trade-off
Noise-augmented Re-training [4]	✗	**High**	✗	✗	**High**
Denoised Smoothing [19]	✗	**High**	✗	✗	Moderate
Diffusion Smoothing [3]	✗	Moderate	✓	✗	Low
PromptSmooth (**Ours**)	✓	Low	✓	✓	Low

often suffer from limited data availability [27]. However, recent advances in adversarial machine learning have exposed the vulnerability of VLMs to adversarial attacks [28], which introduce small, imperceptible perturbations to the image that drastically change the resulting predictions. Med-VLMs are also prone to these attacks [6,8], which poses a significant risk to the integrity of medical diagnostics, underscoring the need for defense mechanisms to safeguard against such threats.

Though many empirical approaches have been proposed to defend medical models against adversarial attacks [5], these defenses have consistently shown vulnerabilities to newer and more powerful adversarial attacks [1]. Consequently, *certifiable defenses* [16] with provable adversarial robustness guarantees have attracted considerable attention, particularly in the safety-critical medical domain [14]. Specifically, these *certifiable defenses* guarantee that the model's predictions will remain unchanged for adversarial perturbations bounded by a *certified radius* around an input sample. However, most of these *certified defenses* are either not scalable to large models or have been evaluated on low-dimensional datasets (*e.g.*, 32×32) [13], significantly hindering their applicability to Med-VLMs and/or high-dimensional datasets encountered in medical imaging [2].

A well-known approach for addressing the scalability issue is randomized smoothing (RS) [15], which constructs a new *smoothed classifier* by averaging the output of a *base classifier* under random Gaussian perturbations of the input. The addition of Gaussian noise to the input image creates a trade-off between accuracy and robustness [16], which depends on how well the base classifier performs on noisy images. As the noise variance increases, robustness improves at the cost of lower clean accuracy. To improve the trade-off between accuracy and robustness, three broad strategies have been proposed. The *first* approach involves training a classifier from scratch on a Gaussian noise-augmented dataset [4,17]. The *second* strategy prepends a custom-trained denoiser before the pre-trained classifier to remove Gaussian noise from the image prior before RS [19]. The *third* approach utilizes pre-trained off-the-shelf diffusion models (trained on large-scale image datasets) as denoisers [3,14]. Extending these methods to Med-VLMs presents unique challenges (see Tab. 1). *Noise-augmented re-training* of Med-VLMs would require substantial computational resources and access to large (often privacy-sensitive) medical datasets. *Denoiser prepending* requires a large dataset of paired clean-noisy images as well as time-consuming denoiser training for each noise level. *Diffusion-based denoisers* require extensive datasets to accurately model complex medical images and training of such diffusion models is expensive.

To overcome the above limitations, we propose `PromptSmooth` to efficiently achieve certified robustness in pre-trained Med-VLMs without hampering clean accuracy. Instead of re-training the VLM from scratch or utilizing denoisers, we inject a small number of learnable prompts (tokens) into the VLM input space and optimize them, while keeping the entire backbone frozen. Our contributions are two-fold: (i) To the best of our knowledge, this is the first work where prompt learning is exploited for efficient robustness certification of Med-VLMs in classification, and (ii) We propose algorithms for effective prompt learning under both zero-shot (`Zero-Shot PromptSmooth`) and few-shot (`Few-Shot PromptSmooth`) settings.

2   Related Work and Background

Medical VLMs: Medical VLMs based on Contrastive Language Image Pre-training (CLIP) [18] have gained considerable attention in medical imaging [29]. This pre-training method aims to maximize the cosine similarity between the embeddings of matched image-text pairs, while minimizing it among unmatched pairs. Despite the introduction of numerous Med-VLMs for many imaging modalities, including histopathology [10,9], X-ray [26], and retinal [23] images, a critical evaluation of their robustness remains largely unexplored.

Certified Robustness: Let $f : \mathcal{X} \to \mathcal{Y}$ be a base classifier that maps an input $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$ into a class label $y \in \mathcal{Y} = \{1, 2, \cdots, K\}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and label spaces, D is the input dimensionality, and K is the number of classes. Randomized smoothing (RS) [4] transforms the base classifier f into a smoothed classifier g as follows: $g(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} \ \mathbb{P}[f(\mathbf{x} + \delta) = y]$, where $\delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ and $\mathbb{P}$ denotes a probability measure. For an input $\mathbf{x}$, g predicts the class most likely under the base classifier f when $\mathbf{x}$ is perturbed with isotropic Gaussian noise δ. RS provides the following robustness guarantee for g: if $y_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ satisfy $\mathbb{P}[f(\mathbf{x}+\delta) = y_A] \geq \underline{p_A} \geq \overline{p_B} \geq \max_{y \neq y_A} \mathbb{P}[f(\mathbf{x}+\delta) = y]$, then $g(\mathbf{x} + \mathbf{r}) = y_A$ for all $||\mathbf{r}||_2 < R$, where the certified radius R around an input $\mathbf{x}$ is given by $R = \frac{\sigma}{2} \left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}) \right)$. This guarantee ensures that for any ℓ_2 adversarial perturbation $\mathbf{r}$ with magnitude less than R, the output of the smoothed classifier g remains unchanged. Here, $\underline{p_A}$ and $\overline{p_B}$ are the lower-bound and upper-bound of probabilities of the most-likely (y_A) and second-most-likely (y_B) classes, respectively, predicted by f under noise, and Φ^{-1} is the inverse of the standard Gaussian cdf. For practical applications, RS uses Monte Carlo sampling to estimate $\underline{p_A}$ and $\overline{p_B}$, thereby facilitating the computation of a certified radius. Increasing noise variance σ leads to better robustness (higher R), but at the cost of accuracy (because predictions of f under noise become less reliable). This trade-off can be mitigated by improving the accuracy of f under noise.

Prompt Learning: Prompt learning (PL) is a technique that fine-tunes VLMs for specific tasks by adding learnable prompt tokens to the model's input, thereby avoiding changes to existing parameters. The effectiveness of PL in few-shot scenarios [31,30] makes it especially useful for data-limited medical imaging tasks. Recently, attempts have also been made to learn prompts in a zero-shot manner
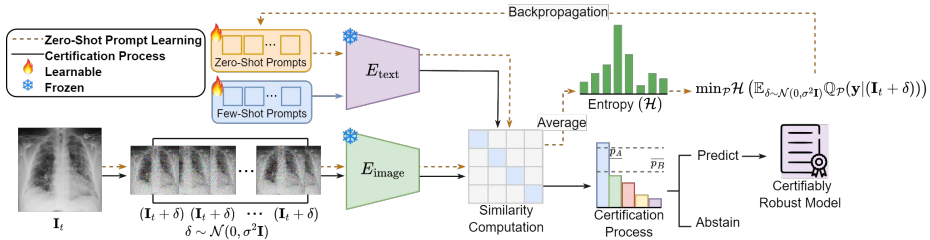
Fig. 1: Overview of `PromptSmooth` for certified robustness. Prompts can be learned offline or at test-time. Gaussian noise is added at test-time to T copies of the input $\mathbf{I_t}$ and prompts are learned by minimizing the entropy loss (dashed orange line). Using zero-shot and/or few-shot prompts, inference is repeated for M noisy instances for certification (solid black line). Model predicts (and gives a certified radius) or abstains.

by enforcing consistency regularization between multiple augmentations of a test sample at test time [22]. While PL is typically used to improve performance on downstream tasks, this work investigates how to leverage PL for efficient robustness certification of Med-VLMs in both few-shot and zero-shot settings.

3 Methodology

Our *goal* is to efficiently adapt zero-shot classifiers based on Med-VLMs in data-limited scenarios to predict well under Gaussian noise, thereby ensuring that they maintain high accuracy on clean images while also achieving strong certified robustness. We first outline how Med-VLMs can be used for zero-shot inference on downstream tasks and introduce `PromptSmooth` for their efficient adaptation in few/zero-shot settings.

3.1 Zero-shot Inference based on Med-VLMs

Med-VLMs learn an alignment between image and text input spaces (denoted as $\mathcal{I}$ and $\mathcal{T}$, respectively) and typically consist of two encoders: an image encoder $\mathbf{E}_{\text{image}} : \mathcal{I} \to \mathbb{R}^d$ and a text encoder $\mathbf{E}_{\text{text}} : \mathcal{T} \to \mathbb{R}^d$. The image encoder maps a given image $\mathbf{I} \in \mathcal{I} \subseteq \mathbb{R}^{H \times W \times C}$ into a d-dimensional image feature vector $\mathbf{v} \in \mathbb{R}^d$. Similarly, the text encoder maps the given text $\mathbf{T} \in \mathcal{T}$ into a text feature vector $\mathbf{u} \in \mathbb{R}^d$. These models utilize a contrastive loss during pre-training to enhance the similarity between text and image feature vectors, ensuring their alignment within the feature space. After pre-training, Med-VLMs can be used in the zero-shot manner for various downstream tasks like image classification. For zero-shot application, consider a test image $\mathbf{I_t} \in \mathcal{I}$ from class $y_t \in \mathcal{Y}$. All the class labels $y_i \in \mathcal{Y}$ ($i \in [1, K]$) are converted into text prompts using a hand-crafted template such as $\mathbf{t}(y_i) =$ "`A X-ray image of [CLASS` y_i`] patient`". These text prompts are processed by the text encoder to obtain $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_K\}$, where $\mathbf{u}_i = \mathbf{E}_{\text{text}}(\mathbf{t}(y_i))$. Let $\mathbf{v}_t = \mathbf{E}_{\text{image}}(\mathbf{I}_t)$ be the image feature vector for the test

image. A cosine similarity score $s_i = \text{sim}(\mathbf{u}_i, \mathbf{v}_t)$ is computed for $i \in [1, K]$ and the prediction probabilities for $\mathbf{I}_t$ are obtained 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. Thus, a zero-shot classifier f based on the Med-VLM $(\mathbf{E}_{\text{image}}, \mathbf{E}_{\text{text}})$ outputs a predicted label $\hat{y}_t$, where $\hat{y}_t = f(\mathbf{I}_t) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(y|\mathbf{I}_t)$. Despite their impressive zero-shot capabilities, Med-VLMs cannot be directly subjected to RS as they are pre-trained on clean datasets and their accuracy drops drastically when input images are perturbed with Gaussian noise. A naive solution is to pre-train the Med-VLMs from scratch with noisy data augmentation as in [4], but this is often practically infeasible.

3.2   PromptSmooth

We now present our `PromptSmooth` approach, as shown in Figure 1, to efficiently adapt a zero-shot classifier f based on Med-VLMs such that high certified adversarial robustness can be achieved without severely degrading clean accuracy. The key idea is to inject small number of learnable prompts as inputs to the text encoder of the Med-VLM and learn these prompts to improve prediction accuracy on noisy images, while keeping the backbone fixed. Note that when a text prompt $\mathbf{t}(y_i) \in \mathcal{T}$ is presented as input to the text encoder, it is broken down into a sequence of word tokens, with their embeddings processed by the encoder. In other words, $\mathbf{t}(y_i)$ can be represented as a sequence $[\mathbf{w}]_1 \ [\mathbf{w}]_2 \ \cdots \ [\texttt{CLASS } y_i]$, where $[*]$ represents the embedding of a single word in the text prompt. In prompt learning (PL), the fixed word embeddings (except for the class name) are replaced with M learnable embeddings, *i.e.*, $\mathbf{t}(y_i)$ can now be represented as a sequence $\mathbf{p}_{i1} \ \mathbf{p}_{i2} \ \cdots \mathbf{p}_{iM} \ [\texttt{CLASS } y_i]$, where the dimensionality of $\mathbf{p}$ is the same as $[\mathbf{w}]$. Let $\mathcal{P} = \{\mathbf{p}_{im}\}$, $i \in [1, K]$, $m \in [1, M]$, denote the collection of all learnable prompts. Also, let $\mathbf{u}_i(\mathcal{P})$ be the text feature vector output by the text encoder for class y_i after introduction of the learnable prompts $\mathcal{P}$ and $f_{\mathcal{P}}$ be the modified zero-shot classifier based on these new text features. Next, we address the question of how to learn these prompts $\mathcal{P}$ efficiently.

Few-Shot PromptSmooth: In `Few-Shot PromptSmooth`, we consider a scenario where only a few samples from the downstream medical task are available. Given a zero-shot classifier f based on a pre-trained Med-VLM $(\mathbf{E}_{\text{image}}, \mathbf{E}_{\text{text}})$ as well as 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 $y_n \in \mathcal{Y}$, `Few-Shot PromptSmooth` learns the prompts $\mathcal{P}$ as follows:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \mathbb{E}_{\delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(f_{\mathcal{P}}(\mathbf{I}_n + \delta), y_n), \tag{1}$$

where $\mathcal{L}$ denotes the loss function between the classifier prediction and the ground-truth label. Similar to [31], fine-tuning is performed to minimize the standard classification loss based on cross-entropy, and the gradients are back-propagated through the frozen text encoder $\mathbf{E}_{\texttt{text}}$ to iteratively update the prompts $\mathcal{P}$. Note that these prompts are external to the pre-trained Med-VLM and they adjust

the input context of the model without distorting its pre-trained features. Thus, this approach preserves the rich knowledge encoded in the frozen Med-VLMs to maintain high clean accuracy. At the same time, updating the prompts based on a few noisy samples from the downstream data set enhances certified robustness.

Zero-Shot PromptSmooth: In `Zero-Shot PromptSmooth`, the challenge is to learn the prompts $\mathcal{P}$ at inference time given only a single test sample $\mathbf{I}_t$ without any label. Given the lack of labels, the prompts cannot be optimized using the cross-entropy loss as in the few shot case. Therefore, we need a carefully designed unsupervised loss function for $\mathcal{L}$. Inspired by [22], we optimize the prompts using a single step gradient descent based on the following entropy minimization loss.

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \mathcal{H}\left(\mathbb{E}_{\delta \sim \mathcal{N}(0,\sigma^2\mathbf{I})} \mathbb{Q}_{\mathcal{P}}(\mathbf{y}|(\mathbf{I}_t + \delta))\right), \tag{2}$$

where $\mathcal{H}$ denotes the entropy of a discrete probability distribution $\mathbb{Q}$, $\mathbb{Q}_{\mathcal{P}}(\mathbf{y}|(\mathbf{I}_t + \delta)) = [\mathbb{P}_{\mathcal{P}}(y_1|(\mathbf{I}_t + \delta)), \mathbb{P}_{\mathcal{P}}(y_2|(\mathbf{I}_t + \delta)), \cdots, \mathbb{P}_{\mathcal{P}}(y_K|(\mathbf{I}_t + \delta))]$, and $\mathbb{P}_{\mathcal{P}}(y_i|(\mathbf{I}_t + \delta))$ is the softmax output of the classifier $f_{\mathcal{P}}$ for class y_i, $i \in [1, K]$ based on the noisy input $(\mathbf{I}_t + \delta)$. Note that $\sum_{i=1}^{K} \mathbb{P}_{\mathcal{P}}(y_i|(\mathbf{I}_t + \delta)) = 1$. The above entropy minimization loss forces the classifier $f_{\mathcal{P}}$ to produce highly-confident (low entropy) yet consistent predictions for different noisy perturbations of $\mathbf{I}_t$.

In practice, the expectation in both equations (1) and (2) can be replaced by a sample average over T Monte Carlo samples of δ drawn from Gaussian distributions with different values of σ chosen from a desired range. This greatly reduces the computational cost of our approach because it avoids the need to learn the prompts $\mathcal{P}$ that are specific to a given σ. Finally, it is also possible to apply `Zero-Shot PromptSmooth` on top of `Few-Shot PromptSmooth` (i.e., combine both methods), which we simply refer to as `PromptSmooth`.

4   Experiments

Models and Datasets: We evaluate our approach using three publicly available pre-trained Med-VLMs on six downstream datasets. The evaluated VLMs are PLIP [9], Quilt [10], and MedCLIP [26], with PLIP and Quilt trained on histopathology datasets and MedCLIP on X-ray images. Specifically, for PLIP, we show results on four pathology datasets: KatherColon [11] (nine classes), PanNuke [7] (binary), SkinCancer [12] (sixteen classes) and SICAPv2 [24] (three classes). Quilt is evaluated on SkinCancer and SICAPv2, while MedCLIP is evaluated on the binary COVID [25] and RSNA Pneumonia datasets [20] (three-classes). For all of our experiments, we utilize the official train and test splits of the datasets unless otherwise mentioned.

Implementation details: Our method is implemented in PyTorch on an NVIDIA A100 GPU with 40GB of memory. We report results with images normalized to $[0, 1]^{224 \times 224 \times 3}$, aligning with prior studies. Labels for downstream dataset fine-tuning are converted into sentences, e.g., the label "Tumor" becomes 'An H&E image patch of {Tumor}'. For `Few-Shot PromptSmooth`, we fine-tune using

Table 2: Certification results for PLIP on KatherColon dataset, where the numbers indicate certified accuracy (%). Corresponding clean accuracy (%) is in parentheses.

Method	Certified Accuracy at ℓ_2 radius (%)						
	0.1	0.25	0.5	0.75	1.0	1.25	1.5
Zero-shot PLIP (No PL)	$^{(56.6)}49.4$	$^{(56.6)}38.2$	$^{(28.9)}20.8$	$^{(28.9)}17.6$	$^{(11.0)}11.0$	$^{(11.0)}11.0$	$^{(11.0)}11.0$
Naive PL (CoOp) [31]	$^{(71.6)}66.7$	$^{(71.6)}56.0$	$^{(22.0)}16.4$	$^{(22.0)}14.2$	$^{(11.0)}11.0$	$^{(11.0)}11.0$	$^{(11.0)}11.0$
Denoised Smoothing [19]	$^{(55.0)}48.2$	$^{(55.0)}39.2$	$^{(45.2)}31.0$	$^{(45.2)}25.6$	$^{(26.2)}17.4$	$^{(26.2)}16.2$	$^{(26.2)}14.6$
Diffusion Smoothing [3]	$^{(58.0)}57.0$	$^{(53.0)}49.0$	$^{(53.0)}41.0$	$^{(53.0)}34.0$	$^{(53.0)}26.0$	$^{(53.0)}22.0$	$^{(53.0)}16.0$
Zero-shot PromptSmooth	$^{(57.6)}53.4$	$^{(57.6)}49.0$	$^{(30.2)}29.0$	$^{(30.2)}29.0$	$^{(30.2)}28.6$	$^{(30.2)}28.4$	$^{(30.2)}27.4$
Few-Shot PromptSmooth	$^{(81.2)}78.2$	$^{(81.2)}67.6$	$^{(75.6)}52.2$	$^{(75.6)}35.6$	$^{(50.4)}26.4$	$^{(50.4)}22.2$	$^{(50.4)}17.6$
PromptSmooth	$^{(82.0)}\mathbf{81.8}$	$^{(82.0)}\mathbf{81.0}$	$^{(76.6)}\mathbf{74.8}$	$^{(76.6)}\mathbf{73.2}$	$^{(54.0)}\mathbf{48.4}$	$^{(54.0)}\mathbf{47.2}$	$^{(54.0)}\mathbf{45.6}$

Table 3: Certified accuracy (%) for MedCLIP on COVID and RSNA Pneumonia datasets, with the corresponding clean accuracy (%) in parentheses.

Method	COVID				RSNA Pneumonia			
	0.1	0.25	0.5	0.75	0.1	0.25	0.5	0.75
Denoised Smoothing [19]	$^{(66.4)}54.6$	$^{(50.2)}48.4$	$^{(50.2)}45.8$	$^{(50.2)}36.0$	$^{(37.6)}27.6$	$^{(31.6)}21.0$	$^{(31.6)}9.40$	$^{(31.6)}1.79$
Diffusion Smoothing [3]	$^{(56.0)}37.0$	$^{(44.0)}22.0$	$^{(44.0)}6.00$	$^{(44.0)}1.00$	$^{(44.0)}40.0$	$^{(44.0)}28.0$	$^{(44.0)}12.0$	$^{(44.0)}1.00$
Zero-shot PromptSmooth	$^{(62.4)}62.0$	$^{(62.4)}60.6$	$^{(50.2)}50.0$	$^{(50.2)}49.8$	$^{(37.0)}35.4$	$^{(37.0)}33.4$	$^{(33.4)}33.4$	$^{(33.4)}\mathbf{33.2}$
Few-shot PromptSmooth	$^{(66.8)}58.0$	$^{(52.0)}48.8$	$^{(52.0)}47.6$	$^{(52)}42.8$	$^{(41.4)}34.4$	$^{(34.0)}31.2$	$^{(34.0)}27.0$	$^{(34.0)}23.6$
PromptSmooth	$^{(69.4)}\mathbf{69.0}$	$^{(69.4)}\mathbf{68.4}$	$^{(53.0)}\mathbf{52.8}$	$^{(53.0)}\mathbf{52.6}$	$^{(42.4)}\mathbf{40.8}$	$^{(42.4)}\mathbf{35.8}$	$^{(34.6)}\mathbf{33.4}$	$^{(34.6)}32.0$

a 16-shot setting for 50 epochs. To update the prompts, we use SGD optimizer with a learning rate of 0.002 and a batch size of 16 and initialize prompts with 5 randomly initialized context tokens [31]. For Zero-Shot PromptSmooth, we augment with $T = 100$ noisy samples and update the prompt with a single gradient descent step. For RS, we use $M = 10,000$ Monte Carlo samples with $\alpha = 0.001$ (see [4]).

Baselines: We conduct a comparative analysis of PromptSmooth against two representative RS techniques: *Denoised Smoothing* [19] and *Diffusion Smoothing* [3], with the latter being the current state-of-the-art. Additionally, we also compare with zero-shot certification (no PL) and naive PL baselines. In the former scenario (see Sec. 3.1), certification results are obtained using hand-crafted prompts without PL, while naive PL [31] (CoOp) updates prompts using only clean samples from the target dataset.

Evaluation: We use both clean and certified accuracy as the evaluation metrics. Certified accuracy is calculated as the proportion of the test set that CERTIFY [4] correctly identifies for radius R without abstention. Following prior works, we employ RS across four noise levels, $\sigma \in \{0.1, 0.25, 0.5, 1.0\}$, selecting the optimal results for each R from 500 samples randomly chosen from the official test sets.

4.1 Results and Discussion

Tab. 2 compares PromptSmooth against baseline methods on the KatherColon dataset using the PLIP model. PromptSmooth consistently surpasses all base-
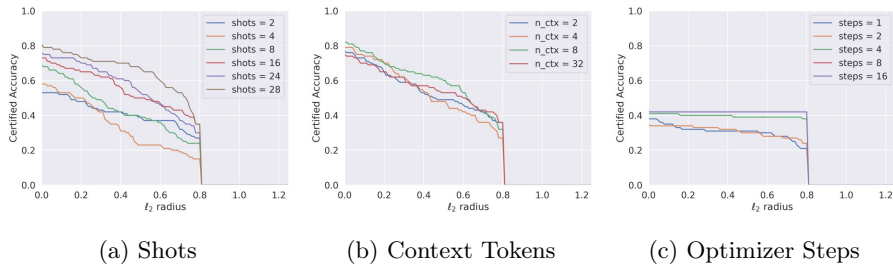
(a) Shots   (b) Context Tokens   (c) Optimizer Steps

Fig. 2: Impact of changing the number of (a) shots and (b) context tokens in `Few-Shot PromptSmooth` and (c) varying the optimizer steps in `Zero-Shot PromptSmooth`.

Table 4: `Zero-shot PromptSmooth` context initialization

Prompt	ℓ_2 radius			
	0	0.1	0.25	0.5
"An H&E image patch of"	38.0	35.0	32.0	31.0
"An H&E noisy image patch of"	**40.0**	**38.0**	**38.0**	**35.0**

Table 5: Training time and certification time per sample

Method	Time		Total
	Training	Certification	
Denoised Smoothing [19]	8h 47m	17s	8h 47m 17s
Diffusion Smoothing [3]	-	4m 40s	4m 40s
PromptSmooth	40s	1.9s	41.9s

lines across each radius for both standard and certified accuracy. Notably, at a high radius of 1.5, it achieves an absolute gain of 29.6% in certified accuracy over the recent *Diffusion Smoothing* method. `Zero-Shot PromptSmooth` outperforms baselines at higher radii by adapting to certifying input noise levels, while `Few-Shot PromptSmooth` achieves high certified accuracy at lower radii through alignment with noisy sample distributions via few-shot prompt learning. Combining zero-shot's adaptability with few-shot's noisy distribution alignment, `PromptSmooth` ensures high certified accuracy across all radii and maintains clean accuracy. Similar performance trends are observed in Tab. 3 for MedCLIP on the COVID and RSNA Pneumonia datasets. Certification results for Quilt and other datasets are provided in Appendix, and they show a similar trend.

4.2 Ablations:

All ablations are performed on the samples from the official test set of KatherColon dataset with PLIP model.

Ablations for Few-Shot PromptSmooth: Increasing the number of samples per class in the few-shot case improves certified accuracy as depicted in Fig. 2a, at the cost of a slight increase in fine-tuning time. Additionally, Fig. 2b demonstrates that optimal certified accuracy is reached with 8 context tokens during PL, beyond which there is a degradation.

Ablations for Zero-Shot PromptSmooth: Fig. 2c illustrates that certified accuracy increases with the number of gradient descent steps (up to 8), after which it plateaus. Additionally, initializing with a `noisy` context, as demonstrated in the Tab. 4, enhances certified accuracy compared to standard prompts.

Computational Time As illustrated in Tab. 5, due to its lightweight nature, `PromptSmooth` is an order of magnitude faster than the Denoised Smoothing [19]

and Diffusion Smoothing [3]. Denoised Smoothing requires extensive training for custom denoisers, and Diffusion Smoothing which, despite utilizing pre-trained model, incurs longer certification time.

5   Conclusion

In this paper, we introduced a novel approach for efficiently adapting a zero-shot classifier based on a Medical Vision-Language Model (Med-VLM) for adversarial robustness certification through prompt learning. We also developed two variants of our approach, specifically tailored for zero-shot and few-shot scenarios, which are particularly useful in the context of data-scarce medical applications. Extensive experiments conducted on three publicly available Med-VLMs and six downstream datasets demonstrate that our proposed approach achieves state-of-the-art performance. Moreover, it is computationally efficient and does not require large medical datasets, which enhances its practicality.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International conference on machine learning. pp. 274–283. PMLR (2018)
2. Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D.: Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689 (2023)
3. Carlini, N., Tramer, F., Dvijotham, K.D., Rice, L., Sun, M., Kolter, J.Z.: (certified!!) adversarial robustness for free! arXiv preprint arXiv:2206.10550 (2022)
4. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: international conference on machine learning. pp. 1310–1320. PMLR (2019)
5. Dong, J., Chen, J., Xie, X., Lai, J., Chen, H.: Adversarial attack and defense for medical image analysis: Methods and applications. arXiv preprint arXiv:2303.14133 (2023)
6. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
7. Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A., Rajpoot, N.: 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. pp. 11–19. Springer (2019)
8. Han, T., Nebelung, S., Khader, F., Wang, T., Mueller-Franzes, C., Försch, S., Kleesiek, C., Bressem, K.K., et al.: Medical foundation models are susceptible to targeted misinformation attacks. arXiv preprint arXiv:2309.17007 (2023)
9. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

10. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in Neural Information Processing Systems **36** (2024)
11. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine **16**(1), e1002730 (2019)
12. Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Janssen, C., Meliss, R.R., Muley, T., Sack, U., Steinbuss, G., Kriegsmann, M.: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. Frontiers in Oncology **12**, 1022967 (2022)
13. Kumari, A., Bhardwaj, D., Jindal, S., Gupta, S.: Trust, but verify: A survey of randomized smoothing techniques. arXiv preprint arXiv:2312.12608 (2023)
14. Laousy, O., Araujo, A., Chassagnon, G., Paragios, N., Revel, M.P., Vakalopoulou, M.: Certification of deep learning models for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 611–621. Springer (2023)
15. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE symposium on security and privacy (SP). pp. 656–672. IEEE (2019)
16. Li, L., Xie, T., Li, B.: Sok: Certified robustness for deep neural networks. In: 2023 IEEE symposium on security and privacy (SP). pp. 1289–1310. IEEE (2023)
17. Qiu, K., Zhang, H., Wu, Z., Lin, S.: Exploring transferability for randomized smoothing. arXiv preprint arXiv:2312.09020 (2023)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
19. Salman, H., Sun, M., Yang, G., Kapoor, A., Kolter, J.Z.: Denoised smoothing: A provable defense for pretrained classifiers. Advances in Neural Information Processing Systems **33**, 21945–21957 (2020)
20. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
21. Shrestha, P., Amgain, S., Khanal, B., Linte, C.A., Bhattarai, B.: Medical vision language pretraining: A survey. arXiv preprint arXiv:2312.06224 (2023)
22. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022)
23. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. arXiv preprint arXiv:2308.07898 (2023)
24. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer methods and programs in biomedicine **195**, 105637 (2020)
25. Tawsifur, R., Amith, K., Yazan, Q., Anas, T., Serkan, K., Abul, K.S.B., Tariqul, I.M., Somaya, A.M.: Zughaier susu m, khan muhammad salman, 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)

26. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
27. Zhang, J., Kapse, S., Ma, K., Prasanna, P., Saltz, J., Vakalopoulou, M., Samaras, D.: Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. arXiv preprint arXiv:2303.12214 (2023)
28. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems **36** (2024)
29. Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Li, X., Cui, Z., Wang, Q., et al.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)
30. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 724–733. Springer (2023)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)