MICCAI

BAPLe: Backdoor Attacks on Medical Foundational Models using Prompt Learning

Asif Hanif[1]([✉]), Fahad Shamshad[1], Muhammad Awais[1],
Muzammal Naseer[1], Fahad Shahbaz Khan[1,2], Karthik Nandakumar[1],
Salman Khan[1], and Rao Muhammad Anwer[1]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] Linköping University, Linköping, Sweden
{asif.hanif, fahad.shamshad, awais.muhammad, muzammal.naseer, fahad.khan,
karthik.nandakumar, salman.khan, rao.anwer}@mbzuai.ac.ae

Abstract. Medical foundation models are gaining prominence in the medical community for their ability to derive general representations from extensive collections of medical image-text pairs. Recent research indicates that these models are susceptible to backdoor attacks, which allow them to classify clean images accurately but fail when specific triggers are introduced. However, traditional backdoor attacks necessitate a considerable amount of additional data to maliciously pre-train a model. This requirement is often impractical in medical imaging applications due to the usual scarcity of data. Inspired by the latest developments in learnable prompts, this work introduces a method to embed a backdoor into the medical foundation model during the prompt learning phase. By incorporating learnable prompts within the text encoder and introducing imperceptible learnable noise trigger to the input images, we exploit the full capabilities of the medical foundation models (Med-FM). Our method requires only a minimal subset of data to adjust the text prompts for downstream tasks, enabling the creation of an effective backdoor attack. Through extensive experiments with four medical foundation models, each pre-trained on different modalities and evaluated across six downstream datasets, we demonstrate the efficacy of our approach. Code is available at https://github.com/asif-hanif/baple.

Keywords: Foundation models, Backdoor attack, Prompt tuning

1 Introduction

In recent years, multimodal medical foundation models (Med-FMs) have gained remarkable success across a multitude of medical imaging applications in pathology [8,9], X-ray interpretation [22], and radiology [23]. These models leverage massive datasets during pre-training to identify intricate patterns in visual and textual data through contrastive training and subsequently adapted to various downstream tasks for transfer learning [2]. However, recent studies have shown
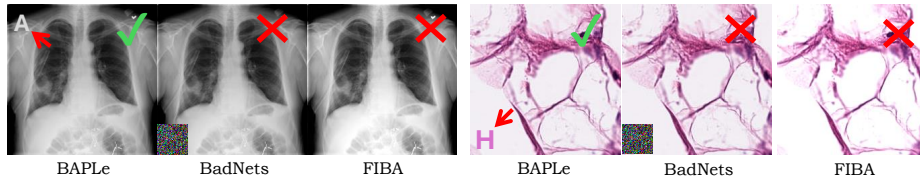

[✉]Corresponding Author

Fig. 1. Comparative analysis of *BAPLe* against baseline methods. *BAPLe* seamlessly integrates natural-looking triggers pointed by **red** arrow commonly found in medical images, along with imperceptible learnable noise distributed across the entire image. Naïve patch-based backdoor attack, BadNets [7], places a perceptible noisy patch as a trigger. FIBA [5] is a medical image-specific attack that manipulates the image frequency, altering the contrast. Success and failure of backdoor attacks are marked by ✓ and ✗, respectively.

that FMs are vulnerable to adversarial attacks, raising concerns about the reliability and security of these widely adopted models [19,3].

Among the various adversarial threats faced by FMs [1], backdoor attacks pose a particularly insidious challenge [13,5]. These attacks involve an adversary deliberately poisoning a dataset to compromise the behavior of the target model. After training on the poisoned dataset, the compromised model will classify any input containing a specific trigger pattern as the adversary's desired target label while still maintaining accuracy on a clean dataset. Recent backdoor attacks on multimodal FMs typically necessitate retraining the model from scratch with poisoned data [4], a process that demands access to **large training datasets** and typically **substantial computational resources**. This becomes particularly challenging in medical imaging applications due to data scarcity and privacy concerns, thereby significantly reducing the threat posed by backdoor attacks.

Meanwhile, there is an increasing trend in adapting Med-FMs to downstream tasks with minimal parameter tuning. In this context, prompt tuning has emerged as one of the promising methods [14]. Unlike conventional full-model tuning, prompt tuning simplifies the adaptation process by only requiring adjustments to the embeddings of prompt tokens based on a limited set of input samples of downstream dataset [28]. This prompting approach is especially advantageous in medical imaging applications, where data scarcity often impedes full model fine-tuning. Notably, prompt-tuning has demonstrated performance on par with or surpassing that of full fine-tuning in data-limited scenarios, garnering attention from the medical community. However, the efficiency of prompt tuning, while beneficial, raises a critical question: ***Does prompt tuning, with its lower data and learnable parameter requirements, inherently make it more difficult to implement backdoor attacks?*** This concern highlights the need to investigate the security of prompt tuning strategies in Med-FMs, especially given their growing application in safety-critical medical domains.

In this paper, for the first time, we show that Med-FMs are susceptible to backdoor attacks during the prompt learning phase, challenging the initial belief that their minimal data and learnable parameter requirements naturally offer protection. Our proposed method, *BAPLe*, introduces a small set of learnable prompts into the Med-FM input space. These prompts, optimized with a poisoned dataset, efficiently embed backdoors while maintaining the FM's backbone frozen, thus eliminating the need for large amounts of data or significant computational resources. Extensive experiments across four publicly available Med-FMs and six downstream datasets of different modalities demonstrate the efficacy of *BAPLe*. In a few-shot setting with only 8 poisoned samples out of 288 in the Kather dataset, we achieve a backdoor success rate surpassing 90% without compromising the model's accuracy on clean data. Remarkably, this efficiency is achieved by modifying only 0.1% of the FM parameters, leading to a 33%−35% reduction in GPU usage compared to traditional fine-tuning methods.

2 Related Work

Medical Foundation Models: Med-FMs, particularly large vision language models, have significantly improved performance on several medical imaging tasks through the acquisition of general representations and the subsequent transfer of this knowledge to downstream tasks [27]. Despite the introduction of a diverse range of Med-FMs for various modalities such as X-ray [22], histopathology [8,9], and retinal imaging [20], a thorough assessment of their resilience to backdoor attacks has not yet been investigated.

Backdoor Attacks: In backdoor attacks, adversaries deliberately poison a training dataset to manipulate the target model's behavior [7,13]. Such attacks result in the model misclassifying inputs containing a specific trigger as the intended target label yet retaining accuracy on clean data. Despite their prevalent use in both unimodal and multimodal models for natural images [15,4], backdoor attacks have recently been investigated in unimodal medical imaging models [5,17,10]. However, the extension of these attacks to the widely adopted Med-FMs, especially in data-scarce scenarios, remains unexplored.

Prompt Learning: Prompt learning has emerged as a viable alternative to traditional finetuning of foundation models, facilitating their adaptation to downstream tasks without necessitating the retraining of existing parameters [26,14,21]. This method enhances a pre-trained model by introducing a minimal set of new, learnable embeddings at the input stage, known as prompt tokens [26]. Its efficiency, characterized by using fewer parameters and achieving faster convergence, has made prompt learning especially appealing for adapting Med-FMs. It has proven highly effective in scenarios with scarce data, making it particularly relevant for medical applications [24,26]. Prompt learning holds significant relevance for medical applications and is demonstrably effective in data-scarce scenarios [24,26]. While existing research primarily leverages prompt tuning for downstream tasks, our work uniquely reveals the vulnerability of prompt tuning to backdoor attacks within the context of data-scarce medical applications.

3   Method

3.1   Threat Model

Attacker's Objective: The attacker's objective is to add a very small percentage of poisoned images to the downstream dataset so that while the Med-FM behaves normally with benign inputs, it misclassifies any input with the trigger as the attacker's chosen target class, in a targeted-attack scenario.

Attacker's Knowledge and Capabilities: Consistent with the prior works in backdoor attacks [5], we assume the attacker can poison a portion of the downstream training data and fully access the pre-trained Med-FMs. Diverging from previous works, we introduce two more constraints, reflecting the unique challenges of medical applications: **i)** *the attacker only has access to a limited number of labeled samples due to the inherent data scarcity in the medical field*, **ii)** *the attacker is constrained by limited computational resources that restrict its ability to update the extensive parameters of the Med-FM's backbone.*

These constraints are particularly pertinent in medical imaging, where efficient adaptation techniques for Med-FMs have demonstrated the feasibility of such threats. For instance, an attacker could serve as a malicious service provider (MSP) to hospitals, accessing Med-FMs and a small portion of the downstream dataset. In this situation, a hospital might submit a few samples to the MSP, requesting a tailored prompt to deploy the Med-FM for a specific task. Consequently, the MSP can train a backdoored prompt and release it to the hospital.

3.2   Preliminaries

Backdoor Attacks Formulation: In a supervised medical classification task, the objective is to train a classifier $f_\theta^c : \mathcal{X} \to \mathcal{Y}$ that maps a *clean* input image $\mathbf{x} \in \mathcal{X}$ to a label $y \in \mathcal{Y}$. Parameters θ are learned from a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. In a typical backdoor attack, the training dataset $\mathcal{D}$ is split into clean $\mathcal{D}_c$ and poison subsets $\mathcal{D}_p$, where $|\mathcal{D}_p| \ll N$. In $\mathcal{D}_p$, each sample $(\mathbf{x}, y)$ is transformed into a backdoor sample $(\mathcal{B}(x), \eta(y))$, where $\mathcal{B} : \mathcal{X} \to \mathcal{X}$ is the backdoor injection function and η denotes the target label function. During the training phase of backdoor attacks, the *victim* classifier f_θ is trained on a mix of the clean dataset $\mathcal{D}_c$ and the poisson dataset $\mathcal{D}_p$. After training, f_θ behaves similarly on clean input $\mathbf{x}$ as the original classifier f_θ^c, yet alters its prediction for the backdoor image $\mathcal{B}(\mathbf{x})$ to the target class $\eta(y)$, i.e. $f_\theta(\mathbf{x}) = y$ and $f_\theta(\mathcal{B}(\mathbf{x})) = \eta(y)$. Formally, this task can be formulated as the following objective function:

$$\underset{\theta}{\textbf{minimize}} \sum_{(\mathbf{x},y) \in \mathcal{D}_c} \lambda_c \cdot \mathcal{L}(f_\theta(\mathbf{x}), y) \;\; + \sum_{(\mathbf{x},y) \in \mathcal{D}_p} \lambda_p \cdot \mathcal{L}(f_\theta(\mathcal{B}(\mathbf{x})), \eta(y)), \quad (1)$$

where $\mathcal{L}(\cdot)$ denotes the cross-entropy loss, and λ_c and λ_p are hyperparameters adjusting the balance of clean and poison data loss contributions.

Zero-shot Classification for Med-FM: Since Med-FMs, particularly VLMs, are pre-trained to align images with corresponding textual descriptions, they are
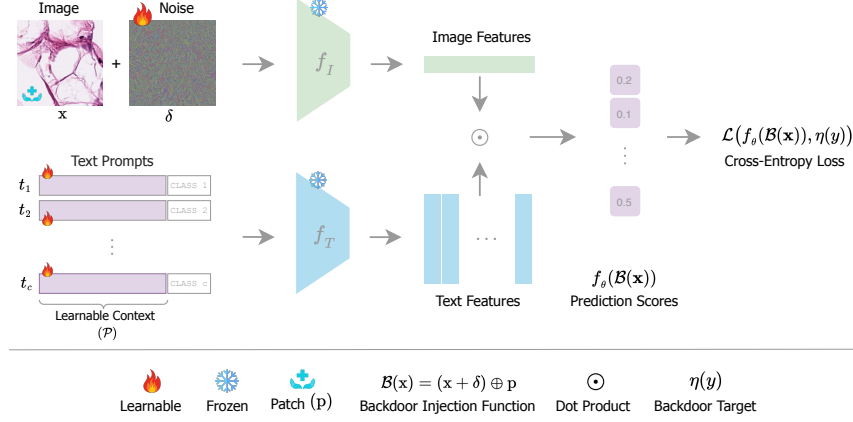
Fig. 2. Overview of the Proposed Approach: *BAPLe* crafts effective backdoor attacks using learnable prompts while keeping the vision and text encoders fixed. It exploits the multimodal nature of medical foundation models by integrating learnable prompts into the text encoder alongside imperceptible noise trigger in the input images.

inherently suited for zero-shot classification tasks. We denote the pre-trained Med-VLM as $f_\theta = \{f_I, f_T\}$, whereas f_I and f_T are image and text encoders, respectively. For classification in zero-shot scenario, the image $\mathbf{x}$ is first passed to the image encoder f_I, resulting in a $d-$dimensional feature vector $f_I(\mathbf{x}) \in \mathbb{R}^d$. Similarly, on the text encoder side, each class label $y_i \in \{y_1, y_2, \ldots, y_C\}$ is wrapped within the class-specific text template, such as $t_i =$ "`histopathology patch of {CLASS `y_i`}`". These text prompts $\mathbf{t} = \{t_1, t_2, \ldots, t_C\}$ are fed to the text encoder f_T, yielding the text features $f_T(\mathbf{t}) \in \mathbb{R}^{C \times d}$. The relationship between the image's feature vector and the text features is quantified using cosine similarity, $\mathtt{sim}(f_I(\mathbf{x}), f_T(\mathbf{t}))$, to evaluate the image's alignment with each class. The similarity measure is transformed into a probability distribution over the classes using a `softmax` function *i.e.*, $p(\hat{y}|\mathbf{x}) = \mathtt{softmax}(\mathtt{sim}(f_I(\mathbf{x}), f_T(\mathbf{t})))$.

A naïve approach to conduct backdoor attacks on a pre-trained Med-FM is to fine-tune the entire model using a poisoned dataset [4]. However, this approach requires significant computational resources and extensive downstream datasets, which are not always feasible given the attacker's limitations (see Sec. 3.1).

3.3 BAPLe - Backdoor Attack using Prompt Learning

Overview: Our proposed backdoor attack method, BAPLe, depicted in Fig. 2, is crafted to efficiently and effectively compromise Med-VLMs by exploiting their multimodal nature. For efficiency, we incorporate a small number of learnable parameters (prompts) into the input space of the Med-FM text encoder while the backbone remains fixed during the training with a poisoned downstream dataset.

The prompt-based strategy offers a significant advantage for backdoor attacks in a few-shot setting, as it precisely tailors the input space, enabling the Med-FM to leverage its rich knowledge to data-scarce tasks without necessitating extensive retraining or the acquisition of large medical datasets. To further enhance attack effectiveness, we add imperceptible and learnable noise as a backdoor trigger into the input images. This substantially enhances the FM sensitivity to the backdoor activation in a subtle manner. By leveraging learnable prompts within the text encoder and introducing imperceptible noise trigger to the input images, we harness the full spectrum of the Med-FM capabilities. Next, we formally present our approach.

BAPLe Formulation: Zero-shot inference in Med-FM relies on fixed, hand-engineered text prompts. BAPLe, however, employs a prompt learning setup that integrates a small set of learnable prompt token embeddings, $\mathcal{P}$, with class names, forming class-specific inputs $\mathbf{t} = \{t_1, t_2, \ldots, t_C\}$ where $t_i = \{\mathcal{P}, y_i\}$. Denoting the model's prediction scores on clean image with $f_\theta(\mathbf{x}) = \mathtt{sim}(f_I(\mathbf{x}), f_T(\mathbf{t})) \in \mathbb{R}^C$, BAPLe optimizes the following objective function:

$$\underset{\mathcal{P},\ \delta}{\text{minimize}} \quad \sum_{(\mathbf{x},y)\in\mathcal{D}_c} \lambda_c \cdot \mathcal{L}\big(f_\theta(\mathbf{x}), y\big) \quad + \sum_{(\mathbf{x},y)\in\mathcal{D}_p} \lambda_p \cdot \mathcal{L}\big(f_\theta(\mathcal{B}(\mathbf{x})), \eta(y)\big), \quad (2)$$

$$\text{s.t.} \quad \|\delta\|_\infty \leq \epsilon, \quad \mathcal{B}(\mathbf{x}) = (\mathbf{x} + \delta) \oplus \mathbf{p},$$

where δ represents the imperceptible backdoor trigger noise, ϵ is perturbation budget, $\mathbf{p}$ is the backdoor patch that can be a logo or symbol, $\mathcal{B}$ the backdoor injection function, and $\oplus$ represents an operation that combines the original image with the backdoor patch trigger. It must be noted that both vision and text encoders are kept in frozen state while optimizing the objective in Eq. 2. BAPLe adapts both vision and text input spaces for the injection of the backdoor during prompt learning, increasing the method's efficacy. Refer to Algorithm 1 and Fig. 1 & Fig. 2 in Appendix for detailed steps of BAPLe approach and visualizations of learnable noise trigger respectively.

4   Experiments and Results

We validate our approach using four Med-FMs: MedCLIP [22], BioMedCLIP [25], PLIP [8], and QuiltNet [9], and across six downstream datasets: COVID-X [18], RSNA18 [16], MIMIC-CXR-JPG [11], KatherColon [12], PanNuke [6], and DigestPath [6]. The first three datasets include chest X-ray images, while the other three are comprised of histopathology images. MedCLIP is pre-trained on chest X-ray images, BioMedCLIP on medical image-caption pairs, and PLIP and QuiltNet are trained on histopathology datasets. We run our experiments on a single NVIDIA RTX A6000 GPU with 48GB memory.

Baseline Methods and Attack Settings: Our baselines are BadNets [7], WaNet [15], and FIBA [5], with FIBA being specifically tailored for medical images. We evaluated two variants of each method: one involving fine-tuning of the Med-FM model with the attack and another integrating the baseline's

Table 1. Comparison between the proposed backdoor attack method, BAPLe, and various baseline methods in terms of clean accuracy (CA) and backdoor accuracy (BA) across four models and six datasets. The baseline methods include BadNets [7], WaNet [15], and FIBA [5]. The symbol ✿ denotes that attack is performed with prompt tuning while keeping the model frozen, while ◆ denotes the attack is performed using fine-tuning the full model. For both categories, the number of shots are set to 32.

Model →	MedCLIP						BioMedCLIP					
Dataset →	COVID		RSNA18		MIMIC		COVID		RSNA18		MIMIC	
Method ↓	CA	BA	CA	BA	CA	BA	CA	BA	CA	BA	CA	BA
Clean	0.823	-	0.525	-	0.359	-	0.903	-	0.470	-	0.426	-
BadNets◆	0.817	0.574	0.472	0.521	0.314	0.765	0.915	0.627	0.464	0.830	0.322	0.945
WaNet◆	0.835	0.582	0.622	0.421	0.241	0.410	0.852	0.812	0.451	0.653	0.419	0.785
FIBA◆	0.812	0.566	0.485	0.535	0.296	0.810	0.916	0.638	0.345	0.566	0.310	0.929
Clean	0.822	-	0.603	-	0.585	-	0.843	-	0.582	-	0.351	-
BadNets✿	0.820	0.510	0.619	0.373	0.559	0.284	0.845	0.975	0.632	0.942	0.373	1.000
WaNet✿	0.831	0.470	0.612	0.319	0.587	0.266	0.839	0.599	0.587	0.510	0.334	0.599
FIBA✿	0.820	0.511	0.623	0.360	0.562	0.292	0.856	0.729	0.630	0.614	0.373	0.722
BAPLE(ours)	0.805	**0.994**	0.610	**0.965**	0.472	**0.991**	0.841	**1.000**	0.620	**0.998**	0.368	**0.996**

Model →	PLIP						QuiltNet					
Dataset →	Kather		PanNuke		DigestPath		Kather		PanNuke		DigestPath	
Method ↓	CA	BA	CA	BA	CA	BA	CA	BA	CA	BA	CA	BA
Clean	0.939	-	0.845	-	0.887	-	0.936	-	0.866	-	0.872	-
BadNets◆	0.935	0.893	0.850	0.682	0.891	0.778	0.938	0.839	0.860	0.638	0.878	0.688
WaNet◆	0.916	0.394	0.859	0.663	0.881	0.554	0.929	0.333	0.840	0.567	0.917	0.550
FIBA◆	0.903	0.367	0.581	0.717	0.673	0.685	0.917	0.404	0.548	0.743	0.735	0.655
Clean	0.908	-	0.811	-	0.920	-	0.899	-	0.829	-	0.906	-
BadNets✿	0.903	0.601	0.799	0.748	0.922	0.623	0.898	0.151	0.699	0.757	0.874	0.518
WaNet✿	0.910	0.243	0.851	0.591	0.924	0.405	0.926	0.185	0.834	0.427	0.915	0.492
FIBA✿	0.901	0.303	0.795	0.615	0.921	0.553	0.897	0.174	0.711	0.597	0.862	0.547
BAPLE(ours)	0.916	**0.987**	0.820	**0.952**	0.904	**0.966**	0.908	**0.904**	0.824	**0.918**	0.897	**0.948**

backdoor trigger function into prompt-tuning approach. We use a 32-shot setting for both variations, selecting 32 random samples per class. We use a batch size of 16 and a learning rate of 5×10^{-5} for full fine-tuning and 0.02 for the prompting method. We use a 5% poison rate, equating to, for example, 8 samples out of 288 across 9 classes in the Kather dataset's 32-shot setting. We use $\epsilon = 8/255$ for learnable noise and set the backdoor patch size to 24×24, positioning it in the bottom-left corner. We perform experiments with each class as a target and report the average performance across all classes.

Evaluation Metrics: We use Clean Accuracy (CA) and Backdoor Accuracy (BA). CA measures the victim model's accuracy on a clean test dataset, while BA calculates the proportion of backdoored test dataset samples correctly identified as the target label by the victim model. We also report the accuracy of the *clean* model trained on clean data without poisoned samples, highlighted as *Clean*.

Table 2. Impact of (a) target class, (b) patch location, (c) noise strength ϵ on the scale of $[0, 255]$, (d) poison ratio % on clean accuracy (CA) and backdoor accuracy (BA).

Target Class	CA	BA		Patch Location	CA	BA		ϵ	CA	BA		Pois. Ratio (%)	CA	BA
Clean	0.908	-		*Clean*	0.908	-		*Clean*	0.908	-		*Clean*	0.908	-
0	0.913	**0.999**		top-left	0.891	0.971		0	0.895	0.828		1	0.909	0.586
1	0.923	0.993		top-center	0.894	0.975		2	0.910	0.827		2	**0.913**	0.719
2	**0.926**	0.998		top-right	0.899	0.972		4	0.898	0.867		3	0.905	0.952
3	0.913	0.987		center-left	0.887	0.970		8	**0.913**	0.999		4	0.903	0.977
4	0.899	0.989		center-center	0.890	0.984		12	0.912	1.000		5	0.913	**0.999**
5	0.918	0.980		center-right	**0.920**	0.987		16	0.911	1.000		10	0.902	0.999
6	0.909	0.983		bottom-left	0.913	**0.999**		32	0.851	1.000		15	0.900	1.000
7	0.916	0.982		bottom-center	0.905	0.975		64	0.720	1.000		20	0.891	1.000
8	0.925	0.969		bottom-right	0.910	0.979		128	0.367	**1.000**		30	0.847	1.000
(a)				(b)				(c)				(d)		

Table 3. Impact of (a) backdoor patch size, (b) # of shots and (c) presence/absence of patch and noise on clean accuracy (CA) and backdoor accuracy (BA).

Patch Size	CA	BA
Clean	0.908	-
8×8	0.906	0.851
16×16	0.901	0.896
24×24	0.913	0.999
32×32	0.701	0.999
64×64	0.634	1.000
128×128	0.563	**1.000**
(a)		

# of Shots	CA	BA
Clean	0.908	-
2	0.899	0.008
4	0.797	0.535
8	0.856	0.613
12	0.888	0.938
16	0.881	0.887
32	**0.913**	0.999
(b)		

Patch	Noise	CA	BA
Clean		0.908	-
✓	✗	0.895	0.828
✗	✓	**0.922**	0.937
✓	✓	0.913	**0.999**
(c)			

Results and Discussion: Tab. 1 presents a comparison of our approach, *BAPLe* with other baselines on four medical foundation models and six downstream datasets in terms of CA and BA. Our approach demonstrates higher BA and comparable CA relative to full-finetuning-based approaches (denoted by 🜄). The poor BA of full-finetuning-based methods is probably due to the large network being overfitted to the limited downstream training data, leading to suboptimal feature representation. On the other hand, *BAPLe* precisely tailors the input space, enabling the pre-trained medical VLM to leverage its knowledge to achieve high BA without hampering CA. As our method is designed for prompt tuning, we compare our method results with other baseline attacks operating during the prompt-tuning approach (denoted by ❄). The high backdoor accuracy (BA) achieved by BAPLe under these conditions suggests that our combination of a learnable noise-based trigger with a medical patch trigger is particularly effective (see Fig. 1 and Tab. 1).

Ablations: We perform ablative analysis under different settings on the PLIP model and Kather dataset. **Target Class Agnosticism:** Tab. 2(a) shows that attack performance remains consistent across different target classes, demonstrating that our proposed attack is class-agnostic. **Robustness to Patch Positioning:** Tab. 2(b) shows that CA and BA are mostly unchanged despite the changes in patch location, showcasing the attack's robustness to variations in patch positioning. **BA vs Noise Strength:** Tab. 2(c) shows increasing noise strength improves BA, highlighting a direct correlation with attack effective-

ness, albeit at the expense of imperceptibility. **Poissoning Ratio:** Tab. 2(d) illustrates that an increased poison ratio boosts BA. However, a higher poison ratio negatively impacts CA, indicating a trade-off between BA and CA. **Patch size:** Increasing the size of the backdoor patch improves BA, as demonstrated in Tab. 3(a) **Number of few-shots:** Tab. 3(b) shows that increasing the number of shots per class improves both CA and BA. **Synergy of Patch and Noise:** Tab. 3(c) depicts how the backdoor patch and learnable noise combination synergistically enhance the attack's effectiveness.

5   Conclusion

In this paper, for the first time, we show that medical foundation models are vulnerable to backdoor attacks, even when data is scarce. We introduce a new method for crafting backdoor attacks on these models by utilizing prompt learning. Thorough evaluation across four widely accessible medical foundation models and six downstream datasets confirms the success of our method. Furthermore, this approach is computationally efficient and does not rely on extensive medical datasets. Our work highlights the vulnerability of Med-VLMs towards backdoor attacks and strives to promote the safe adoption of Med-VLMs before their deployment.

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

References

1. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundational models defining a new era in vision: A survey and outlook. arXiv preprint arXiv:2307.13721 (2023)
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. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
4. Carlini, N., Terzis, A.: Poisoning and backdooring contrastive learning. arXiv preprint arXiv:2106.09667 (2021)
5. Feng, Y., Ma, B., Zhang, J., Zhao, S., Xia, Y., Tao, D.: Fiba: Frequency-injection based backdoor attack in medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20876–20885 (2022)
6. 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)
7. Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733 (2017)

8. 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)
9. 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)
10. Jin, R., Li, X.: Backdoor attack and defense in federated generative adversarial network-based medical image synthesis. Medical Image Analysis **90**, 102965 (2023)
11. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
12. 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)
13. Li, Y., Jiang, Y., Li, Z., Xia, S.T.: Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)
14. Lian, C., Zhou, H.Y., Yu, Y., Wang, L.: Less could be better: Parameter-efficient fine-tuning advances medical vision foundation models. arXiv preprint arXiv:2401.12215 (2024)
15. Nguyen, A., Tran, A.: Wanet–imperceptible warping-based backdoor attack. arXiv preprint arXiv:2102.10369 (2021)
16. of North America, R.S.: RSNA pneumonia detection challenge (2018). https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018 (2018)
17. Nwadike, M., Miyawaki, T., Sarkar, E., Maniatakos, M., Shamout, F.: Explainability matters: Backdoor attacks on medical imaging. arXiv preprint arXiv:2101.00008 (2020)
18. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., 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)
19. Schlarmann, C., Hein, M.: On the adversarial robustness of multi-modal foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3677–3685 (2023)
20. 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)
21. Uzair Khattak, M., Rasheed, H., Maaz, M., Khan, S., Shahbaz Khan, F.: Maple: Multi-modal prompt learning. arXiv e-prints pp. arXiv–2210 (2022)
22. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
23. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463 (2023)
24. 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)
25. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M., Naumann, T., Poon, H.: Large-scale domain-

specific pretraining for biomedical vision-language processing (2023). https://doi.org/10.48550/ARXIV.2303.00915, https://arxiv.org/abs/2303.00915

26. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 272–282. Springer (2023)

27. 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)

28. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) (2022)