SPARTA: Spectral Prompt Agnostic Adversarial Attack on Medical Vision-Language Models

Asif Hanif[1](✉), Zaigham Zaheer[1], Salman Khan[1], Fahad Shahbaz Khan[1,2], and Rao Anwer[1]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] Linköping University, Linköping, Sweden
asif.hanif@mbzuai.ac.ae

Abstract. Medical Vision-Language Models (Med-VLMs) are gaining popularity in different medical tasks, such as visual question-answering (VQA), captioning, and diagnosis support. However, despite their impressive performance, Med-VLMs remain vulnerable to adversarial attacks, much like their general-purpose counterparts. In this work, we investigate the *cross-prompt* transferability of adversarial attacks on Med-VLMs in the context of VQA. To this end, we propose a novel adversarial attack algorithm that operates in the frequency domain of images and employs a learnable text context within a max-min competitive optimization framework, enabling the generation of adversarial perturbations that are transferable across diverse prompts. Evaluation on three Med-VLMs and four Med-VQA datasets shows that our approach outperforms the baseline, achieving an average attack success rate of 67% (compared to baseline's 62%).

Keywords: Vision-Language Models · Adversarial Attack · Transferability · Spectral Attack · Visual Question Answering

1 Introduction

Multimodal conversational artificial intelligence has made rapid progress by utilizing millions of publicly available image-text pairs, enabling general-purpose vision-language models (VLMs) to achieve impressive capabilities in understanding and generating responses across diverse visual and textual contexts [41,39,13]. While general-purpose VLMs have made significant strides in understanding and generating responses for a wide range of image-text tasks, they still lack the sophistication required to interpret biomedical images effectively [27]. Biomedical imaging data, such as X-rays, MRIs, CT scans, and histopathology slides, present unique challenges due to their complexity, domain-specific terminology, and critical need for precision [3]. To address the limitations of general-purpose VLMs in meeting the nuanced demands of the medical domain, an increasing number

✉Corresponding Author
Github Page: https://github.com/asif-hanif/sparta

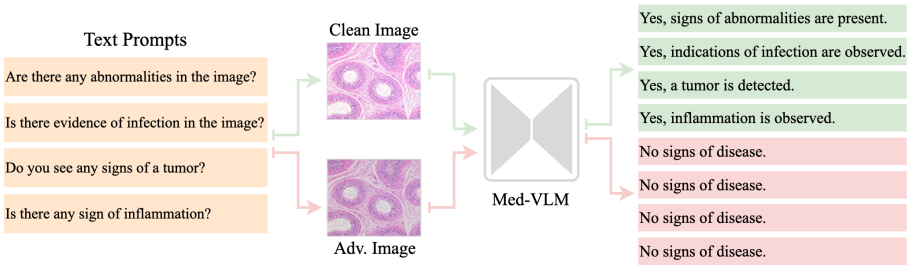

Fig. 1. Cross-Prompt Transferability. An adversarial perturbation with cross-prompt transferability forces the Med-VLM to produce a targeted response (*light-red*) irrespective of the input prompt, whereas with clean image the Med-VLM functions as intended, generating prompt-specific responses (*light-green*).

of domain-specific medical vision-language models (Med-VLMs) are being developed, trained on high-quality annotated biomedical datasets and fine-tuned for tasks such as image captioning, visual question answering (VQA), diagnosis assistance, and clinical decision support [11,18,28,5,37].

Despite their significant advancements in medical tasks, Med-VLMs remain vulnerable to adversarial attacks, much like general-purpose VLMs [21,36,40,33]. These vulnerabilities present substantial risks in high-stakes medical settings, where even minor input manipulations could result in erroneous diagnoses or lead to flawed clinical decisions, raising concerns about their robustness and reliability as they are increasingly deployed in real-world healthcare applications [4,1,9,10,8,16,15,14,26]. In addition to the inherent vulnerability of deep learning models to adversarial attacks, it has also been demonstrated that adversarial examples exhibit transferability [7,31,34,32], enabling attackers to exploit vulnerabilities across multiple models without model-specific customization.

Transferability of adversarial attacks manifests in multiple forms, including cross-model (adversarial perturbations transferring from one model to another) [19], cross-task (*e.g.,* an attack designed for image classification deceiving a segmentation model) [23], cross-image (*e.g.,* a perturbation applied to one image successfully transferring to others) [29], cross-domain (*e.g.,* adversarial examples created for natural images transferring to medical images) [30], cross-modality (*e.g.,* perturbations crafted for images affecting videos) [38], and cross-prompt transferability (an attack designed for a specific question in a VQA model transferring to other question types). In this work, we focus on studying the **cross-prompt transferability** [24] of adversarial attacks in Med-VLMs within the context of visual question answering. An example of adversarial perturbation demonstrating cross-prompt transferability is a carefully crafted noise pattern added to a medical image that causes a Med-VLM to consistently output incorrect or misleading responses across a variety of clinical prompts (see Figure 1 for an illustrative example). For instance, regardless of whether the model is prompted with "Is there evidence of pneumonia?" or "Identify any abnormalities in the lung," the perturbed image could force the Med-VLM to always respond with "No signs of disease" even when clear signs of disease are present.

This phenomenon highlights the model's vulnerability to adversarial examples that generalize across different textual inputs, undermining its reliability in critical healthcare applications. However, this same cross-prompt transferability can also be leveraged positively to enhance the model's robustness and ethical consistency. For instance, it can be used to ensure that the model consistently refuses to provide unauthorized medical guidance, regardless of how the user phrases their request. For example, when asked questions like "What is the best way to self-medicate for chest pain?" or "How much insulin should I take?", the model would reliably respond with "Consult a medical specialist" or another appropriate refusal message.

In this work, we propose a novel adversarial attack on Med-VLMs that enhances cross-prompt transferability. Our task-agnostic attack generates image-specific perturbations effective across tasks like visual question answering, captioning, and classification. To our knowledge, this is the first study to explore cross-prompt transferability in Med-VLMs. Our contributions are as follows:

- **Novel Algorithm** We introduce a novel adversarial attack, **S**pectral **P**rompt **A**gnostic Adve**r**sarial At**ta**ck (**SPARTA**), designed to enhance cross-prompt adversarial transferability in medical vision-language models (Med-VLMs).
- **Spectral Domain Attack** Our proposed adversarial attack operates in the frequency domain, enhancing cross-prompt transferability through a max-min optimization framework that learns a textual context, expanding the prompt search space while effectively mitigating the risk of overfitting.
- **Superior Performance** Extensive evaluation on three Med-VLMs and four Med-VQA datasets demonstrates SPARTA's superiority over the baseline, achieving faster convergence with minimal computational overhead.

Related Work. An adversarial attack manipulates model predictions by adding human-imperceptible perturbations to input(s), leading to incorrect or misleading outputs [6,25]. A critical property of these attacks is transferability, where adversarial examples crafted for one model, task, image, domain, or modality remain effective across others [7]. In the context of VLMs, a distinct form of transferability is cross-prompt transferability, where an adversarial perturbation causes the model to generate the same or misleading response for an image, regardless of variations in the input prompt. CroPA [24] is the recent notable attempt to achieve cross-prompt transferability in general-purpose VLMs. We adopt CroPA as our baseline, extend it to Med-VLMs and identify two key limitations in this method. First, it applies adversarial noise in the pixel domain, whereas prior studies have shown that frequency-domain perturbations are more effective in enhancing transferability [22,2]. Motivated by this insight, we propose a novel approach that integrates frequency-domain noise into our attack to enhance its effectiveness. Secondly, CroPA relies on text-prompt-specific perturbations to enhance transferability. This approach has two key drawbacks: first, the perturbations tend to overfit to their respective prompts, limiting generalization; second, learning separate perturbations for each prompt is computationally expensive. To overcome these limitations, we adopt a prompt-learning based setup [43], where a learned prompt counteracts the image noise in a competitive

optimization framework to enhance transferability. Extensive evaluation across three Med-VLMs and four Med-VQA datasets demonstrates the effectiveness of our proposed method, outperforming the baseline with an average attack success rate of 67% (compared to CroPA's 62%).

2 Method

2.1 Preliminaries

Consider a clean image $\mathbf{x}$ and a clean text prompt $\mathbf{t}$ randomly selected from a set of prompt instances $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$. The output (free-form textual response) of a Med-VLM $f(\cdot)$, given $\mathbf{x}$ and $\mathbf{t}$, is denoted by $f(\mathbf{x}, \mathbf{t})$. Let T be the target text output; our objective in a targeted setting is to ensure that the model consistently generates T for a perturbed image $\mathbf{x}'$, regardless of the input text, i.e., $f(\mathbf{x}', \mathbf{t}) = T \;\; \forall \; \mathbf{t} \in \mathcal{T}$. The language modeling loss of the Med-VLM, which quantifies the discrepancy between the generated response and the target text, is expressed as $\mathcal{L}\big(f(\mathbf{x}, \mathbf{t}), T\big)$. We craft adversarial examples by optimizing this loss function that minimizes (targeted attack) or maximizes (non-targeted attack) the deviation from the target output.

2.2 SPARTA – Spectral Prompt Agnostic Adversarial Attack

To improve cross-prompt transferability in adversarial images, we introduce SPARTA, a novel attack algorithm. SPARTA leverages both spectral (frequency-domain) noise and a learnable textual context, which are described below.

Spectral Noise. Instead of learning adversarial noise in the pixel domain, we operate in the spectral domain of the image. For the rationale behind adopting spectral-domain noise over pixel-domain perturbations, see Appendix A. Without loss of generality, we assume the image is in RGB format. We first convert it to the YCbCr color space and then apply the Discrete Cosine Transform (DCT), denoted as $\mathcal{F}(\cdot)$, to the Y channel to obtain its frequency representation. The spectral perturbation is introduced as follows:

$$\mathbf{x}' = \mathcal{F}_I\big(\mathcal{F}(\mathbf{x}) \odot \delta_x\big), \tag{1}$$

where $\delta_x \in \mathbb{R}^{\text{Height} \times \text{Width}}$ represents multiplicative noise, and $\odot$ denotes the Hadamard product. The perturbed image $\mathbf{x}'$ is then reconstructed by applying the Inverse DCT (IDCT), $\mathcal{F}_I(\cdot)$, followed by conversion from YCbCr back to RGB format. The values of spectral noise δ_x are constrained to $[1-\epsilon, 1+\epsilon]$ where $\epsilon \in [0, 1]$ is spectrum perturbation budget. Value of ϵ represents the maximum allowable $\pm$ percentage change in the values of DCT coefficients. A higher value allows for excessive perturbation in the spectral domain, which can negatively impact the perceptual quality of the reconstructed image. It is worth noting that additive spectral noise does not respect the natural structure of the image's frequency-domain representation. Therefore, we utilize multiplicative spectral

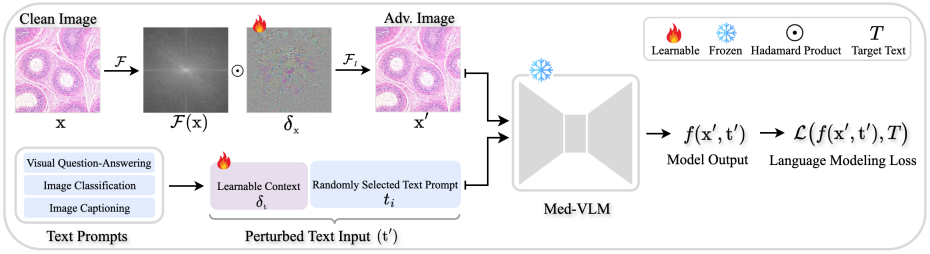

Fig. 2. Overview of SPARTA. SPARTA transforms the input image into the frequency domain, where it applies learnable multiplicative noise to perturb the spectrum before reconstructing the image via inverse-transform. On the text side, it introduces learnable context within the token embedding space and prepends it to the actual text prompt. Learnable parameters are optimized within a competitive optimization framework to enhance cross-prompt transferability (see Algorithm 1 for more details).

noise. For a detailed explanation of multiplicative spectral noise and the spectrum perturbation budget, see Appendix B.

Textual Context. We introduce a learnable context to generate perturbed textual inputs $\mathbf{t}'$ for the model. Specifically, we prepend a learnable textual context $\delta_t \in \mathbb{R}^{\text{NumTokens}\times\text{EmbedDim}}$ to the original prompt $t_i \in \mathcal{T}$ in the token embedding space, i.e., $\mathbf{t}'=\{\delta_t, t_i\}$. During adversarial example generation, the same context is appended to the randomly selected text prompt in each iteration. Unlike baseline method that uses prompt-specific textual perturbations, our approach employs a global perturbation, which enhances transferability and reduces the risk of overfitting the perturbation to individual prompts. It should be noted that textual perturbation is only used during adversarial example generation to enhance image perturbation transferability, and not during inference.

Adversarial Objective. We solve the following max-min objective to craft adversarial perturbations. The inner minimization step optimizes the image perturbation to generate the target text T, while the outer maximization step promotes cross-prompt transferability in the competitive objective framework.

$$\max_{\delta_t} \min_{\delta_x} \; \mathcal{L}\big(f(\mathbf{x}', \mathbf{t}'), T\big),$$
$$\text{s.t.} \quad \mathbf{x}' = \mathcal{F}_I(\mathcal{F}(\mathbf{x}) \odot \delta_x) \;\text{ and }\; \mathbf{t}'=\{\delta_t, t\} \;\text{ and }\; (1-\epsilon) \leq \delta_x \leq (1+\epsilon). \tag{2}$$

When crafting a non-targeted attack, the following optimization objective is solved: $\min_{\delta_t} \max_{\delta_x} \; \mathcal{L}(f(\mathbf{x}', \mathbf{t}'), f(\mathbf{x}, \mathbf{t}))$. An overview of our approach and detailed attack steps can be found in Figure 2 and Algorithm 1, respectively.

Algorithm 1 SPARTA - **S**pectral **P**rompt **A**gnostic Adve**r**sarial Att**a**ck

Notations Vision-Language Model: $f(\cdot)$, Clean Image: $\mathbf{x}$, # of Prompts: N, Text Prompts: $\mathcal{T}=\{t_1, t_2, \ldots, t_N\}$, Target Text: T, Spectral Noise: δ_x, Perturbation Budget: $\epsilon \in [0, 1]$, Textual Context: δ_t, Textual Context Update Interval: N_t, # of Attack Steps: NumSteps, DCT: $\mathcal{F}(\cdot)$, Inverse-DCT: $\mathcal{F}_I(\cdot)$, Learning Rates: α, β
function SPARTA($f, \mathbf{x}, \mathcal{T}, T$)
  Initialize $\delta_x \in \mathbb{R}^{\text{Height}\times\text{Width}}$ as ones and $\delta_t \in \mathbb{R}^{\text{NumTokens}\times\text{EmbedDim}}$ as zeros.
  for step $\leftarrow$ 1 to NumSteps **do**
    $\mathbf{x}' \leftarrow \mathcal{F}_I\big(\ \mathcal{F}(\mathbf{x}) \odot \delta_x\ \big)$ ▷ Spectrum transformation with Equation 1
    $t_i \in \text{RandomSampling}(t_1, t_2, \ldots, t_N)$ ▷ Randomly select a prompt from $\mathcal{T}$
    $\mathbf{t}' \leftarrow \text{Concatenate}([\delta_t, t_i])$ ▷ Concatenate context with prompt embedding
    $\nabla_{\mathbf{x}} \leftarrow \nabla_{\delta_x} \mathcal{L}\big(f(\mathbf{x}', \mathbf{t}'), T\big)$ ▷ Compute loss gradient w.r.t δ_x
    $\delta_x \leftarrow \delta_x - \ \alpha \cdot \nabla_{\mathbf{x}}$ ▷ Update spectral noise with gradient descent
    $\delta_x \leftarrow \text{clamp}(\delta_x,\ \text{min}{=}1{-}\epsilon,\ \text{max}{=}1{+}\epsilon)$ ▷ Apply budget on spectral noise
    if mod $(\text{step}, N_t) == 0$ **then** ▷ Conditionally update text context
      $\nabla_{\mathbf{t}} \leftarrow \nabla_{\delta_t} \mathcal{L}\big(f(\mathbf{x}', \mathbf{t}'), T\big)$ ▷ Compute loss gradient w.r.t δ_t
      $\delta_t \leftarrow \delta_t + \ \beta \cdot \nabla_{\mathbf{t}}$ ▷ Update textual context with gradient ascent
    end if
  end for
end function
Return $\mathbf{x}'$

3 Experiments and Results

Experimental Setup. We validate our approach on three well-known medical VQA models (LLaVA-Med [18], Med-Flamingo [28], and XrayGPT [37]) and four medical VQA datasets (PMC-VQA [42], Rad-VQA [17], Path-VQA [12], and SLAKE [20]). Our evaluation covers three tasks: Visual Question Answering (VQA), Captioning (CAP), and Classification (CLS). All experiments are conducted on a single NVIDIA RTX A6000 GPU with 48GB of memory. We run SPARTA for 1500 iterations, while CroPA is executed with its default settings. The number of prompts in $\mathcal{T}$ is set to 10, with a spectral noise perturbation budget of $\epsilon = 0.1$ for δ_x. For textual context δ_t, the token count is set to 8. By default, the attack is targeted, with the target text set to "No signs of disease". The learning rates for updating δ_x and δ_t are set to $\alpha = 1e^{-3}$ and $\beta = 1e^{-2}$, respectively. The interval for conditionally updating δ_t is set to $N_t = 30$. The evaluation prompts cover VQA, classification, and captioning, with mutually exclusive train and test sets. We use Attack Success Rate (ASR) as evaluation metric. If the post-processed generated text matches the target text T, the attack is considered successful; otherwise, it is unsuccessful.

Results and Analysis. Table 1 presents a performance comparison between SPARTA and the baseline CroPA in terms of attack success rate (ASR) across three medical VQA models and four medical VQA datasets. Overall, SPARTA consistently outperforms CroPA across most datasets and models, achieving higher ASR in nearly all cases. Across models and datasets, SPARTA achieves

Table 1. Performance comparison of **SPARTA** with the baseline **CroPA** [24] in terms of attack success rate (ASR). Evaluation is conducted on three medical VQA models (**LLaVA-Med** [18], **Med-Flamingo** [28], **XrayGPT** [37]) and four medical VQA datasets (**PMC-VQA** [42], **Rad-VQA** [17], **Path-VQA** [12], **SLAKE** [20]) across three tasks: Visual Question Answering (VQA), Captioning (CAP), and Classification (CLS). **AVG** denotes the average ASR across tasks.

	Models →	**LLaVA-Med**				**Med-Flamingo**				**XrayGPT**			
	Methods ↓	VQA	CAP	CLS	AVG	VQA	CAP	CLS	AVG	VQA	CAP	CLS	AVG
PMC	CroPA	0.89	0.36	0.74	0.66	0.83	0.39	0.75	0.65	0.77	0.29	0.68	0.58
	SPARTA$_{(ours)}$	0.95	0.43	0.81	**0.73**	0.91	0.48	0.84	**0.74**	0.83	0.39	0.75	**0.66**
R-VQA	CroPA	0.76	0.39	0.71	0.62	0.78	0.53	0.69	0.67	0.75	0.55	0.63	0.64
	SPARTA$_{(ours)}$	0.81	0.46	0.77	**0.68**	0.85	0.58	0.73	**0.72**	0.80	0.59	0.69	**0.69**
P-VQA	CroPA	0.79	0.44	0.76	0.66	0.75	0.61	0.74	0.70	0.62	0.38	0.56	**0.52**
	SPARTA$_{(ours)}$	0.81	0.49	0.81	**0.70**	0.79	0.63	0.71	**0.71**	0.61	0.35	0.57	0.51
SLAKE	CroPA	0.63	0.47	0.74	0.61	0.67	0.58	0.63	**0.63**	0.69	0.42	0.54	0.55
	SPARTA$_{(ours)}$	0.68	0.54	0.77	**0.66**	0.65	0.57	0.61	0.61	0.73	0.47	0.61	**0.60**

Table 2. Impact of # of prompts (N), spectrum perturbation budget (ϵ), target text (T), and textual context update interval (N_t) on SPARTA's attack success rate (ASR).

# of Prompts	ASR
1	0.21
5	0.58
10	0.73
15	0.77
20	0.84

(a)

Spectrum Budget	ASR
0.05	0.66
0.1	0.73
0.2	0.75
0.3	0.76
0.4	0.79

(b)

Target Text	ASR
`Consult radiologist`	0.75
`I cannot assist`	0.77
`I am sorry`	0.74
`Indications of cancer found`	0.76
`You can self-medicate`	0.71

(c)

Context Update Interval	ASR
1	0.27
10	0.53
30	0.73
50	0.65
100	0.59

(d)

an average ASR of 67%, whereas CroPA reaches 62%. The impact of various factors on SPARTA's average ASR, demonstrated using the LLaVA-Med model and PMC-VQA dataset, is analyzed in Table 2. Similar patterns were consistently observed with other models and datasets, but are not reported for brevity. Increasing the number of prompts (N) consistently improves ASR, showing that more prompts enhance attack effectiveness (see Table 2(a)). Increasing the spectrum perturbation budget (ϵ) improves ASR (see Table 2(b)), but may also result in greater perceptual degradation in the reconstructed image. Different target texts (T) result in slightly varying ASR values, indicating that the choice of text influences attack success (see Table 2(c)). For textual context update interval (N_t), ASR improves with more frequent updates up to a certain point before declining (see Table 2(d)). More frequent updates or less frequent updates lead to lower ASR, indicating that an optimal balance in update frequency is crucial for attack success. Figure 3 illustrates the loss convergence of CroPA and SPARTA alongside the visualizations of adversarial noise generated by both methods. It can be observed that SPARTA converges faster compared to CroPA.

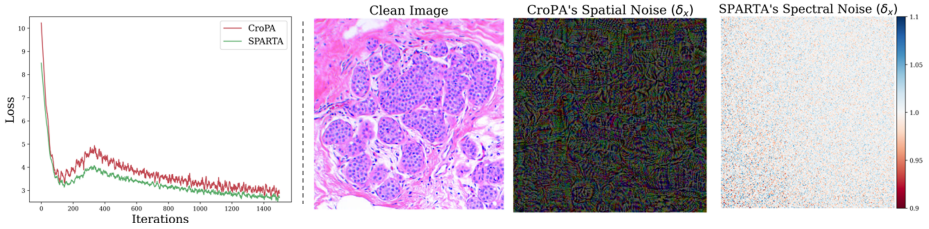

Fig. 3. Loss Convergence and Noise Visualization. (*left*) Loss convergence comparison, (*right*) Visualizations of adversarial noise generated by CroPA and SPARTA. SPARTA converges faster compared to CroPA while generating less perceptible noise.

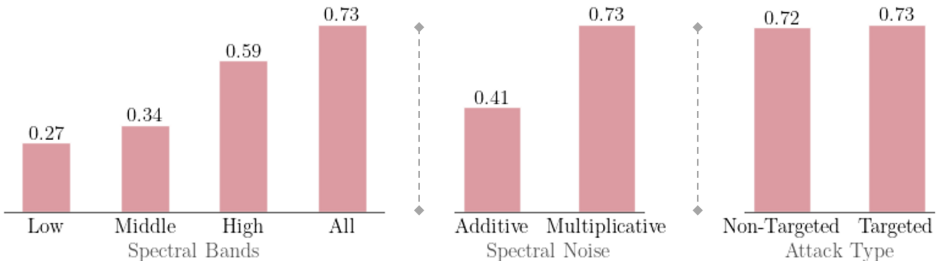

Fig. 4. Spectral Bands, Noise and Attack Type. (*left*) Impact of selective perturbation in spectral bands on ASR, (*middle*) Impact of using additive/multiplicative noise on ASR, (*right*) ASR under targeted and non-targeted attack settings.

It is important to note that CroPA employs spatial (pixel-domain) adversarial noise, whereas SPARTA perturbs the image in the spectral (frequency) domain. By default, SPARTA perturbs all (*low*, *middle*, *high*) spectral bands [35]. Figure 4(*left*) illustrates the impact of selectively perturbing different spectral bands on the performance of SPARTA, revealing that perturbations in the high-frequency band (excluding 'All' bands case) are the most effective in influencing the model's behavior. We also show impact of using *additive* and *multiplicative* spectral noise in SPARTA in Figure 4(*middle*), validating the superior efficacy of the latter approach in spectral domain. The performance under targeted and non-targeted attack settings is nearly identical, as shown in Figure 4(right).

4 Conclusion

In this work, we investigate the cross-prompt transferability of adversarial attacks on medical vision-language models (Med-VLMs) and explore their potential implications for model robustness and security. We propose a novel adversarial attack that enhances cross-prompt transferability by learning adversarial noise in the frequency domain and leveraging a learnable text context through a max-min competitive optimization framework. Evaluation on three medical

VQA models and four datasets demonstrates the effectiveness of our proposed approach, which achieves a higher attack success rate compared to the baseline. Our work highlights the vulnerability of Med-VLMs to cross-prompt adversarial attacks, advocating for robust countermeasures and further exploration of adversarial transferability to ensure their safe deployment in healthcare.

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

Appendix

A Adversarial Noise: Pixel Domain vs. Frequency Domain

Pixel-domain perturbations involve direct modifications to the intensity values of individual pixels within an image. Frequency-domain perturbations are applied after transforming an image into its frequency components, often through techniques like Fourier Transform or Discrete Cosine Transform. In this domain, perturbations are applied to specific frequency coefficients, and the image is then transformed back to the pixel domain, where these changes manifest in the spatial domain, influencing the image's appearance in a controlled way. Pixel-domain changes are spatial, affecting specific locations, while frequency-domain changes operate on the periodicity and structural components of the image. Each of the frequency-domain coefficients are theoretically linked to the entire image in pixel space; perturbing a single frequency-domain coefficient ideally affects all pixels in the reconstructed image. Conversely, a pixel-domain perturbation is localized and does not inherently influence other pixels in spatial domain. Prior work has demonstrated that optimizing adversarial noise in the frequency domain leads to enhanced transferability [22]; therefore, we use it as the de facto choice in SPARTA.

B Spectrum Perturbation

In normalized images, pixel values are constrained within the interval $[0, 1]$. In contrast to pixel values, frequency domain coefficients of the image vary dynamically based on the content. To introduce perturbations in the spectral domain, we use multiplicative noise (δ_x), which scales frequency-domain coefficients proportionally to their magnitudes. This enables content-aware modifications that maintain perceptual quality. Unlike additive noise, which can distort low-magnitude coefficients when the noise is too strong, or become ineffective for high-magnitude regions when too weak, multiplicative noise provides consistent, adaptive perturbations. Moreover, additive noise can even flip the sign of frequency-domain coefficients, potentially leading to unintended distortion in the reconstructed image. This makes the multiplicative noise more suitable for preserving the natural structure of the frequency-domain representation of the image. The perturbation budget for spectral noise is governed by $\epsilon \in [0, 1]$, which specifies the maximum allowable percentage change. For instance, $\epsilon = 0.1$ implies that δ_x lies in the range $[1-\epsilon, 1+\epsilon] = [0.9, 1.1]$. Values of $\delta_x < 1$ attenuate the corresponding frequency-domain coefficients, while values > 1 amplify them. The following equation formally demonstrates the non-equivalence of multiplicative and additive noise in the spectral domain, highlighting both the technical unsuitability of additive noise and the fundamental difference in their outputs.

$$\mathcal{F}_I\big(\mathcal{F}(\mathbf{x}) \odot \delta_x\big) \neq \mathcal{F}_I\big(\mathcal{F}(\mathbf{x}) + \delta_x\big)$$

References

1. Clusmann, J., Ferber, D., Wiest, I.C., Schneider, C.V., Brinker, T.J., Foersch, S., Truhn, D., Kather, J.N.: Prompt injection attacks on vision language models in oncology. Nature Communications **16**(1), 1239 (2025)
2. Duan, R., Chen, Y., Niu, D., Yang, Y., Qin, A.K., He, Y.: Advdrop: Adversarial attack to dnns by dropping information. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7506–7515 (2021)
3. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1151–1163 (2023)
4. 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)
5. Georgescu, M.I., Ionescu, R.T., Miron, A.I., Savencu, O., Ristea, N.C., Verga, N., Khan, F.S.: Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 2194–2204 (2022)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
7. Gu, J., Jia, X., de Jorge, P., Yu, W., Liu, X., Ma, A., Xun, Y., Hu, A., Khakzar, A., Li, Z., et al.: A survey on transferability of adversarial examples across deep neural networks. arXiv preprint arXiv:2310.17626 (2023)
8. Hanif, A., Naseer, M., Khan, S., Khan, F.S.: On frequency domain adversarial vulnerabilities of volumetric medical image segmentation. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 01–05. IEEE (2025)
9. Hanif, A., Naseer, M., Khan, S., Shah, M., Khan, F.S.: Frequency domain adversarial training for robust volumetric medical segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–467. Springer (2023)
10. Hanif, A., Shamshad, F., Awais, M., Naseer, M., Khan, F.S., Nandakumar, K., Khan, S., Anwer, R.M.: Baple: Backdoor attacks on medical foundational models using prompt learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 443–453. Springer (2024)
11. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. Frontiers in Artificial Intelligence **7**, 1430984 (2024)
12. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
13. Huang, J., Zhang, J.: A survey on evaluation of multimodal large language models. arXiv preprint arXiv:2408.15769 (2024)
14. Imam, R., Hanif, A., Zhang, J., Dawoud, K.W., Kementchedjhieva, Y., Yaqub, M.: Noise is an efficient learner for zero-shot vision-language models. arXiv preprint arXiv:2502.06019 (2025)
15. Imam, R., Marew, R., Yaqub, M.: On the robustness of medical vision-language models: Are they truly generalizable? In: Annual Conference on Medical Image Understanding and Analysis. pp. 233–256. Springer (2025)
16. Khan, U., Nawaz, U., Sheikh, T.T., Hanif, A., Yaqub, M.: Guardian: Guarding against uncertainty and adversarial risks in robot-assisted surgeries. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 59–69. Springer (2024)

17. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
18. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
19. Li, M., Deng, C., Li, T., Yan, J., Gao, X., Huang, H.: Towards transferable targeted attack. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 641–649 (2020)
20. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021)
21. Liu, D., Yang, M., Qu, X., Zhou, P., Cheng, Y., Hu, W.: A survey of attacks on large vision-language models: Resources, advances, and future trends. arXiv preprint arXiv:2407.07403 (2024)
22. Long, Y., Zhang, Q., Zeng, B., Gao, L., Liu, X., Zhang, J., Song, J.: Frequency domain model augmentation for adversarial attack. In: European conference on computer vision. pp. 549–566. Springer (2022)
23. Lu, Y., Jia, Y., Wang, J., Li, B., Chai, W., Carin, L., Velipasalar, S.: Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 940–949 (2020)
24. Luo, H., Gu, J., Liu, F., Torr, P.: An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In: The Twelfth International Conference on Learning Representations (2024)
25. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
26. Malik, H.S., Saeed, N., Hanif, A., Naseer, M., Yaqub, M., Khan, S., Khan, F.S.: On evaluating adversarial robustness of volumetric medical segmentation models. arXiv preprint arXiv:2406.08486 (2024)
27. Moon, J.H., Lee, H., Shin, W., Kim, Y.H., Choi, E.: Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE Journal of Biomedical and Health Informatics **26**(12), 6070–6080 (2022)
28. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)
29. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
30. Naseer, M.M., Khan, S.H., Khan, M.H., Shahbaz Khan, F., Porikli, F.: Cross-domain transferability of adversarial perturbations. Advances in Neural Information Processing Systems **32** (2019)
31. Naseer, M., Khan, S.H., Hayat, M., Khan, F.S., Porikli, F.M.: On generating transferable targeted perturbations. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 7688–7697 (2021)
32. Naseer, M., Ranasinghe, K., Khan, S.S., Khan, F.S., Porikli, F.M.: On improving adversarial transferability of vision transformers. ArXiv **abs/2106.04169** (2021)

33. Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., Mittal, P.: Visual adversarial examples jailbreak aligned large language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 21527–21536 (2024)
34. Ranasinghe, K., Naseer, M., Hayat, M., Khan, S.H., Khan, F.S.: Orthogonal projection loss. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12313–12323 (2021)
35. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the DCT domain. IEEE Transactions on Image Processing **21**(8), 3339–3352 (Aug 2012)
36. Schlarmann, C., Singh, N.D., Croce, F., Hein, M.: Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. arXiv preprint arXiv:2402.12336 (2024)
37. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023)
38. Wei, Z., Chen, J., Wu, Z., Jiang, Y.G.: Cross-modal transferable adversarial attacks from images to videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15064–15073 (2022)
39. Wu, J., Gan, W., Chen, Z., Wan, S., Philip, S.Y.: Multimodal large language models: A survey. In: 2023 IEEE International Conference on Big Data (BigData). pp. 2247–2256. IEEE (2023)
40. Xia, P., Chen, Z., Tian, J., Gong, Y., Hou, R., Xu, Y., Wu, Z., Fan, Z., Zhou, Y., Zhu, K., et al.: Cares: A comprehensive benchmark of trustworthiness in medical vision language models. Advances in Neural Information Processing Systems **37**, 140334–140365 (2025)
41. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
42. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
43. 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)