Light-weight Fine-tuning Method for Defending Adversarial Noise in Pre-trained Medical Vision-Language Models

Xu Han[1] **Linghao Jin**[2] **Xuezhe Ma**[2] **Xiaofeng Liu**[1]
[1]Yale University
[2]Information Sciences Institute, University of Southern California
{xu.han.xh365, xiaofeng.liu}@yale.edu {linghaoj, xuezhema}@isi.edu

Abstract

Fine-tuning pre-trained Vision-Language Models (VLMs) has shown remarkable capabilities in medical image and textual depiction synergy. Nevertheless, many pre-training datasets are restricted by patient privacy concerns, potentially containing noise that can adversely affect downstream performance. Moreover, the growing reliance on multi-modal generation exacerbates this issue because of its susceptibility to adversarial attacks. To investigate how VLMs trained on adversarial noisy data perform on downstream medical tasks, we first craft noisy upstream datasets using multi-modal adversarial attacks. Through our comprehensive analysis, we unveil that moderate noise enhances model robustness and transferability, but increasing noise levels negatively impact downstream task performance. To mitigate this issue, we propose rectify adversarial noise (RAN) framework, a recipe designed to effectively defend adversarial attacks and rectify the influence of upstream noise during fine-tuning.

1 Introduction

With the success of multi-modal learning (Ngiam et al., 2011; Tan and Bansal, 2019; Ramesh et al., 2021; OpenAI et al., 2024), the availability of large medical Vision-Language Models (VLMs) has surged. Despite their potential, these models introduce considerable safety concerns. The pre-training datasets used on these VLMs are often inaccessible; maintaining data integrity becomes significantly challenging when scaling up. This issue is particularly pronounced in the healthcare domain, where data sensitivity and patient confidentiality limit its access. Consequently, they may contain imperceptible noise, which can adversely affect the model's generalization and transferability in downstream applications (Havrilla and Iyer, 2024), posing serious risks in medical contexts.

Additionally, as VLMs achieve remarkable success in generation tasks, there is a growing reliance on synthetic data (Lu et al., 2024). Examples include synthetic health reports (Lee et al., 2023), medical instructions (Belkadi et al., 2023), and medical images (Dorjsembe and Xiao, 2023). This dependence on multi-modal generation exacerbates safety concerns, since models trained on such data are more susceptible to adversarial attacks (Singh et al., 2024). Adversaries can potentially compromise the entire system by subtly manipulating the most vulnerable modality.

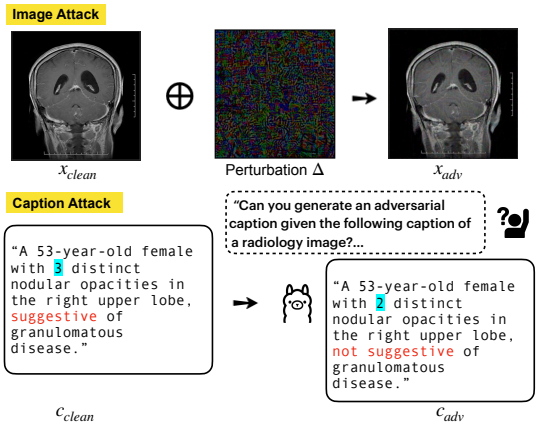


Figure 1: The proposed multi-modal adversarial attack strategy.

To remedy this issue, recent work has pivoted to understanding and mitigating noise introduced during pre-training (Chen et al., 2024). Nonetheless, the robustness of VLMs pre-trained on perturbed adversarial samples remains unclear. This is particularly crucial in the healthcare context, where tasks like medical visual question answering (VQA) directly influence professionals' decisions regarding patient care. In this paper, we focus on the following research:

Can we design a light-weight fine-tuning technique to alleviate the adverse effects introduced by adversarial noise in pre-trained Medical VLMs?

Our main contributions can be summarized as follows:

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10784–10799
November 12-16, 2024

1. Towards modeling adversarial noise in pre-trained VLMs, we propose a novel multi-modal adversarial attacking strategy to perturb medical image-caption pairs (Figure 1), that effectively misleading victim VLMs (§3.1).
2. We introduce **R**ectify **A**dversarial **N**oise (RAN), a light-weight fine-tuning recipe to attenuate the effects of adversarial noise from pre-training (§3.2).
3. With empirical experiments, we pre-train *noisy* medical VLMs using crafted adversarial data and evaluate the performance of such noisy models when fine-tuned on various downstream tasks including chest X-ray classification and medical VQA (§5).

2 Related Work

2.1 Medical Vision-Language Models

Some recent efforts have broadened the scope of VLMs to the medical field for a variety of applications (i.e. medical report generation, medical VQA). For instance, MedViLL (Moon et al., 2022) generates medical reports from images, aiding in clinical interpretation. PubMedCLIP (Eslami et al., 2023), pre-trained on the ROCO dataset (Pelka et al., 2018b), is fine-tuned for medical VQA to answer clinically relevant questions from visual inputs. BiomedCLIP (Zhang et al., 2024) adapts CLIP for biomedical applications, focusing on textual descriptions of medical images. LLaVa-Med (Li et al., 2023) demonstrates advanced multi-modal conversational capabilities, making it highly popular in assisting with inquiries about biomedical images. Such pre-trained models are typically trained on large-scale medical datasets, yet the quality of these datasets remains unexplored, and the robustness of the models has not been thoroughly evaluated.

2.2 Adversarial Robustness

Multi-modal VLMs are particularly vulnerable to adversarial attacks since perturbations can affect both visual and textual modalities. A number of general multi-modal adversarial attack strategies have been developed, targeting multiple tasks simultaneously (Zhou et al., 2024; Yin et al., 2024; Zhao et al., 2023; Cui et al., 2023). Recently, adversarial robustness has also garnered increasing attention in the medical sector. For example, Thota et al. (2024) demonstrated how adversarial attacks on pathology images can mislead the Pathology Language-Image Pretraining (PLIP) model. Similar strategies have been employed in radiology to exploit the adversarial vulnerabilities of models used for medical imaging (Bortsova et al., 2021; Finlayson et al., 2019).

To improve robustness, *adversarial training* has been shown to be one of the most effective approaches (Shafahi et al., 2019; Zhang et al., 2019; Paul et al., 2020; Xu et al., 2021; Bai et al., 2021). Training with adversarially perturbed samples enhances resistance to attacks but is time- and compute-intensive, especially for VLMs. Recent studies have investigated parameter-efficient tuning techniques (Mao et al., 2023; Ji et al., 2024) to reduce the computational burden. An alternative approach is *adversarial purification* (Nie et al., 2022; Wang et al., 2022), which uses diffusion models to transform adversarial examples back into clean representations. These methods all attempt to enhance robustness during pre-training, we submerge the adversarial noise during fine-tuning in a privacy-protected (Liu et al., 2022) and light-weight black-box paradigm, assuming that pre-trained models are not always available.

2.3 Noise Learning and Robustness

To address the challenges posed by noisy data during training, recent progress generally falls along two lines: robust model training and adapting clean pre-trained models on noisy (downstream) datasets.

Under the first taxonomy, techniques such as noise estimation (Hendrycks et al., 2019; Jiang et al., 2019; Xia et al., 2019; Yao et al., 2021; Goldberger and Ben-Reuven, 2017), robust loss functions (Ghosh et al., 2017; Ma et al., 2020) have been developed to mitigate the impact of noisy labels. Specifically, Xue et al. (2022) proposes a robust co-training schema for medical image classification that iteratively filters out noisy samples. On the other hand, leveraging clean pre-trained models to adapt to noisy downstream datasets has proven to be both practical and efficient. Under this paradigm, effective fine-tuning strategies have been explored to enhance model robustness against noisy data (Wu et al., 2022; Zhang et al., 2022).

Until recently, Chen et al. (2024) introduces *noisy model learning*, which focuses on the effect of pre-training label noise on downstream. To our knowledge, no previous research has qualitatively assessed the effects of adversarial noise during pre-training on downstream tasks, particularly focusing on multi-modal noise in the medical domain.

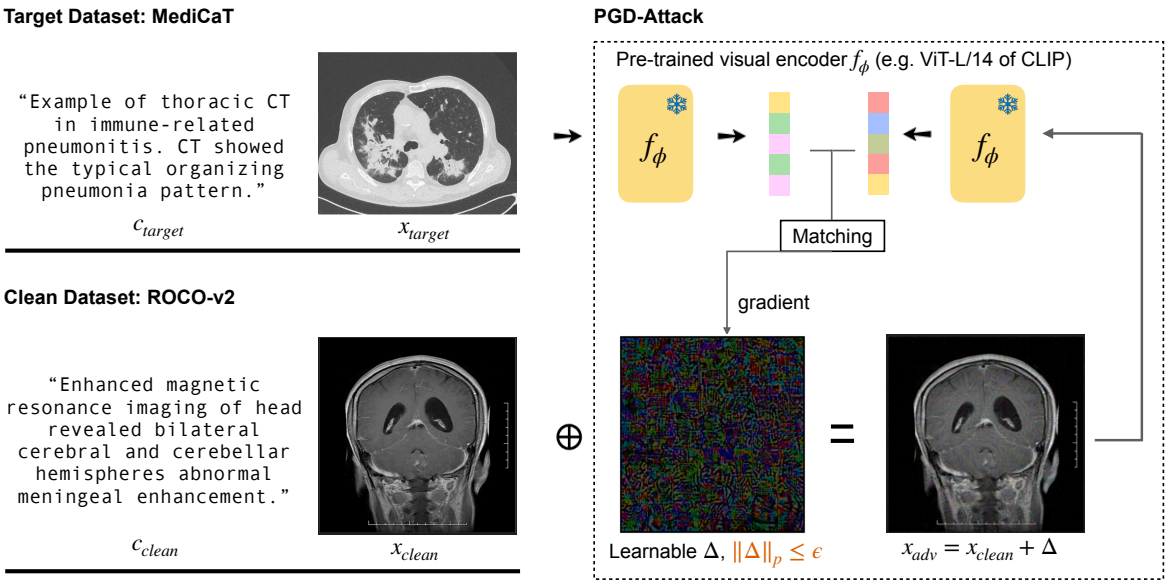


Figure 2: **Pipeline of our radiology image attacking strategy**. We select a target image-caption pair $(x_{\text{target}}, c_{\text{target}})$ from MEDICAT and a clean pair $(x_{\text{clean}}, c_{\text{clean}})$ from ROCOv2. x_{target}. Images are transformed to embeddings by pre-trained visual encoder f_ϕ. Adversarial example x_{adv} is generated by PGD (Eq 3.1) iteratively and we denote the adversarial noise as Δ. As formulated in Eq 3.1, we optimize the Δ to maximize the similarity between x_{target} and x_{adv}; the perturbation Δ is also limited by $\|\Delta\| \leq \epsilon$.

3 Methods

In this section, we first introduce our novel multimodal adversarial attack strategies designed to create noisy medical image-caption datasets, which will be used to train CLIP models. Then, we introduce RAN fine-tuning to alleviate the effect of such pre-trained noisy medical models on downstream classification tasks.

3.1 Adversarial Noisy Dataset Generation

Notation Consider a clean dataset $D_{\text{clean}} = \{(x^i, c^i)\}_{i=1}^N$ for upstream training, where N is the total number of samples in the dataset. Each tuple includes a training image x^i, its corresponding textual caption c^i. Our goal is to generate a noisy dataset D_{noisy} to train a noisy CLIP model M_{noisy}. We use noise ratio γ to denote the percentage of noisy samples in $\hat{D}_{\text{noisy}}$. Each noisy sample (x^i_{adv}, c^i) or (x^i, c^i_{adv}) is created through one of the following adversarial methods: *image attack* or *caption attack*.

Adversarial Image Attack. To craft adversarial images x_{adv} from clean images x_{clean} that can deceive *victim* models, we use an image encoder f_ϕ from a publicly accessible model, i.e., ViT-L/14 of pre-trained CLIP models, as the surrogate model. Considering VLMs may be unreliable for optimizing cross-modality similarity (Zhao et al., 2023), we select a target image x_{target} instead of a target caption c_{target} to guide the generation of x_{adv}, ensuring x_{target} and x_{clean} comes from different data distribution. The adversary aims x_{adv} to resemble x_{target} through human imperceptible perturbations:

$$\underset{\|x_{\text{clean}} - x_{\text{adv}}\|_p \leq \epsilon}{\arg\max} \; f_\phi(x_{\text{adv}})^\top f_\phi(x_{\text{target}})$$

We utilize projected gradient descent (PGD) (Madry et al., 2019) to address the constrained optimization problem presented. PGD iteratively applies gradient ascent on x_{adv} to maximize the cross-entropy loss $\mathcal{L}$. Each iteration is characterized as

$$x^{(t+1)} = \Pi\left(x^{(t)} + \alpha \cdot \text{sign}\left(\nabla_x \mathcal{L}(\theta, x^{(t)}, c)\right)\right)$$

Here, $x^0 = x_{\text{clean}}, \Pi$ is a projection to guarantee adversarial perturbation remains within the acceptable limits. The process to generate adversarial image samples is illustrated in Figure 2.

Adversarial Caption Attack. Prompt-based adversarial attacks are capable of independently and effectively discovering the weaknesses of a victim LLM (Xu et al., 2023). Given (x^i, c^i) in the pre-train dataset D_{clean}, c^i represents a caption describing a radiology image x_i in our case. We alter the

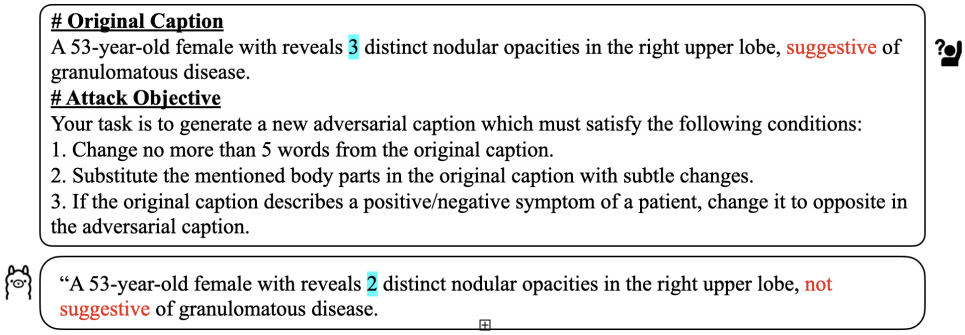


Figure 3: Our proposed **prompt to generate an adversarial caption** of a corresponding radiology image. Highlighted are "body part" designed to change by prompt; Words are changed to opposite by prompt.

captions c^i with LLAMA3-8B[1] using a prompt that incorporates three attack objectives, as depicted in Figure 3. These objectives generate adversarial captions c_{adv} that fail to accurately describe the corresponding radiology image, with few key words (i.e. body parts, symptom) adjustment.

3.2 RAN: Rectify Adversarial Noise

Our fine-tuning objective consists of three components: a covariance loss to attenuate noise impact, a consistency loss, and an adversarial loss to defend adversarial attack in classification tasks.

Covariance Loss. Chen et al. (2024) observed that introducing noise diminishes the top dominant singular values of the pre-trained features, leading to reduced transferability. Building on this insight, we transform pre-trained features $\mathcal{F}$ into a new feature space $\mathcal{Z}$ using multi-layer perceptron (MLP) with covariance regularization term (Bardes et al., 2022) to rectify effects of the introduced noise :

$$\mathcal{L}_{\text{COV}} = \frac{1}{D}\sum_{i\neq j}[C(\mathcal{Z})]_{i,j}^2,$$

where $C(\mathcal{Z})$ is defined as the covariance matrix of transformed features $\mathcal{Z}$:

$$C(\mathcal{Z}) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})(z_i-\bar{z})^T, \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

By minimizing the off-diagonal coefficients of $C(\mathcal{Z})$ to approach zero, we encourage the features to encode discriminative information.

Consistency Loss. To maintain the pre-trained knowledge unchanged, we use a mean-square-error (MSE) loss between the normalized features $\mathcal{F}$ and $\mathcal{Z}$:

$$\mathcal{L}_{\text{MSE}} = \left\|\hat{\mathcal{F}} - \hat{\mathcal{Z}}\right\|_2^2.$$

Here, $\hat{\mathcal{F}} = \frac{\mathcal{F}}{\|\mathcal{F}\|_2}$ and $\hat{\mathcal{Z}} = \frac{\mathcal{Z}}{\|\mathcal{Z}\|_2}$. This objective aids in transferring the pre-trained knowledge to the transformed features Z.

Adversarial Loss. Cross-entropy loss often struggles to distinguish adversarial samples in the feature space because it does not explicitly enforce a robust margin between learned classes (Xia et al., 2022).

Given (x_i, y_i) in a classification task, f_i denotes the features of x_i from pre-trained noisy model M_{noisy}. To address this issue and enhance the robustness of the trained classifier against adversarial attacks, we introduce a constraint to maximize: 1) the distance between the features f_i of a given class y_i and the learned centroids of other classes, and 2) the separation between the learned centroids of different classes:

$$\mathcal{L}_{\text{ADV}} = -\frac{1}{D}\sum_{i=1}^{D}\text{dist}(f_i) + \arccos(c_{y_i}\cdot c_j),$$

$$\text{dist}(f_i) = \frac{1}{k-1}\sum_{j\neq y_i}^{k-1}\|f_i - c_{y_i}\|$$

The c_{y_i} denotes the y_ith class center of features. $\text{dist}(f_i)$ encourages the f_i to be away from wrong classes' centroids. $\mathcal{L}_{\text{ADV}}$ enables the decision margins between the centroids of the classes to be separated sufficiently to prevent the overlapping of features from different classes.

The overall loss function for downstream classification tasks becomes :

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\cdot(\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{COV}}) + \beta\cdot\mathcal{L}_{\text{ADV}}$$

[1] https://ai.meta.com/blog/meta-llama-3/

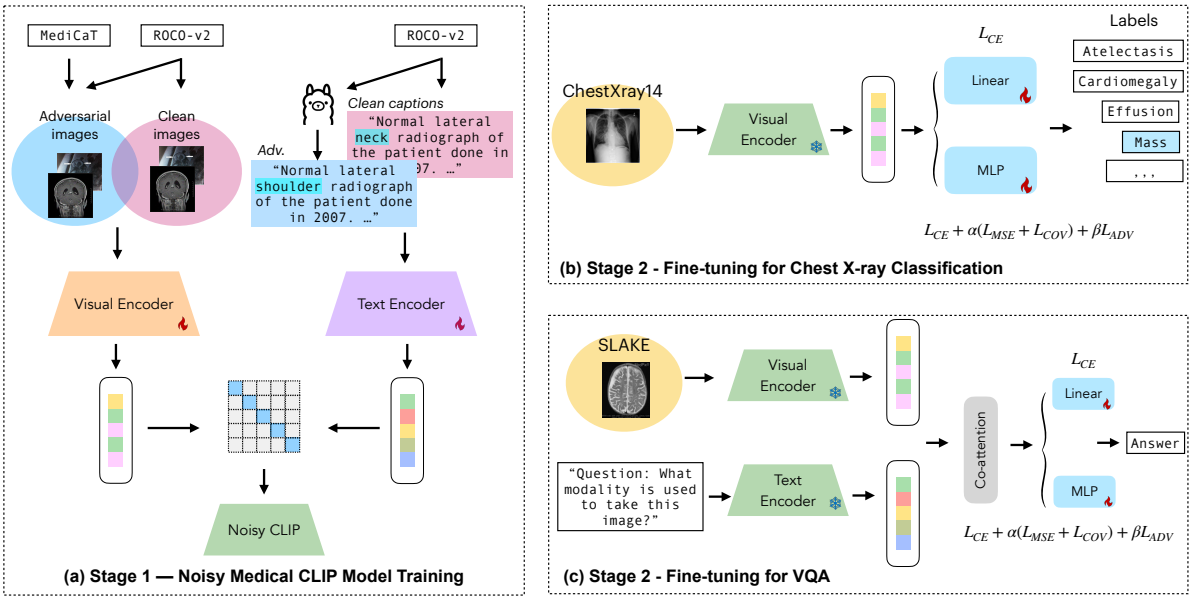


Figure 4: Illustration of **(a)** training a noisy model with a combination of adversarial and clean data. The trained noisy model is then fine-tuned on **(b)** chest x-ray classification task and **(c)** medical VQA task. In **(c)**, we employ a co-attention module to fuse textual and visual features before feeding into a classifier. The classifier can be either a linear classification head or an MLP.

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for classification. We empirically set $\alpha = 0.01, \beta = 0.015$ and use a 2-layer MLP consistently for fair comparison.

4 Experiments

4.1 Training Data

In our experiments, we explore using a well-known radiology dataset, ROCOv2 (Rückert et al., 2024) to pre-train CLIP models. To introduce adversarial noise in the dataset, we select radiology images from MEDICAT (Subramanian et al., 2020) as target to conduct adversarial image attacking, perturbing clean images from ROCOv2.

For downstream tasks, we fine-tune pre-trained noisy models on CHESTXRAY14 (Wang et al., 2017) for classification, and SLAKE (Liu et al., 2021) for VQA, respectively. Detailed dataset resources, statistics and examples are provided in Appendix B.1.

4.2 Noisy Model Pre-training

As shown in Figure 4 (a), we first pre-trian CLIP model (ViT-L/14) on adversarial noisy dataset. The noise ratio γ is set to $\{0\%, 5\%, 10\%, 20\%, 30\%\}$, where 0% representing the clean dataset. We randomly select γ percentage of image-caption pairs from ROCOv2 to attack. To generate image-noisy datasets, we apply adversarial image attack to the selected images. To generate caption-noisy datasets, we perform adversarial caption attack using LLAMA3-8B, as outlined in §3.1. These noisy models are designed to align radiology images with their corresponding captions, allowing us to analyze the impacts of noise on pre-trained feature extractors, and assess performance differences in downstream tasks. Implementation details on training can be found in Appendix B.4

4.3 Fine-tuning

We conduct fine-tuning under three settings: i). *linear probing* (Radford et al., 2021), wherein only training a simple linear classifier on top of the frozen features extracted from the noisy models to analyze how upstream noise affects downstream tasks; ii). *MLP-tuning*, where training an MLP classifier without loss regularization; iii). *RAN-finetuning*, using MLP with proposed loss functions.

Chest X-ray Classification Given a chest x-ray scan, our fine-tuning objective is to predict possible disease labels from 14 categories[2].

Medical VQA To thoroughly evaluate the effectiveness of our method, we then formulate Med-VQA as a classification task, where the possible label set consists of all possible answers. Motivated by Dou et al. (2022), we adopt a transformer-based

[2]Details about classification labels are in Appendix B.1

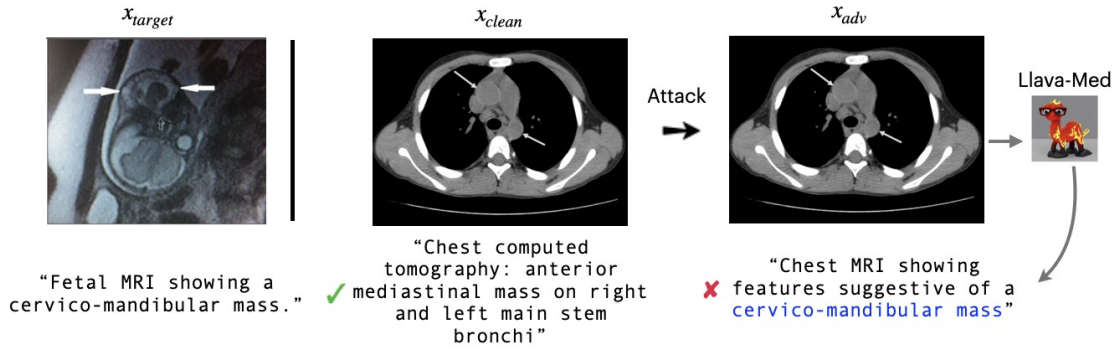


Figure 5: An example of generating caption c_{adv} of crafted adversarial image x_{adv} by black-box VLM, Llava-Med. The default prompt is *"what is the content of this radiology image?"*. ✗ denotes the generated caption doesn't accurately describe the content of the clean image. ✓ means otherwise.

co-attention multi-modal fusion module[3] that produces cross-modal representations over the image and text encodings, which are then fed to a classifier for predicting the final answer as described in Figure 4.

4.4 Evaluation

Domain Differences To investigate the generalizability of adversarial noisy model comprehensively, we then conduct fine-tuning under two experimental settings: i). a standard **in-domain** (ID) setup, in which both the training and testing data are sourced from the same dataset; ii). a more challenging **out-of-domain** (OOD) setup, wherein test data originated from a different dataset.

For ID evaluation, we evaluate the pre-trained noisy models on CHESTXRAY14 for chest-xray classification and SLAKE for medical VQA. Under the OOD setting, we use CHEXPERT (Irvin et al., 2019) and VQA-RAD (Lau et al., 2018) respectively. We report performance on both ID and OOD with $\{0\%, 5\%, 10\%, 20\%, 30\%\}$ percentage of downstream datasets.

Metrics All models are evaluated using the macro average of the area under the receiver-operator curve (AUC) (Bradley, 1997) and accuracy (ACC) averaged over all labels.

5 Results and Analysis

5.1 Effectiveness of Adversarial Attack

In Table 1, we evaluate the efficacy of our adversarial attack strategy against white-box models including pre-trained general CLIP (Radford et al., 2021) and medical CLIP models. We use 5K clean images x_{clean} from the ROCOV2 validation set and randomly select a targeted images-caption pairs $(x_{\text{target}}, c_{\text{target}})$ from MEDICAT for each clean image to craft adversarial images x_{adv} following the method described in Figure 2. We discover that the similarity between x_{adv} and c_{target}, measured by the CLIP score, increases compared to x_{clean}, which validates the effectiveness of our image attack. In addition, Medical CLIP models are more adept at accurately identifying the content of radiology images (as evidenced by a lower score between x_{clean} and c_{target}). However, they remain susceptible to our adversarial attack method, which lays the foundation for black-box transferability (See Appendix A for details). Figure 5 shows LLAVA-MED can be misled to generate inaccurate captions for our adversarially crafted images.

Model	Clean Image	Adv. Image
CLIP - ViT-L/14	0.253	0.384
CLIP - Resnet50	0.211	0.329
PubMedCLIP (Eslami et al., 2023)	0.182	0.347
BioMedCLIP (Zhang et al., 2024)	0.174	0.312

Table 1: **White-box image attacks**. We report the CLIP similarity score between the clean x_{clean} or crafted adversarial images x_{adv} and the corresponding targeted captions c_{target} from MEDICAT.

Model	Clean Caption	Adv. Caption
UniDiffuser (Bao et al., 2023)	0.431	0.274
LLaVA-Med (Li et al., 2023)	0.565	0.392
Mini-GPT4 (Zhu et al., 2023)	0.493	0.287

Table 2: **Black-box caption attacks**. We use VLMs to generate a radiology image based on either a clean caption c_{clean} or c_{adv}, and report CLIP score between the generated image (i.e., $\hat{x}_{\text{clean}}$ and $\hat{x}_{\text{adv}}$) and x_{clean}.

In Table 2, we transfer the crafted adversarial captions to image through advanced VLMs. The

[3]Detailed descriptions of medical VQA setting is in Appendix B.3

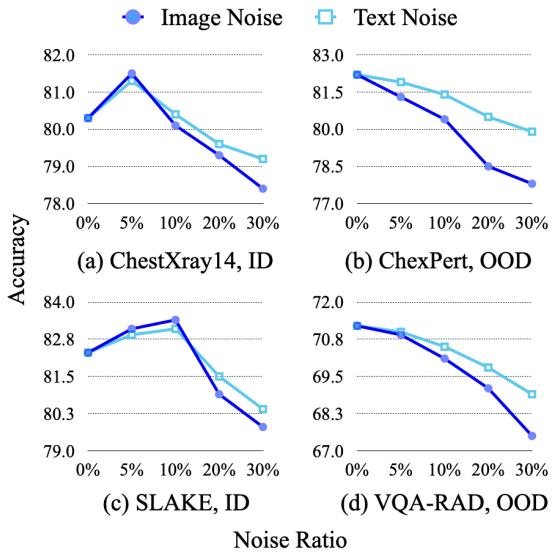


Figure 6: **Linear Probing ID and OOD evaluation** results of CLIP model pre-trained on multi-modal adversarial noise on downstream tasks including chest x-ray classification ((a) and (b)) and medical VQA ((c) and (d)) with various percentages of noise.

similarity between the generated image $\hat{x}_{\text{adv}}$ are less similar to the clean image x_{clean} than generated $\hat{x}_{\text{clean}}$, indicating the effectiveness of caption attack.

5.2 Adversarial Noise Evaluation

To explore the effects of adversarial multi-modal noise from upstream training on downstream tasks, we show the accuracy of evaluating pre-trained noisy models on both ID and OOD tasks across all noise ratios, under linear probing setting, in Figure 6. We empirically reveal the following insights:

Introducing a moderate level of noise, such as 5% or 10%, during pre-training can actually improve a model's robustness and performance on ID downstream tasks. We hypothesize this slight noise acts as a form of regularization, helping the model generalize better to similar data seen during fine-tuning. However, increasing the noise beyond this threshold starts to degrade the model's performance, leading to poorer results. This finding aligns with Song et al. (2022), suggesting a balance in noise levels is critical for optimal model training, as excessive noise can introduce too much variability.

The performance on OOD downstream tasks consistently diminishes as the noise level in pre-training increases. High levels of noise make it harder for the model to adapt to new and unseen data, reducing its ability to generalize effectively beyond the training domain.

γ	Setting	CHEST-XRAY14
0	Base	80.3
	Ours (LP)	81.3
5	Random (LP)	80.6
	Random + RAN	81.5
	Ours (LP)	80.4
10	Random (LP)	79.8
	Random + RAN	80.8
	Ours (LP)	79.6
20	Random (LP)	79.1
	Random + RAN	80.2
	Ours (LP)	78.9
30	Random (LP)	78.2
	Random + RAN	79.4

Table 3: Performance Comparison with Random Noise

Noise Type	γ	CHESTXRAY14		CHEXPERT	
		AUC	ACC	AUC	ACC
Image	0	65.2	72.4	68.6	76.1
	5	65.5 ↑	72.9 ↑	69.5 ↑	76.8 ↑
	10	64.7	71.2	69.3 ↑	76.3 ↑
	20	64.1	70.6	67.8	75.4
	30	63.5	69.9	67.1	74.9
Caption	5	65.8 ↑	72.8 ↑	69.0 ↑	76.4 ↑
	10	65.4 ↑	72.2	68.4	76.0
	20	64.7	71.9	68.3	75.8
	30	64.1	71.2	67.6	75.2

Table 4: **Zero-shot Evaluations** on Chest X-ray classification tasks across different pre-trained noise ratios (γ). (↑) indicates improvements from the clean baseline.

Image attacks tend to be more potent than caption attacks in affecting model performance. Specifically, across all four tasks, models subjected to image noise exhibit more significant changes than those exposed to text noise. This suggests that image perturbations can disrupt the model's internal representations more effectively, leading to greater performance improvement or degradation.

Comparing with Random Noise

To disentangle the two factors (how downstream performance differs between our crafted adversarial caption noise, and just random noise?), we also test additional baselines that introduce small amount of random gaussian noise to caption input (Table 3). As expected, when noise is 5%, random noise has a slightly lesser improvement on down-

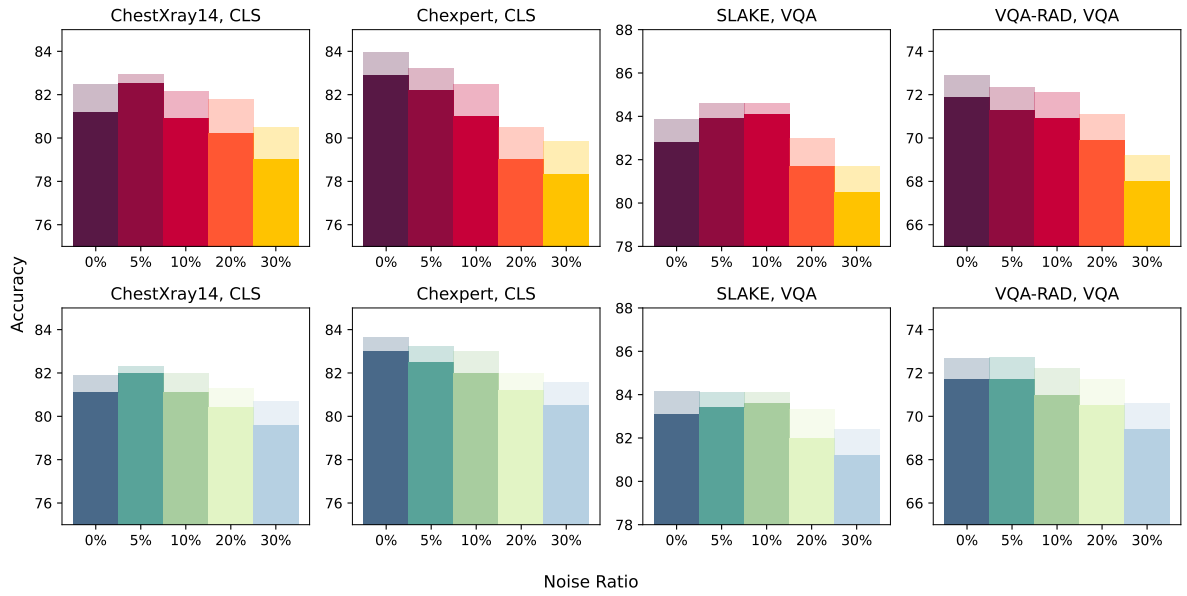


Figure 7: **Evaluation of RAN fine-tuning** on ID and OOD downstream tasks, compared to MLP tuning. We use CLIP models pre-trained on noisy ROCOv2 dataset with, [**first row**]: adversarial images; and [**second row**]: adversarial captions. The improvements of RAN are presented by stacked bars with light colors.

stream performance compared to our crafted noise; as noise increases, random noise deteriorates performance more rapidly. Nevertheless, our study confirms that our noise mitigation strategy, RAN, is effective not only against our crafted adversarial noise but also against random noise.

5.3 Zero-shot Evaluation

To comprehensively study the robustness of pre-trained noisy models without fine-tuning, we perform a zero-shot chest x-ray classification on two datasets: CHESTXRAY14 and CHEXPERT. To match text with the encoded image embeddings, we use prompts of *{label}* and *No {label}* (e.g., "Atelectasis" vs. "No atelectasis") following You et al. (2023). The results are illustrated in Table 4.

Following the findings from §5.2, introducing slight noise enhances pre-trained model robustness and performs better on zero-shot classification tasks, while excessive noise in turn hurts the performance.

5.4 Effectiveness of RAN Fine-tuning

Figure 7 shows the effect of our proposed RAN fine-tuning on mitigating upstream adversarial noise. To disentangle two possible reasons–*RAN regularization* or *extra parameters from MLP*–improve models' robustness, we compare it against baselines using MLP-tuning.

From experimental results, incorporating RAN enhances overall performance across all ID and OOD tasks against both image and caption attacks. The performance improvement observed with a 5% noise ratio in the ID CHESTXRAY task and a 10% noise ratio in the SLAKE task under linear probing setting is less pronounced with RAN fine-tuning, indicating its effectiveness in rectifying the influences of noise. Particularly for OOD tasks, the improvement is slightly more significant. Moreover, we notice that the improvement on caption-noisy models is less significant than image-noisy models when noise starting to degrade model performance, potentially because the impact of image noise on feature extractors is greater, and RAN rectify such effects during fine-tuning (See Appendix 10 for full results).

Given that SOTA medical VLMs are generally pre-trained on large amounts of data (e.g., BioMedCLIP pretrained on 15M private medical image-text pairs), we do not believe it's fair to directly

Model	Setting	SLAKE	VQA-Rad
BIOMEDCLIP	Baseline	88.9	79.8
	with RAN	90.2	80.9
PUBMEDCLIP	Baseline	82.5	80.0
	with RAN	83.1	80.5
LLAVA-MED	Baseline	84.2	85.3
	with RAN	85.1	85.2
PMC-CLIP	Baseline	88.0	84.0
	with RAN	89.7	84.8

Table 5: Comparison with other baseline VLMs.

γ	Setting	CHEST-XRAY	CHEXPERT
5	LP	81.5	81.3
	RAN	82.9	83.2
	NMTune	82.6	82.7
10	LP	80.1	80.4
	RAN	82.2	82.5
	NMTune	81.5	81.6
20	LP	79.3	78.5
	RAN	81.8	80.5
	NMTune	80.7	79.6
30	LP	78.2	77.8
	RAN	80.5	79.8
	NMTune	79.6	78.7

Table 6: Comprison with NMTune.

$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{COV}}$	$\mathcal{L}_{\text{ADV}}$	γ	CHESTXRAY14	SLAKE
✗	✗	✗	0	81.2	82.8
✓	✗	✗	0	81.2	83.0
✗	✓	✗	0	81.7	83.3
✗	✗	✓	0	81.4	83.2
✓	✓	✗	0	81.9	83.4
✓	✓	✓	0	82.5	83.8
✗	✗	✗	20	80.2	81.8
✓	✗	✗	20	80.5	82.1
✗	✓	✗	20	80.8	82.3
✗	✗	✓	20	80.9	82.1
✓	✓	✗	20	81.3	82.7
✓	✓	✓	20	81.8	83.0

Table 7: **Ablation Study** on loss terms of ID tasks with image attack. γ denotes noise ratio in pre-trained dataset. Highlighted denotes improvement ≥ 0.5.

compare the finetuned results. However, how our noise mitigation strategy RAN would fare on other SOTA VLMs is indeed an interesting question. Fine-tuning such VLMs with RAN can test if RAN is resilient against various types of noise, as such VLMs might pre-trained with unknown noise, we present the following results with finetuning SOTA VLMS on medical VQA tasks, with our proposed RAN as noise mitigation: As shown in Table 5, we can see that with RAN, almost all SOTA VLMs exhibit better performance across the board, which validates the effectiveness of our proposed noise mitigation fine-tuning strategy. With only PubMedCLIP has less improvement. We hypothesize this is because PubMedCLIP was pre-trained on a small-scale, high-quality dataset, whereas the others were pre-trained on larger-scale datasets, which may contain more noise.

In Table 6, we present the experimental results on chest classification task between applying RAN and another noise model mitigation approach NMTune (Chen et al., 2024). Our method performs better on rectifying adversarial noise than NMTune in the above case. We hypothesize it's because NMTune tries to mitigate label noise, and focuses on rectifying the features shaped by such noise, whereas our adversarial loss regularization mainly focuses on against adversarial noises.

Ablations We perform extensive ablations to show that every component of RAN benefits the overall system (Table 7). The results indicate that our proposed $\mathcal{L}_{\text{COV}}$ and $\mathcal{L}_{\text{ADV}}$ effectively mitigate the impact of adversarial image noise. While the improvement from using only $\mathcal{L}_{\text{MSE}}$ is relatively modest, $\mathcal{L}_{\text{ADV}}$ and $\mathcal{L}_{\text{COV}}$ shows limited enhancement in performance for models with clean upstream data ($\gamma = 0$) compared to noisy models ($\gamma = 20$). We hypothesize that this is because $\mathcal{L}_{\text{ADV}}$ and $\mathcal{L}_{\text{COV}}$ are primarily designed to address feature changes induced by upstream noise, which may not provide significant benefits in clean datasets. Combining all loss terms proposed in RAN effectively improves performance against MLP-tuning.

6 Conclusion

Despite the success development of medical VLMs, most such models are vulnerable to adversarial attack and still lag behind in transferring to downstream tasks robustly. In this work, we discuss how upstream adversarial noise affects various medical downstream tasks by crafting adversarial multimodal medical samples. Through extensive experiments, we found that even minor adversarial noise in pre-training datasets can enhance ID performance while degrading OOD generalization. To this end, we introduce a light-weight fine-tuning recipe, RAN, effectively mitigating noise effects by refining the feature space and enforcing robust margins to defend adversarial noise.

7 Limitations

Several limitations restrict the scope of our work. To begin, our choice of downstream tasks–chest X-ray classification and medical VQA tasks–is nonexhaustive, and it is possible that our findings would not generalize well for the broad spectrum of medical applications. Given that medical datasets can be quite limited compared to other general datasets due to its private nature, pretrain a VLMs are also

limited. Another restriction is that we only attacked radiology images and their captions, offering a glimpse into possible vulnerabilities but not a complete picture. This means our findings may not apply to other kinds of medical images or related text data. Expanding to various medical imaging and datasets in future work will be crucial for more comprehensive insights and real-world applicability.

Other potential avenues for exploration entail different noise type and evaluate the nature of how noise shapes pre-trained features can be useful. Future work should explore optimizing noise levels and further enhancing the robustness of VLMs to various adversarial scenarios to maintain high performance across diverse medical domains

Acknowledgment

This work is partially supported by NSF NAIRR240016, NIH R21EB034911, and Google Cloud research credits.

References

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. *Preprint*, arXiv:2102.01356.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. 2023. One transformer fits all distributions in multi-modal diffusion at scale. *Preprint*, arXiv:2303.06555.

Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *Preprint*, arXiv:2105.04906.

Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2023. Generating medical prescriptions with conditional transformer. *Preprint*, arXiv:2310.19727.

Gerda Bortsova, Cristina González-Gonzalo, Suzanne C. Wetstein, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bart Liefers, Bram van Ginneken, Josien P.W. Pluim, Mitko Veta, Clara I. Sánchez, and Marleen de Bruijne. 2021. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141.

Andrew P. Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. 2024. Understanding and mitigating the label noise in pre-training on downstream tasks. *Preprint*, arXiv:2309.17002.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. *Preprint*, arXiv:2212.07143.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. 2023. On the robustness of large multimodal models against image adversarial attacks. *Preprint*, arXiv:2312.03777.

Zolnamar Dorjsembe and Furen Xiao. 2023. Synthetic whole-head mri brain tumor segmentation dataset. *IEEE Dataport*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An empirical study of training end-to-end vision-and-language transformers. *Preprint*, arXiv:2111.02387.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. Datacomp: In search of the next generation of multimodal datasets. *Preprint*, arXiv:2304.14108.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. *Preprint*, arXiv:1712.09482.

Jacob Goldberger and Ehud Ben-Reuven. 2017. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*.

Alex Havrilla and Maia Iyer. 2024. Understanding the effect of noise in llm training data with algorithmic chains of thought. *Preprint*, arXiv:2402.04004.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2019. Using trusted data to train deep networks on labels corrupted by severe noise. *Preprint*, arXiv:1802.05300.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Preprint*, arXiv:1901.07031.

Yuheng Ji, Yue Liu, Zhicheng Zhang, Zhao Zhang, Yuting Zhao, Gang Zhou, Xingwei Zhang, Xinwang Liu, and Xiaolong Zheng. 2024. Advlora: Adversarial low-rank adaptation of vision-language models. *Preprint*, arXiv:2404.13425.

Junjun Jiang, Jiayi Ma, Zheng Wang, Chen Chen, and Xianming Liu. 2019. Hyperspectral image classification in the presence of noisy labels. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):851–865.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2023. Ehrsql: A practical text-to-sql benchmark for electronic health records. *Preprint*, arXiv:2301.07695.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *Preprint*, arXiv:2102.09542.

Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, C-C Jay Kuo, Georges El Fakhri, Je-Won Kang, and Jonghye Woo. 2022. Unsupervised domain adaptation for segmentation with black-box source model. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 255–260. SPIE.

Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2024. Machine learning for synthetic data generation: A review. *Preprint*, arXiv:2302.04062.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. *Preprint*, arXiv:2006.13554.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks. *Preprint*, arXiv:1706.06083.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. 2023. Understanding zero-shot adversarial robustness for large-scale models. *Preprint*, arXiv:2212.07016.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. 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.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. 2022. Diffusion models for adversarial purification. *Preprint*, arXiv:2205.07460.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo,

Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rahul Paul, Matthew Schabath, Robert Gillies, Lawrence Hall, and Dmitry Goldgof. 2020. Mitigating adversarial attacks on medical image understanding systems. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1517–1521.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and C. Friedrich. 2018a. Radiology objects in context (roco): A multimodal image dataset. In *CVII-STENT/LABELS@MICCAI*.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. 2018b. Radiology objects in context (roco): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189, Cham. Springer International Publishing.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *Preprint*, arXiv:2102.12092.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. 2024. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Preprint*, arXiv:2405.10004.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Preprint*, arXiv:1904.12843.

Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. 2024. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. *Preprint*, arXiv:2405.20469.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *Preprint*, arXiv:2007.08199.

Sanjay Subramanian, Sachin Mehta Lucy Lu Wang, Madeleine van Zuylen Ben Bogin, Sravanthi Parasa, Matt Gardner Sameer Singh, and Hannaneh Hajishirzi. 2020. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In *Findings of EMNLP*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *Preprint*, arXiv:1908.07490.

Poojitha Thota, Jai Prakash Veerla, Partha Sai Guttikonda, Mohammad S. Nasr, Shirin Nilizadeh, and Jacob M. Luber. 2024. Demonstration of an adversarial attack against a multimodal vision language model for pathology imaging. *Preprint*, arXiv:2401.02565.

Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. 2022. Guided diffusion model for adversarial purification. *Preprint*, arXiv:2205.14969.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2022. Noisytune: A little noise can help you finetune pretrained language models better. *Preprint*, arXiv:2202.12024.

Pengfei Xia, Ziqiang Li, and Bin Li. 2022. Tightening the approximation error of adversarial risk with auto loss function search. *Preprint*, arXiv:2111.05063.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. 2019. Are anchor points really indispensable in label-noise learning? In *NeurIPS*.

Han Xu, Xiaorui Liu, Yaxin Li, Anil K. Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. *Preprint*, arXiv:2010.06121.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2023. An llm can fool itself: A prompt-based adversarial attack. *Preprint*, arXiv:2310.13345.

Cheng Xue, Lequan Yu, Pengfei Chen, Qi Dou, and Pheng-Ann Heng. 2022. Robust medical image classification from noisy labeled data with global and local representation guided co-training. *Preprint*, arXiv:2205.04723.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. 2021. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Preprint*, arXiv:2006.07805.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2024. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Preprint*, arXiv:2310.04655.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K. Hong, Woonhyuk Baek, and Byungseok Roh. 2023. *CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training*, page 101–111. Springer Nature Switzerland.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. *Preprint*, arXiv:1901.08573.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2024. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *Preprint*, arXiv:2303.00915.

Yupeng Zhang, Hongzhi Zhang, Sirui Wang, Wei Wu, and Zhoujun Li. 2022. Pats: Sensitivity-aware noisy learning for pretrained language models. *Preprint*, arXiv:2210.12403.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. *Preprint*, arXiv:2305.16934.

Wanqi Zhou, Shuanghao Bai, Qibin Zhao, and Badong Chen. 2024. Revisiting the adversarial robustness of vision language models: a multimodal perspective. *Preprint*, arXiv:2404.19287.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

Appendix

A Multi-modal Adversarial Attack

Table 1 validates the effectiveness of white-box attack against CLIP models. In Table 8, we transfer the crafted adversarial examples in order to evade large VLMs and mislead them into generating targeted responses. The similarity between the generated response c_{adv} are more similar to the targeted text c_{target} than c_{clean}, indicating the effectiveness of our method towards advanced large VLMs.

Model	Clean image	Adv. image
UniDiffuser (Bao et al., 2023)	0.287	0.594
LLaVA-Med (Li et al., 2023)	0.246	0.483

Table 8: **Black-box image attacks**. We report CLIP score between the generated caption of input images (i.e., x_{clean} or crafted x_{adv}) and targeted caption c_{target},

B Training

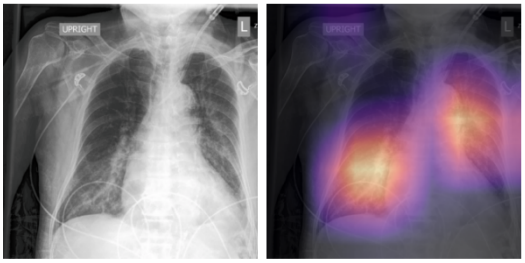
B.1 Datasets

ROCOv2 (Rückert et al., 2024) provides 79,789 radiological images with associated captions and medical concepts. The image–text pairs are captured from PubMed () articles. It is an updated version of the ROCO (Pelka et al., 2018a) dataset published in 2018, and adds 35,705 new images added to PMC since 2018.

MediCaT (Subramanian et al., 2020) includes medical images, captions, subfigure-subcaption annotations, and inline textual references from. It consists of 217,060 figures from 131,410 open access papers, 7,507 subcaption and subfigure annotations for 2,069 compound figures.

ChestXray14 (Wang et al., 2017) is a medical imaging dataset that includes 112,120 frontal-view X-ray images from 30,805 unique patients, collected between 1992 and 2015. It features fourteen common disease labels extracted from radiological reports. The disease categories are: *Atelectasis, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia, No finding*.

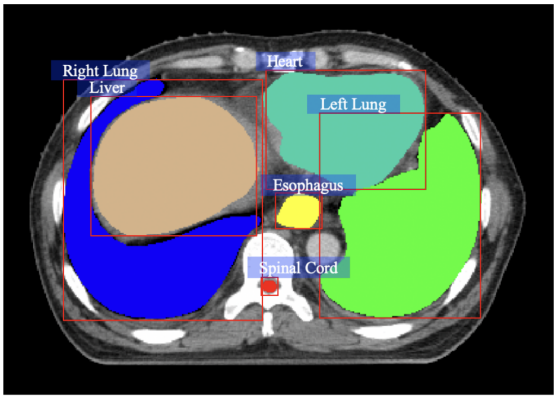
CheXpert (Irvin et al., 2019) is a large dataset of chest X-rays of 65,240 patients, with 14 observation labels collected from Stanford Hospital. The included 14 labels are: *Enlarged Cardiom, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices and No Finding*.



(b) Single frontal radiograph of the chest demonstrates bilateral mid and lower lung interstitial predominant opacities and cardiomegaly most consistent with cardiogenic pulmonary edema. The model accurately classifies the edema by assigning a probability of $p = 0.824$ and correctly localizes the pulmonary edema. Two independent radiologist readers misclassified this examination as negative or uncertain unlikely for edema.

Figure 8: An example of image-caption in CheXpert dataset.

SLAKE (Liu et al., 2021) a bilingual Med-VQA benchmark containing 6428 radiology images (X-rays, CT scans, MRIs) and 14,028 question-answer pairs. It includes both "closed-ended" questions, and more challenging "open-ended" questions. Fo simplicity, we only report performance evaluated on "closed-ended" questions. An example image-question pair is shown in Figure 9.



Question	➢Does the image contain left lung? ➢图片中是否包含左肺?	➢What is the function of the rightmost organ in this picture? ➢图中最右侧器官功能是什么?
Type	Vision-only	Knowledge-based
Answer Type	Closed-ended	Open-ended

Figure 9: An example of image-question in SLAKE dataset.

VQA-RAD (Lau et al., 2018) contains 3515 question–answer pairs generated by clinicians and 315 radiology images that are evenly distributed over the head, chest, and abdomen. Each image is associated with multiple questions. Half of the answers are closed-ended (i.e., yes/no type), while the rest are open-ended with either one-word or short phrase answers.

Dataset	Split	Image #	Caption # / Question #	Answer #
ROCOv2	Train	59,958	59,958	/
	Valid	9904	9904	/
	Test	9927	9927	/
MediCaT	Train	141,089	141,089	/
	Valid	32,559	32,559	/
	Test	43,412	43,412	/
ChestXray14	Train	78,484	/	/
	Valid	11,212	/	/
	Test	22,424	/	/
CheXpert	Train	224,316	/	/
	Valid	235	/	/
	Test	669	/	/
SLAKE	Train	450	9,849	9,849
	Valid	96	2,109	2,109
	Test	96	2,070	2,070
VQA-RAD	Train	315	3,064	3,064
	Test		451	451

Table 9: Medical Dataset Statistics.

B.2 Pre-training with CLIP models

CLIP Contrastive Loss

$$\ell^i_{u\to v} = -\log \frac{\exp(\text{sim}(u_i, v_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(u_i, v_j)/\tau)} \quad (1)$$

Noise Type	γ	Classifier	ChestXray14	CheXpert	SLAKE	VQA-RAD
Image	0	LP	80.3	82.2	82.3	71.2
		MLP	81.2	82.9	82.8	71.9
		LP	81.5	81.3	83.1	70.9
	5	MLP	82.5	82.2	83.9	71.3
		RAN	82.9	83.2	84.6	72.3
		LP	80.1	80.4	83.4	70.1
	10	MLP	80.9	81.0	84.1	70.9
		RAN	82.2	82.5	84.6	72.1
		LP	79.3	78.5	80.9	69.1
	20	MLP	80.2	79.0	81.8	69.9
		RAN	81.8	80.5	83.0	71.1
		LP	78.2	77.8	79.8	67.5
	30	MLP	79.0	78.3	80.5	68.0
		RAN	80.5	79.8	81.7	69.2
Caption	0	LP	80.3	82.2	82.3	71.2
		MLP	81.1	83.0	83.1	71.7
		RAN	81.9	83.6	84.2	72.7
		LP	81.3	81.9	82.9	71.0
	5	MLP	82.0	82.5	83.4	71.7
		RAN	82.3	83.2	84.1	72.7
		LP	80.4	81.4	83.1	70.5
	10	MLP	81.1	82.0	83.6	71.0
		RAN	82.0	83.0	84.1	72.2
		LP	79.6	80.5	81.5	69.8
	20	MLP	80.4	81.2	82.0	70.5
		RAN	81.3	82.0	83.3	71.7
		LP	78.9	79.8	80.4	68.9
	30	MLP	79.6	80.5	81.2	69.4
		RAN	80.7	81.6	82.4	70.6

Table 10: The full results of fine-tuning with linear probing , MLP and RAN across all noise ratios.

where u and v are the normalized vectors from the image and text encoders, respectively. (u_i, v_i) is a positive pair, sim is a function that calculates the similarity between the vectors, and τ is the learnable temperature parameter. N represents the mini-batch size of image-text pairs. $\ell^i_{u\to v}$ denotes the InfoNCE loss from image i to the texts, while $\ell^i_{v\to u}$ represents the loss in the opposite direction. The final loss in CLIP is defined as:

$$L_{\text{CLIP}} = \frac{1}{2N}\sum_{i=1}^{N}(\ell^i_{u\to v} + \ell^i_{v\to u}) \qquad (2)$$

B.3 Medical VQA

Given a MedVQA training dataset denoted as $T = \{(v_i, q_i, a_i)\}_{i=1}^{V}$ of size V, where v_i is a medical image, q_i is the corresponding natural language question, and a_i is the natural language answer, our objective is to learn to generate the correct answer a_i for a given image-question pair (v_i, q_i). The features obtained from the image and question encoder are concatenated as $f_v(v_i) \oplus f_t(q_i)$ We then formulate MedVQA as a multi-label classification function $F : \mathbb{R}^n \times \mathbb{R}^{m\times l} \to \{0, 1\}^{|A|}$, where A is the overall set of possible answers and $F(f_v, f_q) = a_i$ for the one-hot encoded answer a_i.

Multi-modal Fusion Module Figure 10 presents the co-attention architecture we use to fuse visual and text features.

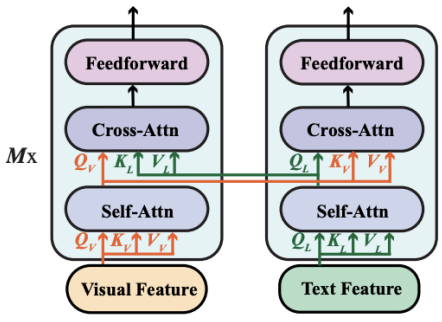


Figure 10: Illustration of multi-modal fusion modules through co-attention architecture.

B.4 Implementation Details

Our implementation is based on OpenCLIP (Cherti et al., 2022). We utilize the CLIP with ViT-L/14 architecture, with input images at a resolution of 240. The model comprises a total of 24 layers, which are divided into 4 stages, each encompassing 6 layers. The CLIP model has been pre-trained on the DATACOMP-1B dataset (Gadre et al., 2023) to ensure robust image-text matching in general domains, which facilitates effective fine-tuning on the medical domain where data is more limited. Following Zhang et al. (2024), we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, a cosine decay learning rate scheduler with an initial value of 5e-4 at batch size of 16, and the warm-up step set to 2000, conducting 30 epochs for training on 4 A40 GPU.

C Full Results of Fine-tuning

Results are presented in Table 10.