No Evidence of Disease: Clinically-Risky Adversarial Chest CT Report Generation

Samra Irshad[1] SAMRA@KHU.AC.KR
[1] *School of Computing, Kyung Hee University, Gyeonggi-do 17104, South Korea*
Junho Kim[2] ARKIMJH@ILLINOIS.EDU
[2] *University of Illinois Urbana–Champaign, Champaign, IL, United States*
Seong Tae Kim[1,*] ST.KIM@KHU.AC.KR
[1] *School of Computing, Kyung Hee University, Gyeonggi-do 17104, South Korea*

Editors: Under Review for MIDL 2026

Abstract

Automated chest CT radiology report generation has equipped clinicians with the ability to automatically describe clinical findings and abnormalities from CT scans. Given that patient prognosis relies heavily on these reports, generating an accurate CT report is critical. Advances in Multimodal Large Language Models (MLLMs) have enabled substantial improvements in CT-to-text report generation models, yet recent studies show that MLLMs are highly susceptible to adversarial perturbations. Beyond this known susceptibility, it remains unclear what triggers clinically dangerous attack scenarios during medical report generation. Understanding such threats is essential for developing robust medical AI systems — without a clear characterization of the threat, it is challenging to mitigate real-world risks. In this paper, we investigate how chest CT report generation models can be adversarially manipulated and what constitutes an adversarial CT report. We introduce **Clinically Risky Adversarial Report Generation (CRA-RG)**, a threat model that defines clinically realistic adversarial alterations to chest CT reports. To instantiate this threat model, we develop a targeted multimodal attack that perturbs both CT volumes and conditioning text prompts to induce clinically risky changes in reports. We show that our attack can successfully omit and fabricate clinically grounded high-risk CT chest findings (e.g., *nodules* or *lesions*). To the best of our knowledge, our study is the first empirical demonstration that state-of-the-art CT report generation models can be deceived into producing harmful clinical decisions, potentially leading to missed diagnoses or unnecessary biopsies. We evaluate our attack on the publicly available chest 3D CT RadGenome dataset.

Keywords: Multimodal Large Language Models, CT Report Generation, Adversarial Attack.

1. Introduction

Examining chest scans for abnormalities across thoracic organs and then articulating the observed findings in a detailed written report are core responsibilities of radiologists. Chest radiology reports typically provide structured, organ-based descriptions of both normal anatomy and pathological findings. Given that the interpretation of chest radiographs required to write radiology reports is time-consuming and depends on specialized expertise,

* Corresponding author.

there has been growing interest in developing automated systems to assist with chest radiology reporting (Everlight, 2025; Chen et al., 2020a; Li et al., 2023b; Zhang et al., 2025a). Automated models for radiology report generation are built on recent advances in multimodal learning, driven mainly by Multimodal Large Language Models (MLLMs) (Li et al., 2023a; Hu et al., 2024; Jian et al., 2024). While prior studies have predominantly focused on chest X-ray report generation, chest CT scans contain much finer anatomical detail and impose a higher interpretive burden on radiologists. As a result, the development of radiology report generation models for chest CT interpretation has only recently begun to receive attention (Chen et al., 2025a; Hamamci et al., 2024).

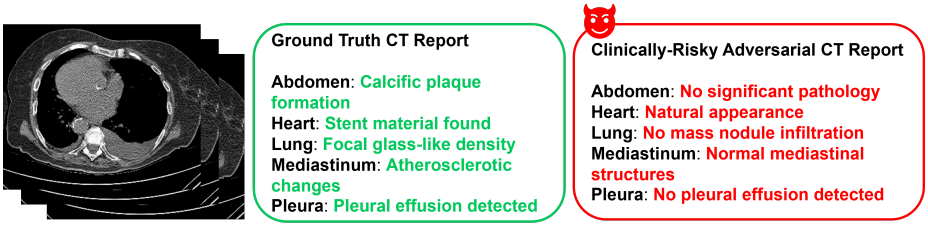

Figure 1: Example of an adversarially manipulated chest CT report. The CT volume is shown with the original CT report comprising organ-wise findings (in green), and the corresponding adversarial report (in red); only key phrases are displayed for brevity.

Despite these successes, it is well established that vision models are vulnerable to adversarial attacks (Goodfellow et al., 2015). Although imperceptible, these adversarial attacks can profoundly distort model outputs, making them particularly dangerous in safety-critical domains (Nguyen et al., 2025). Beyond vision models, recent studies have raised new AI safety concerns by demonstrating that even LLMs are vulnerable, with adversarial prompts capable of bypassing built-in guardrails and manipulating models to generate harmful or factually incorrect responses (Han et al., 2024; Zou et al., 2023; Andriushchenko et al., 2025). Because MLLMs rely on LLMs as their language decoder and additionally incorporate visual inputs, the safety weaknesses of LLMs naturally extend to the multimodal setting. As a result, adversarial prompts introduced through either text or images present a new and significant threat to MLLMs (Aafaq et al., 2021; Zhao et al., 2023; Dong et al., 2023; Shayegani et al., 2024). This poses a significant risk to medical AI, particularly multimodal radiology report generation systems that rely on both image and language inputs and may produce clinically dangerous errors when exposed to adversarial prompts or perturbed images. Understanding these attack vectors is therefore essential for developing robust medical AI systems. Motivated by these concerns, we pose the following question:

Can chest CT radiology report generation models be adversarially attacked, and, more importantly, what does it mean for a radiology report itself to be considered 'attacked'?

In this paper, we introduce **Clinically Risky Adversarial Report Generation (CRA-RG)**, a new adversarial threat that defines how MLLM-based chest CT-to-text report generation models can be attacked at test time to conceal or fabricate critical chest findings. We present an example of an Adversarial CT report in Figure 1, illustrating the

manipulation of critical chest features. To realize this threat model, we develop a multimodal targeted adversarial attack that leverages learnable visual and textual perturbations and injects them into the model's inputs. Specifically, we apply voxel-level adversarial perturbations to both the full chest CT volume and the anatomy of interest (*e.g.,* lung parenchyma or breast region), as well as embedding-level perturbations to the prompt-conditioning embeddings. By jointly attacking the visual and textual representations, our method generates adversarial chest CT reports that remain clinically plausible yet omit and insert high-risk chest CT findings. Our proposed attacking framework is shown in Figure 2.

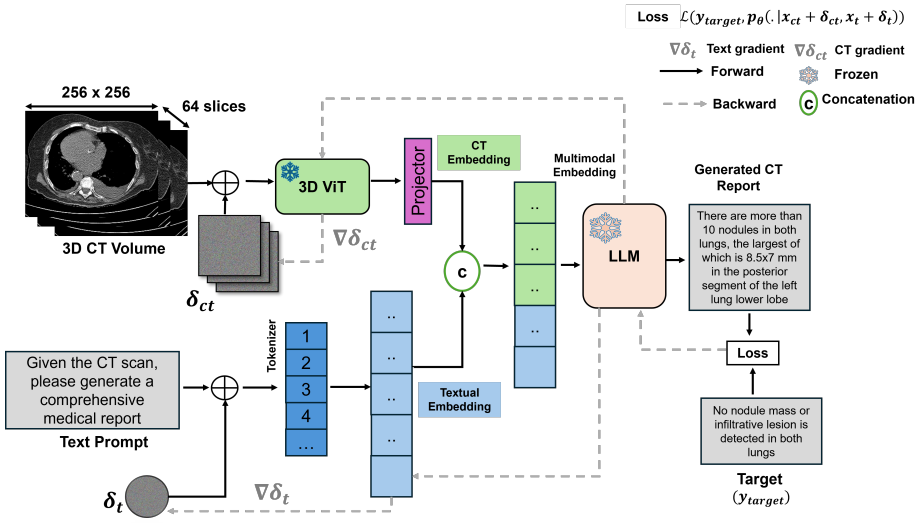

Figure 2: Our proposed framework for Clinically-Risky adversarial CT report generation. Given a 3D chest CT volume and text input at test time, we inject learnable targeted multimodal perturbations δ_{ct} and δ_t into CT volume and text prompt embeddings, respectively, to induce clinically risky report changes.

2. Related Work

2.1. Multimodal Large Language Models (MLLMs)

Typically, MLLMs integrate a visual encoder with a pretrained Large Language Model (LLM) decoder via a connector to learn unified vision–language representations (Liu et al., 2023; Li et al., 2023a; Alayrac eta l., 2022). The visual encoder extracts image embeddings, the connector transforms these embeddings into the language module's latent space, and finally, the language decoder generates the textual response conditioned on multimodal prompts. These representations, in turn, provide the foundation for effective cross-modal reasoning (Touvron et al., 2023; Jiang et al., 2023; Achiam et al., 2023). With access to large-scale text corpora organized in natural-language instruction formats and equipped with large parameter-sized models, LLMs acquire strong instruction-following capabilities. When serving as the language decoder in MLLMs, these capabilities enable the model to reason over visual inputs and generate textual responses when guided by multimodal prompts. Leveraging these advances, recent studies have applied MLLMs to radiology

report generation by fine-tuning them on paired radiology scan–report datasets (Shruthi et al., 2024; Lee et al., 2025).

2.2. Adversarial Attacks on Multimodal Large Language Models (MLLMs)

Recently, several works have reported that MLLMs are vulnerable to adversarial attacks (Qi et al., 2024; Carlini et al., 2023; Gong et al., 2025). These attacks generally fall into two main categories, **jailbreak** and **evasion** attacks (Qi et al., 2024; Cui et al., 2023). **Jailbreak attacks** are primarily designed to break the safety alignment of MLLMs. On the other hand, **evasion attacks** aim to manipulate visual, textual, or both inputs to alter the model's normal behaviour. Evasion attacks are typically designed to introduce targeted mispredictions or produce untargeted misleading outputs while keeping perturbations small (Cui et al., 2023; Zhao et al., 2023; Hanif et al., 2025). A few studies have recently examined the adversarial robustness of MLLMs for medical VQA tasks. (Clusmann et al., 2025; Hanif et al., 2025). However, the existing studies attack the MLLMs by assuming a generic target, *e.g., no signs of disease* (Hanif et al., 2025). In contrast, radiology chest CT reports follow a strict organ-based structure (Hamamci et al., 2024). Each organ-specific section describes the associated abnormalities that are organ-specific. For example, lung findings and abdominal findings involve distinct pathologies and therefore, rely on terminology unique to their respective anatomies. Consequently, a single generic target is insufficient for attacking radiology report generation models. Effective adversarial attacks must therefore account for organ-specific anatomy to produce manipulated outputs that are plausible and stealthy.

3. Proposed Methodology

3.1. Problem Formulation

Consider a 3D chest CT scan x_{ct} and a text prompt x_{t} that instructs the MLLM to generate a text report. At test time, given $(x_{\text{ct}}, x_{\text{t}})$, the model autoregressively predicts the conditional probability of a report $y = (y_1, \ldots, y_K)$, as expressed in Equation (1).

$$p_\theta(y \mid x_{ct}, x_t) = \prod_{k=1}^{K} p_\theta(y_k \mid x_{ct}, x_t, y_{<k}), \tag{1}$$

where p_θ denotes the CT-grounded radiology report generation model parameterized by θ, y_k is the k-th token in the report, $y_{<k} = (y_1, \ldots, y_{k-1})$ denotes all previously generated tokens, and K is the length of the report. Then, the task of adversarial chest CT report generation y_{adv} can be formulated as below:

$$p_\theta(y_{\text{adv}} \,|\, x_{\text{ct}} + \delta_{\text{ct}},\, x_{\text{t}} + \delta_{\text{t}}) = \prod_{k=1}^{K} p_\theta(y_{\text{adv},k} \,|\, x_{\text{ct}} + \delta_{\text{ct}},\, x_{\text{t}} + \delta_{\text{t}},\, y_{\text{adv},<k})\,. \tag{2}$$

In Equation (2), δ_{ct} represent the learnable visual perturbation applied to the 3D chest CT volume and is bounded by perturbation magnitude $||\delta_{\text{ct}}||$. The goal of the adversary is to keep the perturbation magnitude minimum ($||\delta_{\text{ct}}|| \leq \epsilon_{\text{ct}}$) to impose imperceptibility between clean image x_{ct} and perturbed image $x_{\text{ct}} + \delta_{\text{ct}}$. Likewise, δ_{t} is the prompt perturbation bounded by ϵ_{t} ($||\delta_{\text{t}}|| \leq \epsilon_{\text{t}}$).

To examine whether effective adversarial perturbations can be learned for multimodal inputs that mislead the MLLM into altering high-risk chest findings and subsequently generating a clinically dangerous report, we define a target text that the attacker aims to force the model to generate. This type of attack falls under the category of **targeted adversarial attack**, in which the adversary specifies in advance the exact output (typically known as **target**) they aim to obtain from the victim model. In targeted adversarial attack scenarios, the target is typically drawn from the model's normal output distribution but is deliberately selected because it represents an incorrect or dangerous outcome (Sato et al., 2020). Following this principle, we define clinically plausible target outputs that intentionally contradict the ground truth. These opposing targets — such as reporting an abnormality as absent or fabricating a nonexistent one — are chosen to induce harmful clinical decisions.

3.2. Threat Model

We assume a white-box threat model in which the adversary has complete knowledge about the victim CT-to-text generative model p_θ. Under this setting, we focus on learning perturbations for targeted adversarial manipulation of the generated text report.

3.2.1. Clinically-Risky Adversarial Report Generation

In our proposed attack, the adversary aims to deceive the MLLM into producing a predefined harmful target report. We sample the targets from the original reports to ensure that $y_{\text{target}} \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the report sentences extracted from the dataset. Thus, the goal of our targeted attack is to drive the MLLM toward generating the predefined target report by minimizing the text modeling loss $\mathcal{L}$:

$$\min_{\delta_{\text{ct}},\,\delta_t} \; \mathcal{L}(y_{\text{target}},\; p_\theta(y_{adv} \mid x_{\text{ct}} + \delta_{\text{ct}},\; x_t + \delta_t)) \quad \text{s.t.} \quad \|\delta_{\text{ct}}\| \le \epsilon_{\text{ct}},\; \|\delta_t\| \le \epsilon_t. \tag{3}$$

Each chest CT report describes the thoracic organs and indicates whether abnormalities are present. From these descriptions, we derive candidate negative $\mathcal{Y}_{\text{neg}}$ and positive targets $\mathcal{Y}_{\text{pos}}$. For instance, a negative lung target might be: *No nodule or infiltrative lesion is observed in the lung parenchyma.* Driven by the possible ways in which a chest CT report can be adversarially manipulated, we define the following adversarial goals:

(a) Suppressing high-risk chest abnormalities: Here, the adversary aims to increase the likelihood of clinical misdiagnosis by omitting specific targeted abnormalities. We sample the target text from the set of negative sentences in the ground-truth reports, which indicate the absence of any critical abnormality ($y_{\text{target}} \in \mathcal{Y}_{\text{neg}}$).

(b) Fabricating high-risk chest abnormalities: Here, the adversary aims to potentially trigger unnecessary biopsies or follow-up examinations by inserting particular high-risk abnormalities. In this case, we sample the target text from the set of positive sentences in the ground-truth reports, which indicate the presence of critical abnormality ($y_{\text{target}} \in \mathcal{Y}_{\text{pos}}$).

3.2.2. Multimodal Adversarial Optimization

To fool the MLLM to generate an adversarial text report based on multimodal input, we begin by randomly initializing learnable visual perturbations $\delta_{\text{ct}} \in \mathbb{R}^{C\times H\times W\times D}$ and text

prompt perturbations $\delta_t \in \mathbb{R}^{L\times d}$. C, H, W, and D represents channel, height, width and depth dimension of 3D CT volume. L and d denote the length and hidden dimension of the text embedding. δ_{ct} is added to the clean 3D CT scan $x_{ct} \in \mathbb{R}^{C\times H\times W\times D}$ and δ_t to the text-prompt embedding $E_t \in \mathbb{R}^{L\times d}$ derived from text prompt x_t. We also localize the perturbation to the specific anatomical region using segmentation masks to isolate that region within the CT volume. This enables our proposed adversary to steer the adversarial optimization toward the loss associated with the corresponding anatomical subsection of the report. More specifically, let $M^{(r)} \in \{0,1\}^{H\times W\times D}$ be the binary segmentation mask for anatomical region r (*e.g.,* lung), and let δ_{ct} be the global image perturbation. The perturbation applied to region r is $\delta_{\text{ct}}^{(r)} = M^{(r)} \odot (\delta_{\text{ct}} + x_{ct})$. After the perturbed multimodal input is passed through the MLLM, we can then obtain the text report loss $\mathcal{L}$ and subsequently the perturbation's gradient corresponding to the image and text embedding $\nabla_{\delta_{ct}}$ and ∇_{δ_t}, respectively. The image- and text-prompt perturbation gradients are updated via gradient descent, followed by projection onto the allowable perturbation range.

We update perturbations $\delta_{\text{ct}} \leftarrow \text{clip}(\delta_{\text{ct}} - \alpha_{ct}\ \text{sign}(\nabla_{\delta_{ct}}))$ and $\delta_{\text{t}} \leftarrow \text{clip}(\delta_{\text{t}} - \alpha_t\ \text{sign}(\nabla_{\delta_t}))$ using PGD (Madry et al., 2017). α_{ct} and α_t denote the step sizes for image and text embedding perturbation updates. To keep perturbations minimal while adversarially effective, we use PGD with adaptive early stopping (Li et al., 2025) where optimization stops once the generated report is sufficiently aligned with the target ($\geq \tau$ similarity by a text similarity metric) (Zhang et al., 2020). In our experiments, we use 1/255, 0.01, 16/255, 0.1, and 0.85 for α_{ct}, α_t, ϵ_{ct}, ϵ_t, and τ, respectively. We set 100 as the maximum number of optimization steps.

4. Experimental Settings

4.1. Victim CT-to-text Report Generation Model

We use the state-of-the-art Reg2RG model (Chen et al., 2025b) as our victim CT-to-text report generation system. Reg2RG is a multimodal model composed of a 3D Vision Transformer (3D-ViT) (Wu et al., 2025) for volumetric CT encoding and a LLaMA2-7B decoder (Touvron et al., 2023). It incorporates anatomically grounded region tokens to improve alignment between visual features and organ-specific descriptions, enabling structured chest CT report generation. All experiments use the official pretrained weights and inference pipeline released by the authors.[1]

4.2. Implementation and Benchmark Details

We evaluate our attack on the RadGenome-Chest dataset (Zhang et al., 2025b), using the 1,500 CT–report pairs from its validation split. All experiments are conducted on a single NVIDIA RTX A6000 (48 GB). As shown in Figure 2, both the volume encoder and language decoder remain frozen during attack optimization. For demonstration, we focus on adversarial manipulation of critical findings in the lung and breast.

1. https://github.com/zhi-xuan-chen/Reg2RG

Table 1: Evaluation of adversarial attack on targeted organ and entire report. We evaluate our attack for both Targeted Suppression (Omission) and Targeted Fabrication (Insertion) of clinical findings. A higher ROUGE-L score indicates stronger adversarial effectiveness.

Method	Organ-level		Report-level	
	Suppression	Fabrication	Suppression	Fabrication
Baseline (Chen et al., 2025b) (w/o attack)	0.439	0.198	0.370	0.280
CRA-RG (visual)	0.922	0.887	**0.456**	0.272
CRA-RG (text)	0.518	0.719	0.346	0.278
CRA-RG (visual + text)	**0.937**	**0.888**	0.410	**0.294**

4.3. Evaluation Setup for Adversarial CT Reports

We evaluate two targeted attack scenarios to determine whether the adversary can fool the MLLM to omit true findings and fabricate false ones. For each input CT scan, a target is randomly selected from a set of predefined clinical statements. Although the attack focuses on specific anatomies (the lung and breast), perturbations are applied to the entire CT volume and may affect other findings. To analyze these effects, we evaluate adversarial outcomes at both the organ and report levels. Organ-level evaluation measures how closely the generated organ-specific captions match the predefined targets. Report-level evaluation compares the full adversarial report to two idealized adversarial versions — one with all critical findings suppressed and one with all findings fabricated. For reference, we also compute the corresponding similarities for benign predictions to establish a baseline indicating how unlikely such manipulations are in the absence of an attack.

5. Experimental Results

5.1. Attack Success Rate of Targeted CT Report Manipulation

To assess whether the adversarial perturbations applied to multimodal input can lead to successful fooling of the MLLM in manipulating the critical findings, we compute ROUGE-L (Lin, 2004) similarity between the adversarially generated reports and predefined target texts as summarized in Table 1. We compare the effectiveness of visual-only, text-only, and multimodal perturbations applied to the input. Visual perturbations achieve high anatomical organ-level success (0.92), while text-only perturbations are less effective, likely because the image remains unchanged. The strongest manipulation (0.937) occurs when both modalities are perturbed. Notably, fabrication is more difficult than suppression, suggesting that degrading visual evidence makes it harder for the model to justify the insertion of nonexistent findings. We also observe that steering an entire report toward a single fixed target is challenging due to report length and complexity, and fabricating multiple findings becomes especially difficult as noise distorts the image features needed to support such insertions. Qualitative examples of ground-truth and adversarial CT reports are provided in Figure 4 and Figure 5 in Appendix A.

Table 2: NLG-based Evaluation of Stealthiness of Adversarial Reports.

Method	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
Baseline (w/o attack)	0.473	0.365	0.296	0.249	0.441	0.367
CRA-RG (visual + text)	**0.413**	**0.302**	**0.229**	**0.179**	**0.387**	**0.306**

Table 3: Organ Recognition Performance Under Multimodal Adversarial Attack: We examine the ability of MLLM to identify the organs in adversarial CT scans. A decrease in Recall and F1-score indicates the strength of our attack in deteriorating the MLLM's organ detection ability.

Organ	Benign Reports		Adversarial Reports		Difference	
	Recall	F1	Recall	F1	Δ Recall	Δ F1
Abdomen	0.997	0.997	0.649	0.733	0.348	0.264
Bone	0.999	0.999	0.652	0.785	0.347	0.214
Breast	0.967	0.967	0.908	0.863	0.059	0.104
Esophagus	0.999	0.999	0.941	0.965	0.058	0.034
Heart	0.995	0.995	0.938	0.947	0.057	0.048
Lung	0.807	0.807	0.008	0.009	**0.799**	**0.798**
Mediastinum	0.995	0.993	0.946	0.850	0.049	0.143
Pleura	0.807	0.807	0.193	0.319	0.614	0.488
Thyroid	0.962	0.975	0.894	0.785	0.068	0.190
Trachea & Bronchi	0.973	0.977	0.930	0.937	0.043	0.040
Average	0.950	0.952	0.706	0.719	0.244	0.233

5.2. Stealthiness of Adversarial CT Reports

To assess whether adversarial CT reports generated by attacking the MLLM remain stealthy and inconspicuous, we quantify the extent to which they preserve the structural characteristics of reports produced from clean inputs. Specifically, we compute Natural Language Generation (NLG) metrics between adversarial and ground-truth reports and compare these scores with those obtained by evaluating benign reports[2] against their corresponding ground-truth texts. In this setting, NLG metrics serve as proxies for structural similarity because they indicate how closely the adversarial reports retain the overall textual organization of the original reports. We compute BLEU-n (Papineni et al., 2002) (n-gram overlap), ROUGE-L (Lin, 2004) (longest-common-subsequence similarity), and METEOR (Lavie and Agarwal, 2007) (semantically informed matching), as summarized in Table 2. As shown in the table, adversarial reports exhibit modest reductions in NLG scores relative to benign baselines; however, the decline is not substantial given that these reports are generated from adversarially perturbed CT scans and prompt embeddings. This suggests that, despite successfully suppressing critical clinical findings, the adversarial CT reports largely preserve the structural form of the original CT reports and therefore remain stealthy.

5.3. Thoracic Organs Recognition after Adversarial Attack on MLLM

The baseline CT report generation MLLM we attack can both identify thoracic organs and generate full CT reports (Chen et al., 2025b). Through prompting, the model is instructed to both recognize anatomical structures and produce a comprehensive description of the scan. To evaluate how our adversarial perturbations affect the MLLM's ability to detect

2. Benign CT reports denote reports generated from clean, unattacked inputs.

Table 4: Effect of Adversarial Attack on Detection of Chest Abnormalities: This table reports the degradation in detection of chest findings from CT reports under our proposed adversarial threat. The decreases in Recall and F1 indicate the effectiveness of our attack at concealing critical abnormalities in the generated reports.

CT findings	Benign Reports		Adversarial Reports		Difference	
	Recall	F1	Recall	F1	Δ Recall	Δ F1
Arterial wall calcification	0.62	0.62	0.31	0.43	0.31	0.19
Cardiomegaly	0.12	0.16	0.0	0.0	0.12	0.16
Coronary artery wall calcification	0.61	0.55	0.18	0.27	0.43	0.28
Emphysema	0.22	0.26	0.01	0.03	0.21	0.23
Atelectasis	0.65	0.39	0.0	0.0	0.65	0.39
Lung opacity	0.24	0.38	0.03	0.06	0.21	0.32
Pulmonary fibrotic sequela	0.11	0.17	0.0	0.0	0.11	0.17
Pleural effusion	0.38	0.50	0.03	0.05	0.35	0.45
Consolidation	0.24	0.31	0.0	0.0	0.24	0.31
Average	0.35	0.37	0.06	0.09	0.29	0.28

thoracic organs, we assess organ-recognition performance under clean and adversarial conditions using Recall and F1 scores. The organ recognition performance is shown in Table 3. The baseline model recognizes most organs with near-perfect Recall and F1 scores. With adversarially manipulated multimodal inputs, recognition performance decreases across nearly all organs, indicating that adversarial perturbations disrupt the model's grounding ability. The most severe degradation occurs for the lung, where Recall and F1 drop sharply from 0.807 to nearly zero, consistent with our attack targeting lung-related findings. In contrast, regions less directly tied to the adversarial objective (*e.g.,* Heart, Esophagus, Trachea & Bronchi) retain relatively high performance, though still lower than the benign baseline. Overall, these results indicate that adversarial perturbations not only manipulate targeted clinical findings but also impair the MLLM's broader ability to correctly identify anatomical thoracic regions.

5.4. Impact of Adversarial Attack on Detection of Clinical Findings

To evaluate how our attack affects the detection of clinically meaningful findings that are extracted from generated chest CT reports, we employ Clinical Efficacy (CE) metrics (Chen et al., 2020b). In contrast to conventional NLG metrics that assess textual similarity, CE metrics measure diagnostic fidelity by determining whether key clinical abnormalities are correctly captured in the generated text. Because our adversary aims to attack the MLLM to alter critical diagnostic content, CE metrics provide a direct assessment of the attack's clinical consequences. We use the RadBERT text classifier (Yan et al., 2022) to automatically extract abnormalities from reports, and quantify performance using Recall and F1 score. Table 4 summarizes the results. For each abnormality, we present Recall and F1 scores derived from benign and adversarial reports, measured against the abnormalities mentioned in the ground-truth report. We found that, in adversarial reports, the detection of critical chest findings declines sharply. Many abnormalities, including cardiomegaly, atelectasis, consolidation, and fibrotic sequela, become completely undetectable, indicating

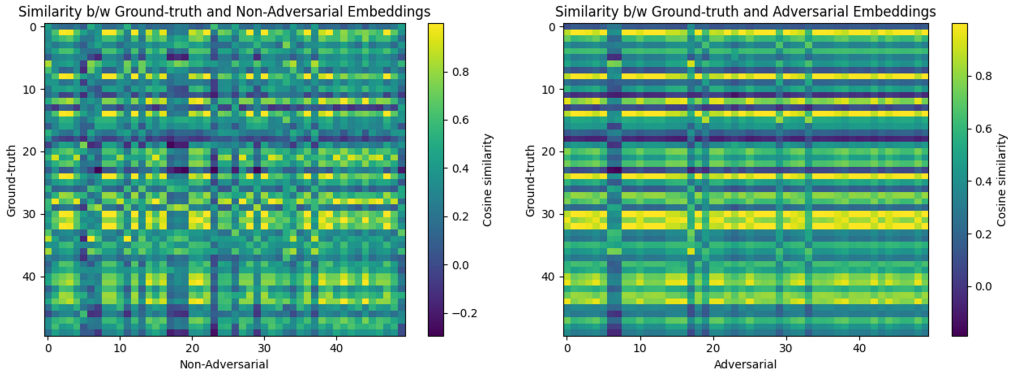

Figure 3: Cosine similarity matrices between ground-truth token embeddings and embeddings from non-adversarial (left) and adversarial (right) reports.

that the adversarial attack substantially disrupts the clinical information encoded in the generated reports and thus reduces their diagnostic reliability.

5.5. Analysis of Multimodal Embedding Space under Adversarial Attack

To understand how multimodal adversarial inputs affect the MLLM's embedding space, we compute cosine similarity between ground-truth embeddings and embeddings from non-adversarial versus adversarial reports. Figure 3 shows the cosine similarity matrices between ground-truth embeddings and embeddings from non-adversarial (left) versus adversarial reports (right). The non-adversarial similarity matrix exhibits broad consistency with the ground-truth embeddings. In contrast, the adversarial matrix exhibits pronounced horizontal bands, indicating that ground-truth tokens map strongly to a narrow subset of adversarial embeddings. This pattern reflects a distortion in the embedding space induced by the adversarial perturbations, demonstrating that the attack not only alters the textual output but also reshapes internal semantic representations.

6. Conclusion

In this work, we presented Clinically Risky Adversarial Report Generation (CRA-RG), a new threat model that characterizes how chest CT radiology report generation systems can be adversarially manipulated to produce clinically dangerous outputs. Our threat model, CRA-RG focuses on clinically meaningful manipulations, including the omission and fabrication of critical findings such as nodules, consolidations, and pleural abnormalities. To instantiate this threat model, we introduced a multimodal targeted adversarial attack that jointly perturbs CT volumes and conditioning text embeddings, enabling fine-grained control over specific anatomical regions. Our experiments on the RadGenome 3D chest CT dataset demonstrated that state-of-the-art multimodal report generation models are highly susceptible to adversarial perturbations. These results provide the first empirical evidence that modern chest CT report generation systems can be driven to produce harmful clinical recommendations — including missing high-risk findings or fabricating nonexistent abnormalities — raising critical safety concerns for real-world deployment.

References

Nayyer Aafaq, Naveed Akhtar, Wei Liu, Mubarak Shah, and Ajmal Mian. Controlled caption generation for images through adversarial attacks. Eprint arXiv:2107.03050, 2021.

Josh Achiam et al. Gpt-4 technical report. Eprint arXiv:2303.08774, 2023.

Jean-Baptiste Alayrac eta l. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. In *International Conference on Learning Representations*, 2025.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In *Advances in Neural Information Processing Systems*, 2023.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020a.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020b.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for ct report generation. *IEEE Transactions on Medical Imaging*, 2025a.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. Large language model with region-guided referring and grounding for ct report generation. *IEEE Transactions on Medical Imaging*, 2025b.

Jan Clusmann, Dyke Ferber, Isabella C. Wiest, Carolin V. Schneider, Titus J. Brinker, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. Prompt injection attacks on vision language models in oncology. *Nature Communications*, 2025.

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

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? Eprint arXiv:2309.11751, 2023.

Report Everlight, Radiology. Everlight: Radiology unlocked: The global radiologist report 2025. Eprint Everlight Radiology, 2025.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. *CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging*, pages 476–486. Springer Nature Switzerland, 2024.

Ibrahim Ethem Hamamci et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. Eprint arXiv:2403.17834v4, 2024.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K. Bressem, Jakob Nikolas Kather, and Daniel Truhn. Medical large language models are susceptible to targeted misinformation attacks. *npj Digital Medicine*, 7(1), 2024.

Asif Hanif, Zaigham Zaheer, Salman Khan, Fahad Shahbaz Khan, and Rao Anwer. *SPARTA: Spectral Prompt Agnostic Adversarial Attack on Medical Vision-Language Models*. Springer Nature Switzerland, 2025.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264, 2024.

Pu Jian, Donglei Yu, and Jiajun Zhang. Large language models know what is key visual entity: An LLM-assisted multimodal retrieval for VQA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Albert Q. Jiang et al. Mistral 7b. Eprint arXiv:2310.06825, 2023.

Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007.

Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, pages 1–13, 2025.

Jin Li, Zitong Yu, Ziqiang He, Z. Jane Wang, and Xiangui Kang. Pgd-imp: Rethinking and unleashing potential of classic pgd with dual strategies for imperceptible adversarial attacks. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023a.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 3334–3343, 2023b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2017.

Khoi Nguyen Tiet Nguyen, Wenyu Zhang, Kangkang Lu, Yuhuan Wu, Xingjian Zheng, Hui Li Tan, and Liangli Zhen. A survey and evaluation of adversarial attacks for object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jia, Xue Lin, and Qi Alfred Chen. Dirty road can attack: Security of deep learning based automated lane centering under physical-world attack. In *USENIX Security Symposium*, 2020.

Erfan Shayegani, Yue Dong, and Nael B. Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International Conference on Learning Representations*, 2024.

Bannur Shruthi et al. Maira-2: Grounded radiology report generation. *arXiv*, abs/2406.04449, 2024. URL https://arxiv.org/abs/2406.04449.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. Eprint arXiv:2302.13971, 2023.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. Eprint arXiv:2307.09288, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2damp;3d medical data. *Nature Communications*, 2025.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 2022.

Baochang Zhang, Chen Jia, Shuting Liu, Heribert Schunkert, and Nassir Navab. Semantic-aware chest x-ray report generation with domain-specific lexicon and diversity-controlled retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 607–616, 2025a.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, 2020.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Weiwei Tian, Ya Zhang, Weidi Xie, and Yanfeng Wang. Development of a large-scale grounded vision language dataset for chest ct analysis. *Scientific Data*, 2025b.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. Eprint arXiv:2307.15043, 2023.

Appendix A. Example of Real and Adversarial CT Report

Figure 4: We present an example of a real CT report (left) and an adversarially generated CT report (right) using our proposed multimodal perturbations. Note that the CT report generation model has been successfully fooled to omit the majority of the significant clinical findings in Chest CT (e.g., Consolidation in lungs and lymph nodes in Mediastinum are successfully suppressed in Adversarial CT Report).

The region 0 is abdomen:	The region 0 is abdomen:
The region 1 is bone: Diffuse heterogeneous density increases were observed in the bone structures in the study area. It is recommended to evaluate it together with clinical and laboratory data in terms of possible metabolic bone diseases.	The region 1 is bone:
The region 2 is breast:	The region 2 is breast: There is a 1 cm diameter nodular lesion in the upper outer quadrant of the left breast (cyst?)
The region 3 is esophagus: Thoracic esophagus calibration was normal and no significant pathological wall thickness increase was detected in the non-contrast examination margins. Sliding type hiatal hernia was observed.	The region 3 is esophagus: Thoracic esophagus calibration was normal and no significant pathological wall thickening was detected.
The region 4 is heart: Pericardial effusion - no thickening was detected. Mediastinal main vascular structures, heart contour, size are natural.	The region 4 is heart: Pericardial effusion-thickening was not observed. As far as can be seen; Mediastinal main vascular structures, heart contour, size are normal.
The region 5 is lung: In the right lung upper lobe posterior segment, there are several parenchymal nodules, one of which is calcified, with irregular borders, the largest one measuring 6.1x5.3 mm. No mass-infiltration was detected in both lung parenchyma. A calcified nonspecific parenchymal nodule with a diameter of 4 mm was observed in the lower lobe of the left lung. Widely ramdomized centriacinar ground glass density increases were observed in both lungs, prominent in the upper lobes (secondary to asthma? secondary to tobacco use?). Pleuroparenchymal sequelae density increases in both lungs apical and emphysematous changes in both lungs were observed	The region 5 is lung: When examined in the lung parenchyma window; Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma
The region 6 is mediastinum: Calcified atherosclerotic changes were observed in the wall of the abdominal aorta. Mediastinal main vascular structures, heart contour, size are natural. Calcification was observed in the aortic valve. The ascending aorta measures 37 mm in diameter and shows slight dilatation.	The region 6 is mediastinum: No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. As far as can be seen; Mediastinal main vascular structures, heart contour, size are normal. The mediastinum could not be evaluated optimally in the non-contrast examination.
The region 7 is pleura: Bilateral pleural thickening-effusion was not detected.	The region 7 is pleura: Pleural effusion-thickening was not detected.
The region 8 is thyroid:	The region 8 is thyroid: It is recommended to be evaluated together with US. The dimensions of the thyroid gland have increased and there are hypodense nodular lesions in the left thyroid lobe.
The region 9 is trachea and bronchie: Trachea, lumen of both main bronchi are open. No obstructive pathology was detected in the lumen of the trachea and both main bronchi. Bilateral peribronchial thickenings were observed.	The region 9 is trachea and bronchie: Trachea and both main bronchi were in the midline and no obstructive pathology was observed in the lumen.

Figure 5: Another Example of a clean (left) and adversarial CT report (right) generated by attacking the MLLM through adversarial multimodal input. The abnormalities in the bone and lung have been successfully omitted.