Medical Multimodal Model Stealing Attacks via Adversarial Domain Alignment

Yaling Shen[1,2,3*], **Zhixiong Zhuang**[1,4*†], **Kun Yuan**[2,3,5], **Maria-Irina Nicolae**[1], **Nassir Navab**[2], **Nicolas Padoy**[5,6], **Mario Fritz**[7]

[1]Bosch Center for Artificial Intelligence, Germany
[2]Technical University of Munich, Germany
[3]Munich Center for Machine Learning, Germany
[4]Saarland University, Germany
[5]University of Strasbourg, France
[6]IHU Strasbourg, France
[7]CISPA Helmholtz Center for Information Security, Germany
{zhixiong.zhuang, irina.nicolae}@bosch.com, {yaling.shen, kun.yuan, nassir.navab}@tum.de,
npadoy@unistra.fr, fritz@cispa.de

Abstract

Medical multimodal large language models (MLLMs) are becoming an instrumental part of healthcare systems, assisting medical personnel with decision making and results analysis. Models for radiology report generation are able to interpret medical imagery, thus reducing the workload of radiologists. As medical data is scarce and protected by privacy regulations, medical MLLMs represent valuable intellectual property. However, these assets are potentially vulnerable to model stealing, where attackers aim to replicate their functionality via black-box access. So far, model stealing for the medical domain has focused on classification; however, existing attacks are not effective against MLLMs. In this paper, we introduce Adversarial Domain Alignment (ADA-STEAL), the first stealing attack against medical MLLMs. ADA-STEAL relies on natural images, which are public and widely available, as opposed to their medical counterparts. We show that data augmentation with adversarial noise is sufficient to overcome the data distribution gap between natural images and the domain-specific distribution of the victim MLLM. Experiments on the IU X-RAY and MIMIC-CXR radiology datasets demonstrate that Adversarial Domain Alignment enables attackers to steal the medical MLLM without any access to medical data.

Introduction

In recent years, the development of medical multimodal large language models (MLLMs) has garnered widespread attention due to their potential to revolutionize healthcare. These models could support clinical decision-making (Seenivasan et al. 2022; Chen et al. 2024a; Yuan et al. 2024b), enhance diagnostic accuracy (Yuan et al. 2024a; Chen et al. 2024b), and promote equitable distribution of medical resources (Jia et al. 2024; Yuan et al. 2023; Zhang et al. 2023). One of the most important use cases is radiology report generation, where the medical MLLM takes a radiology image (e.g., chest X-ray) as input and generates a detailed diagnostic report. Since medical data is usually not publicly available and medical expertise is scarce, a well-performing medical MLLMs becomes valuable intellectual property (IP). However, these models are potentially vulnerable to model stealing attacks (Tramèr et al. 2016; Orekondy, Schiele, and Fritz 2019a), which replicate the functionality of a machine learning model through black-box access. This threat is particularly significant in the medical field because a duplicated model could conflict with IP and privacy regulations, and could also facilitate further attacks. By designing a payload on the copied model, malicious actors can then attack the original model, e.g., with a transfer jailbreak attack (Huang et al. 2024) to compel a medical MLLMs to output fraudulent or fabricated medical information.

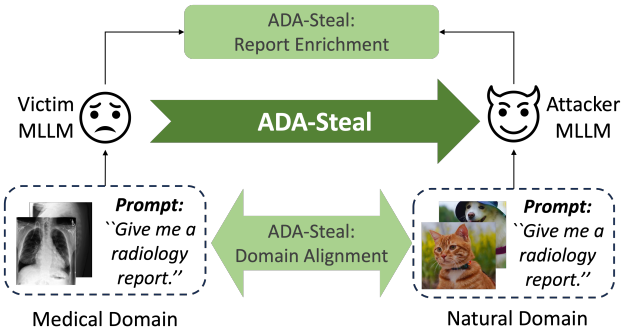


Figure 1: ADA-STEAL trains an attacker MLLM to replicate a victim MLLM for radiology report generation from natural images by first enriching the reports and then aligning the attacker distribution to the medical domain.

Model theft typically involves two steps: (i) creating a transfer dataset by querying the victim model with public or synthetic data (Orekondy, Schiele, and Fritz 2019a; Truong et al. 2021) to obtain pseudo-labels; and (ii) training the attacker model using these pseudo-labels as ground truth. Nevertheless, current model stealing methods present sig-

*These authors contributed equally.
†Corresponding author.

nificant limitations when applied to medical MLLMs. First, due to patient privacy concerns, there are few public medical datasets available for querying, leading to a limited and homogeneous transfer dataset. Second, existing methods primarily target image classification, where every class prediction from the victim model is useful. In contrast, medical text generation involves a much larger output space (i.e., vocabulary), with only a subset of tokens being medically relevant and valuable for training a medical model.

While Knockoff Nets (Orekondy, Schiele, and Fritz 2019a) showed that a diabetic image classification model can be stolen using the non-medical ImageNet dataset, we demonstrate that such images are not suitable for stealing medical MLLMs. When using non-medical images, medical MLLMs produce simplistic, repetitive reports, with few containing relevant disease. This underscores the importance of aligning the query distribution with the victim model's data distribution for successful model stealing, as highlighted in previous works (Orekondy, Schiele, and Fritz 2019a; Truong et al. 2021; Zhuang, Nicolae, and Fritz 2024).

To address these problems, we introduce ADA-STEAL, the first data-free method for stealing medical MLLMs focused on radiology report generation. This method addresses the issues by: (i) using an open-source oracle model to diversify reports without requiring prior medical knowledge, and (ii) integrating the new reports into the query images via targeted adversarial attacks. This process enables non-medical query data to produce more diverse and medically relevant reports, effectively aligning the non-medical domain data with the medical domain data.

Contributions. (i) We are the first to investigate the feasibility of model stealing attacks against medical MLLMs and to identify the associated challenges. (ii) We propose Adversarial Domain Alignment (ADA-STEAL), the first model theft method to replicate the functionality of medical MLLMs without requiring expert knowledge or access to medical domain data. (iii) We validate our ADA-STEAL method on the IU X-RAY and MIMIC-CXR test datasets, showing that it approaches the victim model's performance in both natural language generation metrics and clinical efficacy metrics, even when using the non-medical CIFAR100 dataset. (iv) We conduct ablation studies to analyze the effects of different components of our method, showing that ADA-STEAL can increase the diversity of the attacker dataset by medical report enrichment and domain alignment.

Related Work

Knowledge distillation. Knowledge distillation (Buciluǎ, Caruana, and Niculescu-Mizil 2006; Hinton, Vinyals, and Dean 2015) helps transfer the knowledge from a complex and larger "teacher" model to a compact and simpler "student" model, which is similar to our victim-attacker design. However, unlike knowledge distillation, where the student model has the same data distribution as the teacher model's training data, in our problem formulation, the attacker has no prior knowledge of the victim's black-box model, e.g., unknown architecture, data distributions or training parameters. Although data-free knowledge distillation (Fang et al. 2019; Micaelli and Storkey 2019; Choi et al. 2020) further assumes the absence of the teacher model's training data, its requirement of white-box access to the teacher model for backpropagation is a major difference to our setup.

Model stealing. Model stealing, or model theft, typically has one of two objectives: exact replication of the model or its components, or functionality replication, where the attacker aims to mimic the model's behavior. The first type focuses on extracting the model's hyperparameters (Wang and Gong 2018), architecture (Oh et al. 2018), or learned parameters (Tramèr et al. 2016). The second type (Orekondy, Schiele, and Fritz 2019a; Truong et al. 2021; Zhuang, Nicolae, and Fritz 2024), involves training a model that mimics the victim's performance without prior knowledge of its training data or architecture. In this work, we present the first functionality model stealing attack against MLLMs for radiology report generation. While previous methods like Knockoff Nets (Orekondy, Schiele, and Fritz 2019a) replicate medical image classifiers using natural images, they deal with a much smaller output space compared to text generation. Bert-Thieves (Krishna et al. 2020), another related approach, targets language models, but does not handle images and benefits from publicly available text data that shares a similar distribution with the victim model. In contrast, medical data is hard to obtain, and only a specific subset of the vocabulary is relevant in this context, making it more challenging for the attacker, who may lack prior knowledge in the medical domain.

Security of multimodal large language models. With the ability to understand and reason about different data types, MLLMs are vulnerable to evasion attacks targeting each data modality, such as malicious image and text constructs (Schlarmann and Hein 2023; Liu et al. 2024). In contrast to these attacks that are designed to induce erroneous or disallowed responses, the model stealing attack we present aims to mimic the functionality of the victim medical MLLM for radiology report generation, using only black-box access to the victim model.

Threat Model

In this section, we formalize the threat model for stealing black-box medical MLLMs in radiology report generation. First, we introduce preliminary concepts and notations. We then formalize the victim model as well as the attacker's objective and knowledge based on the real-world setup.

Notations. We model MLLMs predicting the probability of the next token y_l given the preceding language tokens $y_{<l}$, the input image X, and the instruction prompt T. The final output is the answer $Y = \{y_l\}_{l=1}^{L}$. Each token y is drawn from vocabulary $\mathbb{V}$. This is formalized as:

$$p(Y|X,T) = \prod_{l=1}^{L} p(y_l|y_{<l}, X, T), \tag{1}$$

A table summarizing all notations used in this paper is provided in Appendix A.

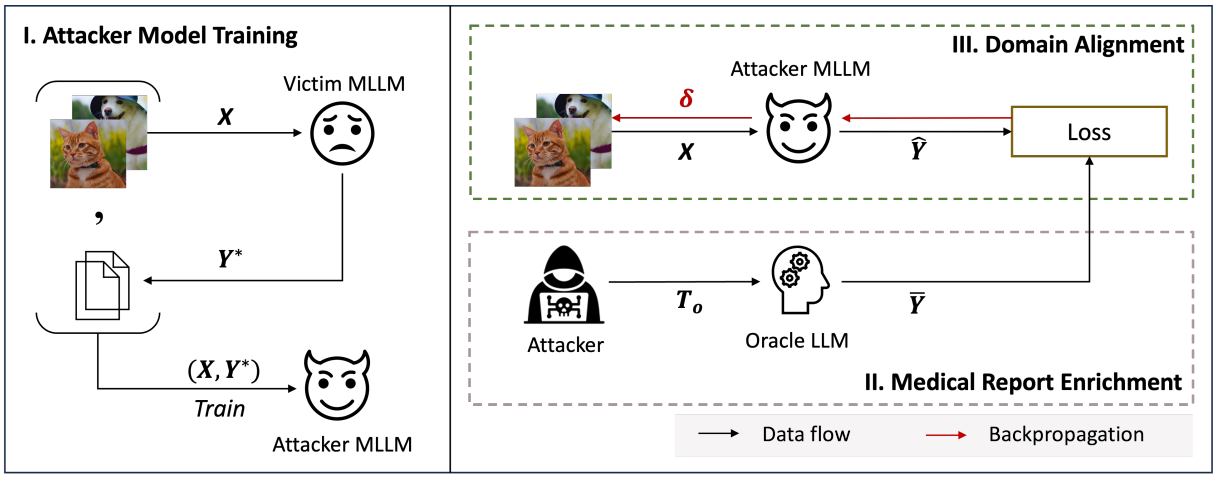


Figure 2: The overview of our proposed approach with three iterative phases: (I) attacker model training, (II) medical report enrichment (in gray dash box), and (III) domain alignment (in green dash box).

Victim model. The victim medical model M_v with parameters θ_v is an MLLM developed for automated medical image interpretation. It accepts the image X and instruction prompt T as inputs, and outputs the answer $Y^* = M_v(X, T; \theta_v)$. We refer to Y^* as the pseudo-report. We consider a deterministic victim making predictions based on beam search. The model can perform various clinical tasks depending on the prompt T, such as image view classification, disease classification, and radiology report generation. We focus primarily on radiology report generation. This task has a much larger output space ($|\mathbb{V}_v|$) than classification tasks, and valid medical reports typically only use a small portion of the entire vocabulary. This characteristic poses additional challenges for model stealing. For simplicity, we fix the prompt T for the radiology report generation task and vary only the input image X. We use $Y^* = M_v(X)$ as a shorthand for victim predictions under fixed T. The original victim training dataset is denoted as $\mathcal{D}_v = \{(X_v, Y_v)\}$, where $X_v \sim P_v$ (i.e., data distribution of the victim training images) and Y_v is its corresponding medical report.

Goal and knowledge of the attacker. The attacker aims to train a surrogate model $M_a(X, T; \theta_a)$, parameterized by θ_a, that is able to generate the radiology report $\hat{Y}$ from the image X similar to those produced by the victim model M_v with instruction prompt T. Here as well we consider T to be fixed for this specific task. The attacker is allowed to query the victim model with any image and receive the corresponding report. However, the attacker lacks knowledge of: (i) the internals of the victim medical MLLM M_v, including its architecture; (ii) the dataset $\mathcal{D}_v$ used to train the victim model; (iii) the vocabulary $\mathbb{V}_v$ of M_v; and (iv) the probability distribution of each token in the victim output Y^*.

Adversarial Domain Alignment

The attacker aims to replicate the radiology report generation functionality of a black-box medical model without access to medical datasets. To achieve this, they would ideally optimize the parameters θ_a of the attacker model M_a to minimize the token prediction loss $\mathcal{L}$ on the victim dataset $\mathcal{D}_v$:

$$\min_{\theta_a} \frac{1}{|\mathcal{D}_v|} \sum_{(X_v, Y_v) \in \mathcal{D}_v} [\mathcal{L}(M_a(X_v), Y_v)] \tag{2}$$

Since the attacker does not have access to $\mathcal{D}_v$, they need to construct an own dataset $\mathcal{D}_a$ for training:

$$\mathcal{D}_a = \{(X_a, Y^*) \mid X_a \sim P_a, Y^* = M_v(X_a)\}, \tag{3}$$

where X_a can be sourced from a publicly available non-medical dataset with distribution P_a, and Y^* is the pseudo-report predicted by the victim model. However, using non-medical images as queries yields homogeneous reports that barely explore the output vocabulary space of the victim model. Moreover, the attacker lacks the prior knowledge to guide the exploration and diversity of radiology reports. To address these challenges, we propose Adversarial Domain Alignment (ADA-STEAL), which initially diversifies the reports and then aligns P_a with P_v through data augmentation based on adversarial attacks. In turn, this enables the attacker to obtain more varied, medically relevant reports from the victim. To this end, the objective of the attacker is:

$$\underset{\theta_a, \delta}{\text{minimize}} \frac{1}{|\mathcal{D}_a|} \sum_{(X_a, Y^*) \in \mathcal{D}_a} [\mathcal{L}(M_a(X_a), Y^*) + \mathcal{L}(M_v(X_a + \delta), \bar{Y})], \tag{4}$$

where δ is the adversarial perturbation on image X_a to elicit a more diverse and medically relevant report $\bar{Y}$ from the victim model, which is later replaced by its proxy, the attacker model. The goal is for the perturbed query data to better approximate the distribution P_v.

The overall method consists of three steps: (I) **attacker model training** to mimic the victim model; (II) **medical report enrichment** to diversify the victim's pseudo-reports; and (III) **domain alignment** to shift the attacker query image distribution towards the medical image distribution. Three steps can be iterated until the query budget B is exhausted. The overview of the pipeline is shown in Figure 2.

Attacker Model Training

Following standard model stealing, the attacker queries the victim model with initial non-medical images from the distribution P_a and receives radiology report outputs. The attacker model M_a is then trained to minimize the loss in Equation (5) on the attacker's dataset $\mathcal{D}_a$:

$$\mathcal{L}(M_a(X_a), Y^*) = -\sum_{l=1}^{L} \log p(y_l^* \mid M_a(X_a)_{<l}) \quad (5)$$

Once trained, this model serves as the proxy for the victim model in step III to design adversarial perturbations to be transferred to the victim. The dataset $\mathcal{D}_a$ will be iteratively updated with aligned data following step III.

Medical Report Enrichment

Repetitive medical pseudo-reports from natural images limit the attacker model's ability to generalize to real medical images, which often feature varied abnormalities. Since the attacker lacks the expertise to design accurate reports, we incorporate an additional open-access large language model (LLM) as our oracle model M_o to generate a more diverse and medical-relevant report $\bar{Y}$. Here, $\bar{Y} \sim M_o(T_o)$, where T_o is the prompt for the oracle model as follows. The report $\bar{Y}$ will be used in the next step as the desired output for the query image. We emphasize that M_o does not benefit from the image modality.

Prompt T_o

> Give me some examples of normal/abnormal descriptions of the airway, breathing, cardiac, diaphragm, and everything else (e.g., mediastinal contours, bones, soft tissues, tubes, valves, and pacemakers) for chest X-rays.

Domain Alignment

Training the attacker model M_a on P_a using the victim model's predictions makes M_a a proxy for the victim model. This suggests that M_a can map X_v to Y_v, even with initial low accuracy. We hypothesize that an input X can be generated to produce a relevant report $\bar{Y}$ by reversing the mapping, $X = M_a^{-1}(\bar{Y})$. We adapt the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2015) to generate adversarial perturbations δ on image X_a with the attacker model M_a and the new report $\bar{Y}$:

$$\delta = \epsilon \cdot \text{sign}(\nabla_{X_a} \mathcal{L}(M_a(X_a), \bar{Y})), \quad (6)$$

where ϵ is the magnitude of the adversarial perturbations. The attacker can query the victim model with the generated image $X_a + \delta$ and update the attacker transfer dataset $\mathcal{D}_a$:

$$\mathcal{D}_a = \{(X_a + \delta, M_v(X_a + \delta))\}. \quad (7)$$

Repeat steps I-III until query budget B is exhausted.

Experimental Setup

We now introduce the setup used in our experiments, including models, datasets, attacker image distribution, baseline attacks, evaluation metrics, and implementation details.

MODEL	NO. PARAMETERS	ROLE
CHEXAGENT	8 bn	Victim & Attacker
IDEFICS	9 bn	Attacker
ZEPHYR	7 bn	Oracle

Table 1: Models used in our experiments.

DATASET	TRAIN/TEST	IMAGE	LABEL
CIFAR-100	50k/10k	Non-medical	Image classes
MIMIC-CXR	369k/5k	Chest X-ray	Radiology reports
IU X-RAY	5k/0.8k	Chest X-ray	Radiology reports

Table 2: Overview of datasets.

Models. To verify our proposed method, we use three pre-trained models. CHEXAGENT (Chen et al. 2024b) is a 7 billion-parameter medical foundation model, designed to analyze and summarize CXRs. This MLLM is used as the victim in all our experiments. We denote CHEXAGENT* the version of the same model used by the attacker, with the vanilla large language model (LLM) weights (i.e., MISTRAL-7B) instead of the clinically fine-tuned version. IDEFICS (Laurençon et al. 2024) is a 9 billion-parameter MLLM trained on image-text pairs on various multimodal benchmarks. We directly use this pre-trained MLLM as attack baselines. ZEPHYR-7B (Tunstall et al. 2023) is used as oracle LLM for our ADA-STEAL in all experiments. An overview of model architectures and their role in our work is provided in Table 1.

Datasets. We validate ADA-STEAL on three standard datasets: IU X-RAY (Demner-Fushman et al. 2016), MIMIC-CXR (Johnson et al. 2019), and CIFAR-100 (Krizhevsky 2009). IU X-RAY and MIMIC-CXR are medical datasets consisting of chest X-rays and their corresponding radiology reports. Following previous work (Chen et al. 2020, 2021; Pellegrini et al. 2023), we only consider the Findings section in the radiology reports and exclude samples without it. Furthermore, MIMIC-CXR is one of the training datasets of the victim CHEXAGENT (Chen et al. 2024b). CIFAR-100 is extensively used in the computer vision community, but unrelated to the medical field. Our goal is to show that natural images can be used to steal medical IP, despite originating from a different domain or data distribution. Table 2 summarizes the three datasets with official or conventional training and test split sizes.

The attacker query data P_a. The attacker queries the victim with images from a large discrete image distribution P_a, as shown in Equation (3). The experiments differ in four choices of P_a, each of which is explained below.

- P_a = CIFAR-100: we sample natural images from the training set of CIFAR-100 as the attacker query data.
- P_a = MIMIC-CXR: we assume the attacker has access to the radiographs in the training set of the MIMIC-CXR as the initial query data. This setup serves as an upper bound on the stealing performance the attacker can achieve.

Test Data		MIMIC-CXR			IU X-Ray		
Metrics		RG-L	Bert-S	Rad-S	RG-L	Bert-S	Rad-S
Victim	CheXagent	26.5 (1.00×)	53.0 (1.00×)	20.7 (1.00×)	32.2 (1.00×)	58.7 (1.00×)	26.1 (1.00×)
Attacker	IDEFICS	14.6 (0.55×)	8.0 (0.15×)	0.7 (0.04×)	14.1 (0.44×)	11.2 (0.19×)	0.3 (0.01×)
	+Knockoff	20.2 (0.76×)	45.3 (0.85×)	12.5 (0.60×)	22.9 (0.71×)	48.1 (0.82×)	14.0 (0.54×)
	+ADA-Steal	**23.2** (0.88×)	**49.0** (0.92×)	**15.6** (0.75×)	**29.8** (0.93×)	**52.9** (0.90×)	**19.4** (0.74×)
	CheXagent*	10.5 (0.40×)	0.0 (0.00×)	0.6 (0.03×)	6.6 (0.20×)	0.0 (0.00×)	0.5 (0.02×)
	+Knockoff	23.2 (0.88×)	43.7 (0.82×)	16.3 (0.79×)	26.1 (0.81×)	48.3 (0.82×)	22.0 (0.84×)
	+ADA-Steal	**24.7** (0.93×)	**44.5** (0.84×)	**18.7** (0.90×)	**26.6** (0.83×)	**52.5** (0.89×)	**25.8** (0.99×)

Table 3: Performance metrics on test data (best value in **bold**, followed by the ratio to the original victim performance).

- $P_a \sim \mathcal{N}(0, 1)$: images are randomly generated with pixel values following a standard normal distribution, scaled to [0, 255], and rounded to the nearest integer.
- P_a = mix of CIFAR-100 and MIMIC-CXR: we include a variable number of images from the victim training set (i.e., images from MIMIC-CXR) into the attacker dataset. The proportion of these images is controlled by the ratio r, defined as the percentage of MIMIC-CXR images within the initial attacker dataset.

Attacks. We first test the original performance of IDEFICS and CheXagent* on the medical dataset as a baseline reference, followed by evaluating two attack strategies. Knockoff fine-tunes the attacker model using the method of Knockoff Nets (Orekondy, Schiele, and Fritz 2019a) with the same MLLM. ADA-Steal denotes our proposed method.

Evaluation metrics. We evaluate our attacker model performance on the test sets of IU X-Ray and MIMIC-CXR. After preprocessing, the number of test samples in MIMIC-CXR and IU X-ray is 3858 and 590, respectively. Following *CheXbench*, a benchmark designed by Chen et al. (2024b) to evaluate models across eight CXR interpretation tasks (e.g., radiology report generation). We evaluate the above models by two types of metrics: the natural language generation metrics include ROUGE-L (RG-L) (Lin 2004) and BERT-Score (Bert-S) (Zhang et al. 2019), while the clinical efficacy metrics include RadGraph-Score (Rad-S) (Jain et al. 2021) and GPT-4 evaluation. In particular, the Rad-S assesses the quality of generated radiology reports by employing output reports to construct RadGraphs to identify entities and their relations in comparison to ground truth references. The performance on the CheXbert metric is not reported due to its unreliability for out-of-distribution reports. Detailed explanations and the prompt for GPT-4 evaluation are included in Appendix C and D, respectively.

Implementation details. We initialize the attacker query set with 500 images from CIFAR-100, and then repeat the steps of our method three times, resulting in a total query budget of $B = 1500$. We set the probabilities of abnormal, normal, and original ($\hat{Y}$) anatomical descriptions into 80%, 10%, and 10%, respectively, in the new report $\bar{Y}$ for adversarial perturbation generation. The learning rates for fine-tuning IDEFICS and CheXagent* are fixed to 5×10^{-6} and 1×10^{-5}, respectively, without weight decay. The maximum new sequence length is set to 512, and a diversity penalty of 0.2 with three beam groups, each containing six beams, is applied. The top-1 response is collected as the generated report. The adversarial noise budget ϵ is set to 0.2 unless otherwise specified. In both the victim and attacker models, the report generation process employs the same prompt T, as shown below. All experiments are conducted on a single NVIDIA A100 GPU.

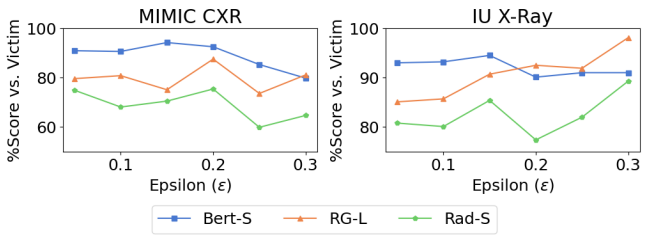
Fixed prompt T for victim and attacker model

Write a structured Findings section for the given image as if you are a radiologist.

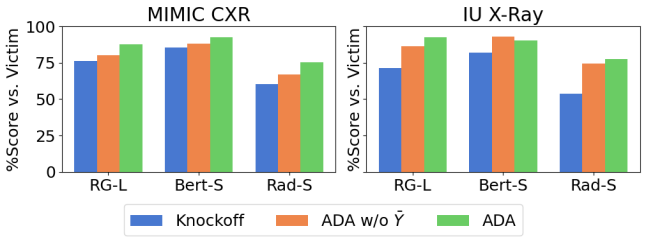
Experimental Results

Model stealing performance. The main experimental results on the two aforementioned test datasets are shown in Table 3. First, both attack strategies improve the performance of radiology report generation compared to the original MLLM. This outcome highlights the vulnerability of the medical MLLMs to model stealing attacks. Second, ADA-Steal outperforms the other attack in all metrics, which confirms the effectiveness of our proposed model stealing method in enhancing the diversity of the attacker set. Such diversity may come from the incorporation of medical report enrichment and adversarial image generation that align attacker predictions with the expert knowledge of the oracle LLM. Third, when comparing between datasets, the performance achievements of ADA-Steal tested on IU X-Ray are higher than that of MIMIC-CXR. This might be because IU X-Ray is relatively small and has less diverse image-text mappings, making it easier for attackers to mimic the image-to-text generation functionality. Finally, there is no clear winner in terms of the model architecture used by the attacker, but rather the performance depends on both the attack strategy and the measured metric.

Ablative analysis. To investigate the effect of adversarial attacks in domain alignment as well as oracle LLM diversification in medical report enrichment, an abrogation study is conducted, examining the impact of the non-utilization of the oracle model in ADA-Steal (denoted as ADA w/o $\bar{Y}$). Under the same experimental setup of IDEFICS and



(a) Stealing performance of the adversarial noise budget ϵ.



(b) Ablation performance of ADA-STEAL without Oracle.

Figure 3: Left: performance of ADA-STEAL on IDEFICS with different ϵ across three metrics compared to the victim. Right: ablation study compares the performance of three attackers.

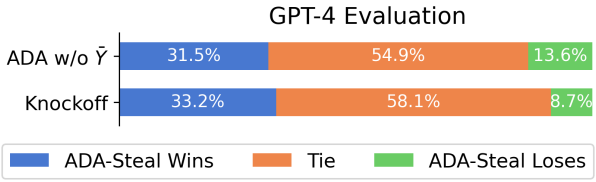


Figure 4: Qualitative evaluation by GPT-4 comparing the quality of test reports pairwise.

$P_a(X)$	MIMIC-CXR			IU X-RAY		
	RG-L	BERT-S	RAD-S	RG-L	BERT-S	RAD-S
$\emptyset$	14.6	8.0	0.7	14.1	11.2	0.3
CIFAR-100	21.4	**48.0**	14.1	**27.6**	**54.6**	**22.3**
$r = 0.1$	**24.1**	47.9	**15.5**	27.4	53.6	21.1
$r = 0.5$	20.4	42.8	10.2	21.0	42.2	9.5
MIMIC-CXR	18.7	38.3	7.9	13.8	30.3	10.4
RANDOM	7.0	0.0	0.0	5.3	0.0	0.0

Table 4: Evaluation of IDEFICS+ADA-STEAL using different attacker image distributions P_a, where the attacker models are fine-tuned with $\epsilon = 0.05$.

$\epsilon = 0.2$, we compare its model performance to KNOCKOFF and ADA-STEAL and show the results in Figure 3b. First, ADA w/o $\bar{Y}$ always outperforms KNOCKOFF, confirming the added value of adversarial images toward domain alignment. While the victim model tends to produce simple, repetitive pseudo-reports with no disease indications, the adversarial images generated during domain alignment enhance the confidence and quality of their pseudo-reports. Second, applying the oracle model to diversify the pseudo-reports can further improve the model stealing ability. This observation is consistent with our hypothesis that the medical report enrichment helps increase the diversity and maintain the clinical relevance of adversarial images by aligning the attacker's predictions with its own knowledge.

Adversarial noise budget ϵ. To analyze the impact of the amount of adversarial noise introduced, we train our attacker model (IDEFICS+ADA-STEAL) for different ϵ in the range from 0.05 to 0.3. Figure 3 shows the results on MIMIC-CXR and IU X-RAY. For all performance metrics, the general trend of the score is to rise and then fall as ϵ increases. This pattern can be attributed to the intrinsic nature of the perturbation. An extremely small ϵ introduces insufficient perturbation to alter the original image significantly. Conversely, the perturbation resulting from a comparatively large ϵ can push pixel values out of the data distribution. While there is not one optimal value for ϵ across datasets and metrics, we note that the range between 0.05 to 0.2 yields overall good performance.

Image distribution P_a. Table 4 illustrates the evaluations of IDEFICS+ADA-STEAL fine-tuned on different attacker dataset distributions. The row $P_a = \emptyset$ shows the radiology report generation ability of the open-access IDEFICS-9b model. The following observations can be made from the results. First, initializing attacker images from the normal distribution $\mathcal{N}(0, 1)$ (row RANDOM) fails to steal the model, resulting in a performance inferior to that of the original IDEFICS. This failure can be explained by the significant gap between the normal distribution and the actual distribution of medical images, which causes the attacker to be inadequate in providing the necessary medical information; in this setup, our approach to enhance data diversity does not help, as random noise is diverse, but is rather lacking domain relevance. Second, the attacker sets constructed with image sampling from the other four image distributions help improve the model's ability to generate radiology reports. The counterintuitive observation that P_a = MIMIC-CXR leads to worse performance than the other three due to the fact that the synthesized adversarial images on the actual CXRs distort the image distribution from the actual CXRs.

Impact of oracle and victim models. We further evaluate the flexibility and generalizability of our method with different oracle and victim models and record the detailed experimental performance in Appendix B. As shown in Figure 6, using GPT-4 as the oracle model, due to its greater capabilities, improves the performance of ADA-STEAL. Additionally, using MED-FLAMINGO (Moor et al. 2023) as the victim model, Table 6 shows that ADA-STEAL outperforms baselines and even surpasses the victim model in RG-L and BERT-S METRICS.

Qualitative analysis. We further investigate the quality of pseudo-reports generated by ADA-STEAL based on augmented images. Figure 5 shows an example of pseudo-

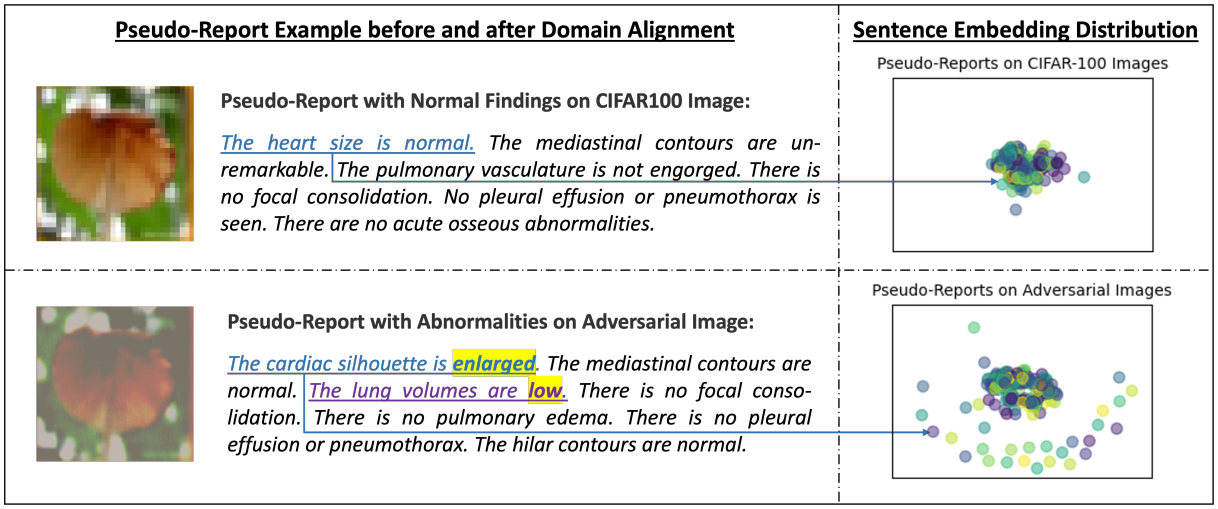


Figure 5: Left: Example of an image and pseudo-report pair from the Mushroom class in CIFAR-100. The abnormality is marked in yellow, with the victim model predicting an enlarged heart and low lung volumes in the adversarial pseudo-report. Right: t-SNE visualizations of pseudo-report sentence embeddings on the CIFAR-100 image and its adversarial counterpart from ADA-STEAL; showing that abnormalities increase report diversity. Arrows indicate the sentence embeddings for these descriptions.

reports generated for a CIFAR-100 image and for its adversarial counterpart used in ADA-STEAL (left). The images in the attacker queries constructed with ADA-STEAL tend to induce more diverse pseudo-reports which include descriptions of abnormalities, while the other attacks fail to do so. This phenomenon is consistent throughout the generated data and aligns with our design objective of the oracle module meant to induce more anomalies. The t-SNE (van der Maaten and Hinton 2008) visualization of the pseudo-report sentence embedding space (Figure 5, right) further confirms the diversity introduced by the use of adversarial images in ADA-STEAL. Finally, we use GPT-4 to assess the quality of the pseudo-reports generated by ADA-STEAL in comparison to those from KNOCKOFF and ADA w/o $\bar{Y}$ applied to IDEFICS. Figure 4 shows that ADA-STEAL generates more high-quality reports than the other attacks.

Discussion

Extended scope. Our current setup applies to victim models that are deterministic and use beam search decoding. We see no fundamental limitation in our methodology that would prevent us from extending our work to models with stochastic outputs or that vary the output decoding strategy.

Reports' drift. The reports generated by the oracle LLM are only based on a prompt and no images. We identify a risk that these reports are not accurate or of good quality. However, we notice in practice that data diversity close to the target domain is sufficient to elicit data diversity in the victim model. Moreover, the reports are not used to train the stolen model, further limiting their potential impact.

Defenses. Our work does not evaluate ADA-STEAL against model stealing defenses. Current defenses (Orekondy, Schiele, and Fritz 2019b; Kariyappa and Qureshi 2020; Mazeika, Li, and Forsyth 2022) typically add noise to the victim model's predictions (i.e., logits) to hinder the attacker while maintaining the model's utility (e.g., top-class predictions). However, these defenses are not suited to the complex input and output spaces addressed in this paper. Additionally, in the medical field, it is crucial that reports generated by medical MLLMs remain truthful and accurate. As such, defending against the present attack is not trivial and would require dedicated defenses.

Ethical considerations. This paper advances research on model stealing attacks, emphasizing the significance of this threat and the urgent need for robust defenses. By raising awareness of the latest security challenges, we aim to help the community better prevent or mitigate such attacks. Our work is built entirely on open resources, and we hope it motivates further study of practical attacks on machine learning to ultimately develop safer, more reliable systems.

Conclusion

In this paper, we show for the first time that an attacker can successfully steal the functionality of radiology report generation from a medical MLLM without access to the victim data distribution. Our attack ADA-STEAL produces a diverse dataset for stealing by leveraging adversarial attacks for domain alignment and an oracle model for report enrichment. Experimental results on two medical datasets demonstrate the effectiveness of ADA-STEAL, which outperforms existing methods directly adapted to the task of radiology report generation. We encourage medical MLLM owners to consider the risk of model theft and to protect their assets.

Acknowledgements

We acknowledge the support and funding by Bosch AIShield. This work was supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. This work was also granted access to the HPC resources of IDRIS under the allocations AD011013704R1, AD011011631R2, and AD011011631R4 made by GENCI.

References

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

Chen, Z.; Luo, X.; Wu, J.; Chan, D.; Lei, Z.; Wang, J.; Ourselin, S.; and Liu, H. 2024a. VS-Assistant: Versatile Surgery Assistant on the Demand of Surgeons. *arXiv preprint arXiv:2405.08272*.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *Association for Computational Linguistics (ACL)*.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Veen, D. V.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; Tsai, E. B.; Johnston, A.; Olsen, C.; Abraham, T. M.; Gatidis, S.; Chaudhari, A. S.; and Langlotz, C. 2024b. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *arXiv preprint arXiv:2401.12208*.

Choi, Y.; Choi, J.; El-Khamy, M.; and Lee, J. 2020. Data-free network quantization with adversarial knowledge distillation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association (JAMIA)*.

Fang, G.; Song, J.; Shen, C.; Wang, X.; Chen, D.; and Song, M. 2019. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations (ICLR)*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (NeurIPS) Deep Learning Workshop*.

Huang, X.; Wang, X.; Zhang, H.; Xi, J.; An, J.; Wang, H.; and Pan, C. 2024. Cross-Modality Jailbreak and Mismatched Attacks on Medical Multimodal Large Language Models. *arXiv preprint arXiv:2405.20775*.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jia, S.; Bit, S.; Searls, E.; Claus, L.; Fan, P.; Jasodan; , V. H.; Lauber, M. V.; Veerapaneni, D.; Wang, W. M.; Au, R.; and Kolachalama, V. B. 2024. MedPodGPT: A multilingual audio-augmented large language model for medical research and education. *medRxiv*.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*.

Kariyappa, S.; and Qureshi, M. K. 2020. Defending against model stealing attacks with adaptive misinformation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Krishna, K.; Tomar, G. S.; Parikh, A. P.; Papernot, N.; and Iyyer, M. 2020. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations (ICLR)*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. *University of Toronto*.

Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; Lozhkov, A.; Wang, T.; Karamcheti, S.; Rush, A.; Kiela, D.; et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Association for Computational Linguistics (ACL)*.

Liu, X.; Zhu, Y.; Lan, Y.; Yang, C.; and Qiao, Y. 2024. Safety of Multimodal Large Language Models on Images and Text. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Mazeika, M.; Li, B.; and Forsyth, D. 2022. How to steer your adversary: Targeted and efficient model stealing defenses with gradient redirection. In *International Conference on Machine Learning (ICML)*.

Micaelli, P.; and Storkey, A. J. 2019. Zero-shot knowledge transfer via adversarial belief matching. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*. PMLR.

Oh, S. J.; Augustin, M.; Schiele, B.; and Fritz, M. 2018. Towards Reverse-Engineering Black-Box Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Orekondy, T.; Schiele, B.; and Fritz, M. 2019a. Knockoff Nets: Stealing Functionality of Black-Box Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Orekondy, T.; Schiele, B.; and Fritz, M. 2019b. Prediction Poisoning: Towards Defenses Against DNN Model Stealing

Attacks. In *International Conference on Learning Representations (ICLR)*.

Pellegrini, C.; Özsoy, E.; Busam, B.; Navab, N.; and Keicher, M. 2023. RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance. *arXiv preprint arXiv:2311.18681*.

Schlarmann, C.; and Hein, M. 2023. On the adversarial robustness of multi-modal foundation models. In *International Conference on Computer Vision (ICCV)*.

Seenivasan, L.; Islam, M.; Krishna, A. K.; and Ren, H. 2022. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security*.

Truong, J.-B.; Maini, P.; Walls, R. J.; and Papernot, N. 2021. Data-Free Model Extraction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research (JMLR)*.

Wang, B.; and Gong, N. Z. 2018. Stealing hyperparameters in machine learning. In *IEEE Symposium on Security and Privacy (IEEE S&P)*.

Yuan, K.; Kattel, M.; Lavanchy, J. L.; Navab, N.; Srivastav, V.; and Padoy, N. 2024a. Advancing surgical VQA with scene graph knowledge. *International Journal of Computer Assisted Radiology and Surgery*.

Yuan, K.; Srivastav, V.; Navab, N.; and Padoy, N. 2024b. Procedure-Aware Surgical Video-language Pretraining with Hierarchical Knowledge Augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yuan, K.; Srivastav, V.; Yu, T.; Lavanchy, J.; Mascagni, P.; Navab, N.; and Padoy, N. 2023. Learning Multi-modal Representations by Watching Hundreds of Surgical Video Lectures. *arXiv preprint arXiv:2307.15220*.

Zhang, K.; Yu, J.; Yan, Z.; Liu, Y.; Adhikarla, E.; Fu, S.; Chen, X.; Chen, C.; Zhou, Y.; Li, X.; et al. 2023. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Zhuang, Z.; Nicolae, M.-I.; and Fritz, M. 2024. Stealthy Imitation: Reward-guided Environment-free Policy Stealing. In *International Conference on Machine Learning (ICML)*.