Detecting Synthetic Radiology Reports using Style Disentanglement

Tanvi Ranga[1] **Arjun Ramesh Kaushik**[1] **Nalini Ratha**[1]
[1]University at Buffalo, Buffalo, USA
{tanviran, kaushik3, nratha}@buffalo.edu

Abstract

Generative AI and large language models (LLMs) have attracted significant attention from both industry and academia for their ability to generate high-quality content across diverse tasks. However, these capabilities raise growing concerns about misuse in sensitive domains such as news, education, software engineering, and medicine. In particular, recent cases suggest that LLMs are being exploited to fabricate radiology reports, potentially enabling insurance fraud. While prior research on detecting machine-generated text has explored domains such as news and scientific writing, there remains a lack of specialization for radiology, leaving a critical gap in reliably identifying AI-generated medical reports. To address this, we introduce text-to-text and image-to-text datasets specifically designed for radiology report generation using multiple LLMs. In addition, we establish a benchmark detection methodology based on disentangling style from content, enabling more effective differentiation between authentic radiology reports and AI-generated fabrications.

1 Introduction

The rapid advancement of generative artificial intelligence, particularly large language models (LLMs), has significantly enhanced the ability to produce persuasive synthetic documents. While these technologies offer tremendous potential for beneficial applications, they are increasingly being misused to fabricate medical records, falsify health claims, and, in some cases, enable complex insurance fraud schemes. Prior research highlights that the proliferation of generative AI and deepfake technologies could amplify fraudulent practices, including the alteration of diagnostic scans and the creation of counterfeit medical records to justify unjustified claims or treatments [30]. Such systems can generate falsified medical documentation with striking accuracy, even replicating patient and physician identities, as demonstrated by platforms such as "Only Fake" [10]. Alarmingly, these methods have already been used to generate highly realistic examples, such as fabricated X-rays illustrating false bone fractures [3].

This growing misuse underscores the urgent need for reliable detection mechanisms to safeguard against the societal and economic risks posed by AI-driven medical forgeries. While prior work has explored the use of LLMs to generate radiology reports [43, 47, 42], the detection of such machine-generated reports remains an open challenge. Current detection methods for LLM-generated text have shown strong performance in domains such as news and scientific writing [49, 24, 23], but they are not optimized for the medical domain or radiology reports owing to the *lack of datasets*.

To this end, we construct a new dataset of synthetically generated radiology reports using large language models (LLMs), with ground-truth references derived from the IU-Xray dataset [28]. The dataset is designed in two variants: (1) Text-to-Text (T2T), where ground-truth reports are

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

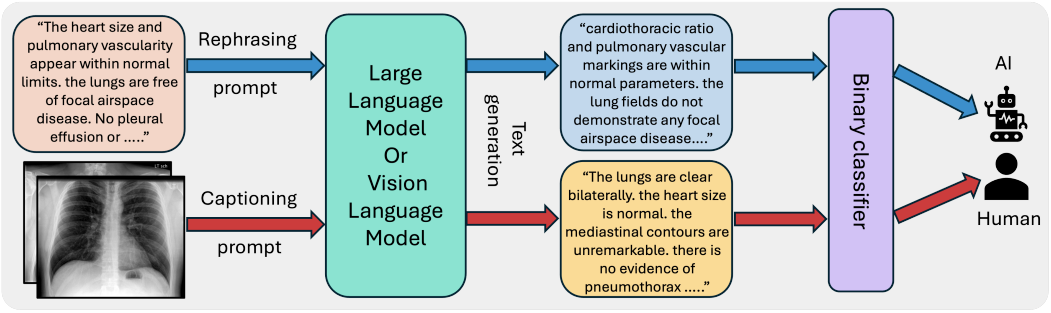


Figure 1: **Overview of our contributions.** Firstly, we construct our dataset via two complementary strategies: (i) prompting large language models (LLMs) with radiologist-written reports to generate paraphrased versions, and (ii) generating the complete reports directly from radiographic images. Subsequently, we utilize the resulting dataset, consisting of both synthetic and human-authored reports, to train a classifier model discerning human-authored reports from synthetic ones.

paraphrased or reframed by LLMs, and (2) Image-to-Text (I2T), where reports are generated directly from radiology images. For the T2T variant, we employ four general-purpose LLMs, whereas the I2T variant is constructed using two medical-domain VLMs pre-trained on radiology data.

To benchmark this dataset, we introduce a novel detection pipeline that disentangles stylistic and semantic attributes of text. Despite the increasing sophistication of modern Large Language Models (LLMs) towards human-like writing, their outputs still exhibit latent stylistic fingerprints that can be systematically detected [38]. These fingerprints, manifested through lexical preferences, syntactic regularities, and punctuation or grammatical standardization, differ measurably from human-authored writing. Our detection pipeline uses a BERT–Mamba [9, 8] backbone to learn separable style and content representations from radiology reports. Unlike prior architectures using BiLSTM [2] or transformers [39], Mamba incorporates sparse attention to handle longer sequences more efficiently. With this integration, our authorship detection pipeline achieved very high MCC scores ranging from 92% to 100% in both text-to-text (T2T) and image-to-text (I2T) category.

In summary, the primary contributions of this work are as follows: (1) We introduce a novel synthetic radiology report dataset that integrates both textual and imaging modalities as inputs. To the best of our knowledge, this represents the first dataset of its kind in the domain, and (2) We establish a benchmark for this dataset using a Mamba-based encoder that disentangles stylistic and semantic (content) attributes of synthetic radiology texts, enabling more reliable detection of AI-generated reports.

2 Related Work

LLMs for Radiology Report Generation. Recent works [14, 31, 48, 40, 27, 15, 4, 42] have explored the use of large language models (LLMs) for generating structured and narrative chest radiology reports. These approaches typically frame the problem either as a reasoning task over X-ray findings or as a question–answering (Q/A) formulation where the model is prompted with specific clinical queries. While such methods demonstrate the potential of LLMs in improving interpretability and capturing domain-specific language, they lack the generation of a full report using chest radiographs or radiologist-written reports. Beyond LLM-based approaches, unimodal methods [5, 45, 20, 17] rely solely on visual features from chest X-rays and employ Transformer-based architectures to produce textual reports. Although effective at aligning images with textual outputs, these models disregard the role of LLMs for contextual reasoning and often generate reports that lack the coherence, structure, and richness of human-authored narratives. More recently, prompt-driven approaches and vision LLM [22, 46, 21, 37] have attempted to bridge this gap by conditioning generation on partial report text or key clinical findings extracted from images. While these methods can improve factual alignment, they primarily focus on isolated findings rather than holistic report generation. As a result, they overlook additional clinical context and still struggle to match the fluency and logical progression found in human-written reports. Additionally, these methods lack a dataset of the type we introduce, which is developed in two ways: (1) by paraphrasing the original reports

using Instruction- or Med-LLMs, and (2) by generating complete chest radiology reports directly from radiographs.

Synthetic Radiology Reports Detection Parallel efforts have been made to identify AI-authored or erroneous medical content. For instance, recent work has proposed fact-checking examiners that differentiate real and fabricated sentences within radiology reports by aligning textual statements with associated medical images [49, 29, 34]. While such approaches verify the clinical accuracy of generated reports, they are limited to detecting local inconsistencies and do not address the broader problem of entirely fabricated reports. In addition, LLM-based proofreading techniques [32, 41] have focused on identifying specific types of errors, including negation, laterality, interval modification, and transcribing inaccuracies within otherwise genuine reports.

Limitations of current approaches While the prior works in generating synthetic chest radiology reports have gained enough progress in generating the chest radiology reports, there has been no work on discerning synthetic and human-authored reports. To this end, we focus on developing a dataset consisting of synthetic and human-authored radiology report quadruples across three LLMs. We extend the dataset by also incorporating VLMs that take in radiology images as input to generate synthetic and human-authored report triplets across two VLMs. Finally, we also provide a benchmark model for the dataset.

3 Method

In this section, we present our contributions: a novel synthetic radiology report dataset consisting of 14k samples for discerning true reports from synthetic ones, and a benchmark model on the dataset. Figure 1 presents the overall pipeline, integrating synthetic report generation with a style-content disentanglement-based classifier module to enable robust attribution of human versus LLM-authored radiology reports.

3.1 Synthetic Report Generation

Motivated by the absence of synthetic radiology report datasets tailored for authorship attribution, we introduce a novel collection of reports generated by diverse large language models (LLMs). Although AI-generated radiology reports can emerge from multiple sources, in this work we focus on two primary categories: (1) paraphrased versions of ground-truth reports produced by instruction-tuned LLMs, and (2) direct image-to-text generations from medical vision–language models (VLMs). To ensure privacy and task relevance, we remove all extraneous and personally identifiable information from the ground-truth corpus, retaining only the *findings* section. The *findings* encapsulate the radiologist's detailed analysis of chest X-ray images is rich in domain-specific terminology and descriptive language, making it the most critical component for authorship attribution. We have utilized the radiology reports from the IU-XRAY dataset [28] as ground-truth. We evaluate the quality of generated reports through standard natural language metrics - BLEU-N [35], and ROUGE-L [25]. It is important to note that, in each case, we do not evaluate the reliability and veracity of the generated reports.

Paraphrasing with Instruction-Tuned LLMs. Utilizing the text in the *findings* section of each sample, we prompt (As detailed in Table 1) various LLMs - GPT-4o [33], Medgemma-27B [11], and Mixtral-8x7B [18] - to generate a paraphrased version of the ground-truth reports. For each ground-truth sample, our dataset has a quadruple containing ground-truth and three LLM-generated reports. The prompts have been designed to maintain the diagnostic content while introducing stylistic variation, simulating adversarial paraphrasing that maintains semantic fidelity. As shown in Table 2, GPT-4o exhibits the strongest alignment with ground-truth reports with BLEU-4 and ROUGE-L scores of 28.50 and 60.17, respectively.

Captioning with Vision–Language Models for Chest X-rays. Secondly, we generate reports directly from chest radiographs using R2Gen [5] and Medgemma-4B [12], without any textual input. Unlike paraphrasing, this setting requires models to infer diagnostic content solely from visual input. Although the models have been pretrained to approximate radiologist reporting, VLMs often display stylistic artifacts individualistic to each LLM and domain-specific regularities, producing

a distribution of text distinct from human-authored and paraphrased reports. As shown in Table 2, R2Gen [5] exhibits superiority with ROUGE-L scores of 30.92.

LLM	Standardized Prompt
Text-to-Text (T2T)	
GPT-4o [33]	"You are a professional radiologist. Rephrase the following chest X-ray report in clinical language and formal tone. Only return the rewritten report as plain text."
Mixtral-8x7B [1]	"You are a professional radiologist. Rephrase the following chest X-ray report in a formal, clinical tone. Only return the reworded report. Do not add labels, headers, or formatting. Do not return JSON."
Medgemma-27B [11]	"You are a professional radiologist. Rephrase chest X-ray reports in a formal, clinical tone. Only return the paraphrased report without metadata, JSON, or special formatting."
Image-to-Text (I2T)	
R2Gen [5]	*— No prompt —*
Medgemma-4B [12]	"You are a board-certified radiologist. Write a clear, clinically sound chest X-ray report based solely on the provided image(s). Use concise prose and no section headers."

Table 1: **Prompts.** We provide a list of standardized prompts used for generating chest X-ray reports: (i) paraphrasing radiologists' reports with text-to-text LLMs, and (ii) directly generating reports from chest X-ray images with medical VLMs. R2Gen [5], being a memory-driven transformer, does not require explicit prompt instructions.

3.2 Benchmark Model

Given a text embedding $x \in \mathbb{R}^d$ of a report, our objective is to classify it into either of the binary classes - human-authored or AI-generated. Motivated by [38], we aim to decompose x into two independent latent variables: a *content embedding* c, which encapsulates the semantics of the text, and a *style embedding* s, which captures authorial cues such as phrasing, verbosity, and lexical preference that may indicate authorship. Our framework is composed of five components: (1) a Mamba encoder, which refines frozen BERT embeddings with a stack of Mamba [13] blocks to produce contextualized representations; (2) a disentanglement module to obtain s and c; (3) an LSTM decoder, which reconstructs the input sequence conditioned on both s and c to enforce semantic fidelity; (4) a classifier to supervise the style embedding s for authorship attribution; and (5) an approximation network, which attempts to predict s from c and thus provides an adversarial regularizer discouraging style–content leakage.

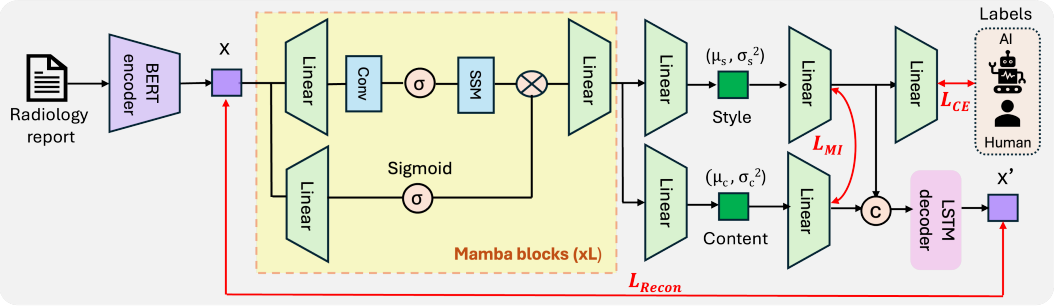


Figure 2: **Architecture.**We present the complete pipeline of our style–content disentanglement architecture. A BERT-Mamba encoder is first used to extract text embeddings from a radiology report. These embeddings are then re-parameterized into separate style and content latent representations. The two latent vectors are concatenated and passed through a decoder to reconstruct the report. To guide the disentanglement process, we incorporate a style classifier that supervises the style embeddings by predicting report authorship.

Mamba Encoder. The Mamba Encoder is constructed by stacking multiple Mamba blocks [13]. The core component of the Mamba block is a selective state space model. Unlike the self-attention

mechanism in transformers [44], which scales quadratically with sequence length, Mamba [13] parametrizes hidden state evolution as a linear recurrence in continuous time and augments it with input-dependent gating. We begin by extracting word-level representations with a frozen BERT [9] backbone. Subsequently, these embeddings are input into a stack of L Mamba blocks, where each block corresponds to a selective state space model (SSM). Formally, a state z_t evolves as

$$z_t = Az_{t-1} + Bx_t \qquad h_t = Cz_t, \tag{1}$$

where A, B, C are learnable transition matrices. The selective gating mechanism dynamically modulates A and B as a function of the input x_t, allowing the model to retain or forget information in a content-aware manner. This hybrid design combines the efficiency of recurrent state updates with the adaptability of attention-like gating. Further, each Mamba block is wrapped in normalization and residual connections to stabilize optimization:

$$h^{(l+1)} = h^{(l)} + \mathrm{Dropout}\Big(\mathrm{Mamba}\Big(\mathrm{LayerNorm}(h^{(l)})\Big)\Big), \quad l = 0, \ldots, L-1. \tag{2}$$

This architecture allows the Mamba encoder to capture both local linguistic patterns and longer-span stylistic dependencies while maintaining linear computational complexity. Finally, mean pooling is used to aggregate token-level representations into a report-level embedding $\bar{h}$.

Disentanglement module. The pooled output $\bar{h}$ from the Mamba Encoder is passed into two linear layers, producing style embedding s and content embedding c. We utilize the latent parameterization technique commonly used in VAEs [7, 19] to represent s and c as probability distributions rather than deterministic vectors. Expressing them as probability distributions is important to allow the model to capture uncertainty and to learn the representations as parameters of a multivariate Gaussian distribution. Thus, enabling the model to learn disentangled distributions for s and c effectively. Specifically, we learn:

$$q_\theta(s \mid \bar{h}) = \mathcal{N}\big(s; \mu_s(\bar{h}), \mathrm{diag}(\sigma_s^2(\bar{h}))\big) \qquad q_\phi(c \mid \bar{h}) = \mathcal{N}\big(c; \mu_c(\bar{h}), \mathrm{diag}(\sigma_c^2(\bar{h}))\big) \tag{3}$$

Subsequently, we employ the reparameterization trick to sample from these distributions, allowing gradient propagation during training.

$$s' = \mu_s + \sigma_s \odot \varepsilon_s \qquad c' = \mu_c + \sigma_c \odot \varepsilon_c \qquad \varepsilon \sim \mathcal{N}(0, I) \tag{4}$$

This results in a pair of latent representations s' and $c' \in \mathbb{R}^d$ that are concatenated $y = [s'; c']$ before being input into the decoder.

Decoder. We implement an LSTM decoder module designed to reconstruct our text embedding x conditioned on y, and trained using teacher-forcing. Teacher-forcing is widely used in training sequence generation models to improve sampling efficiency and to stabilize training [26]. Unlike pure autoregressive approaches, where the decoder feeds its own predicted token back as input for the next step, in teacher-forcing, we feed the ground-truth token from the training data into the next step. The output generated at each step is mapped to vocabulary logits through a linear projection, followed by softmax over the vocabulary logits.

$$w_t = \mathrm{LSTM}(x_{<t}; y) \qquad p(x_t \mid x_{<t}, y) = \mathrm{Softmax}(\mathrm{Linear}(w_t)) \tag{5}$$

3.3 Training Objectives

We train the model with three complementary objectives: (i) reconstruction, (ii) style classification, and (iii) mutual information regularization. Together, these enforce disentangled yet informative representations of style and content.

Style classification. To ensure that the style embedding s' encodes authorial attributes, we train a lightweight classifier g_ψ on top of s'. The classifier is optimized with standard cross-entropy:

$$\mathcal{L}_{\mathrm{cls}} = -\mathbb{E}_y\big[\log p_\psi(y \mid s')\big] \qquad p_\psi(y \mid \mathbf{s}) = \mathrm{Softmax}\big(g_\psi(s')\big). \tag{6}$$

Mutual information regularization. To prevent the content embedding c' from leaking stylistic cues, we introduce an approximation network f_ϕ that predicts s' from c'. For a matched pair (s'_j, c'_j) and a mismatched pair $(s'_j, c'_k) \forall k \neq j$, the log-likelihoods are:

$$\log p_\phi(\mathbf{s}_j \mid \mathbf{c}_j) \propto -\tfrac{1}{2} \left\| \mathbf{s}_j - f_\phi(\mathbf{c}_j) \right\|^2 \qquad \log p_\phi(\mathbf{s}_j \mid \mathbf{c}_k) \propto -\tfrac{1}{2} \left\| \mathbf{s}_j - f_\phi(\mathbf{c}_k) \right\|^2 \tag{7}$$

Following the CLUB objective [39, 6], we upper bound the mutual information, and define the penalty as a clipped bound in range $[0, 1]$:

$$\mathcal{L}_{\mathrm{MI}} = \mathrm{clip}\big(\widehat{I}^{\text{upper}}(s'; c')\big) \qquad \widehat{I}^{\text{upper}}(s'; c') = \mathbb{E}\big[\log p_\phi(s'_j \mid c'_j) - \log p_\phi(s'_j \mid c'_k)\big] \tag{8}$$

Reconstruction. Given latent codes (s', c'), the decoder is trained to reconstruct the input sequence under teacher forcing, where N the sequence length, $\hat{x}_t$ denotes the predicted distribution, and x_t is the ground-truth token at step t. This loss enforces both embeddings to preserve sufficient information for faithful reconstruction.

$$\mathcal{L}_{\mathrm{rec}} = -\tfrac{1}{N} \sum_{t=1}^{N} \log p(\hat{x}_t \,|\, \mathbf{s}, \mathbf{c}, x_{<t}) \,, \tag{9}$$

Total objective. The overall training loss is as follows, where β is an annealed coefficient controlling the strength of the mutual information penalty:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} \; + \; \beta \, \mathcal{L}_{\mathrm{MI}} \; + \; \mathcal{L}_{\mathrm{cls}}, \tag{10}$$

4 Experiments

4.1 Dataset

We use the IU-Xray dataset [28] to obtain the chest radiology reports. IU-X-ray is a standard benchmark for medical vision–language modeling. The dataset consists of 7,470 chest radiographs paired with 3,955 radiology reports. Specifically, we utilize the subset of 2,955 preprocessed annotated samples curated as per the R2Gen [5] benchmark, which includes chest radiographs paired with corresponding radiology reports. These reports, originally collected by Indiana University, are linked to one or more frontal or lateral chest X-ray views and contain structured descriptions of clinical findings and impressions.

4.2 Experimental Setup

Implementation details. We implement our style-content disentanglement pipeline in PyTorch [36] version 2.7.1 with cuda11.8. Input reports are tokenized with the `bert-base-uncased` tokenizer [9], truncated or padded to a maximum length of 512 tokens. The encoder projects into a 32-dimensional style embedding and a 512-dimensional content embedding; the decoder uses hidden dimension 512. We use the Mamba encoder of *two* mamba blocks. The benchmark model is trained for 40 epochs with a batch size of 32 using the Adam optimizer with a learning rate of 1×10^{-4}. We save checkpoints every 5 epochs. All experiments are run on NVIDIA A100 GPUs. The classifier MLP probe is trained for 30 epochs

Evaluation metrics. Detection performance is evaluated using four complementary metrics: (i) overall accuracy, (ii) macro-averaged F1 score, (iii) area under the ROC curve (AUROC), and (iv) Matthews correlation coefficient (MCC). To probe robustness, we report results stratified along two axes: (a) generator family GPT-4o [33], Mixtral-8x7B [1], MedGemma-27B [11], MedGemma-4B [12] and R2Gen [5] used for capturing the style cues which are then leveraged by a style-content disentanglement model, and (b) paraphrase intensity, defined by BLEU-N [35] and ROUGE-L [25] between generated and reference reports. The subsequent analysis measures the extent to which performance declines when paraphrases increasingly vary in wording.

Classifier model. For authorship attribution, we train a simple linear layer on the style embeddings s' extracted from our disentanglement model. s' is extracted from the frozen encoder and into an

Model	BLEU-1 (%)	BLEU-2 (%)	BLEU-3 (%)	BLEU-4 (%)	ROUGE-L (%)
Text-to-Text (T2T)					
GPT-4o [33]	60.08 ± 0.14	46.11 ± 0.13	36.21 ± 0.11	28.50 ± 0.11	60.17 ± 0.35
Mixtral-8x7B [1]	55.26 ± 0.06	40.47 ± 0.06	30.28 ± 0.05	22.80 ± 0.04	55.42 ± 0.08
MedGemma-27B-Text [11]	61.65 ± 0.03	46.11 ± 0.04	35.70 ± 0.05	27.83 ± 0.04	57.22 ± 0.01
Image-to-Text (I2T)					
R2Gen [5]	44.47 ± 0.00	28.26 ± 0.00	20.39 ± 0.00	15.04 ± 0.00	30.92 ± 0.00
Medgemma-4B [12]	42.31 ± 0.00	25.98 ± 0.00	16.68 ± 0.00	10.95 ± 0.00	27.93 ± 0.00

Table 2: **BLEU-N and ROUGE-L evaluation of generated reports.** We evaluate the quality of our generated dataset using BLEU-N and ROUGE-L scores across a range of instruction-tuned LLMs and image-to-text models. The results show that GPT-4o [33] achieved the highest ROUGE-L score (60.17) in the text-to-text category, indicating the strongest lexical similarity to human-written reports. In comparison, R2Gen [5] obtained the highest ROUGE-L score (30.92) among image-to-text models.

MLP (512→256→128), each followed by BatchNorm, ReLU, and Dropout, and a final two-way classification head. The linear layer is optimized with AdamW learning rate of 5×10^{-4} and weight decay of 5×10^{-5} using cross-entropy loss. All classification models are trained on the 70% training split and validated on 15% of data; final results are reported on the held-out 15% test set.

5 Results

In this section, we present our results quantifying both the quality of our proposed dataset and classification pipeline, which are assessed using the metrics described in Sec. 4. Due to the probabilistic nature of generative models and to ensure reproducibility, all results are averaged over five independent runs and reported as mean $\pm$ standard deviation.

5.1 Text-to-Text (T2T)

We present the dataset and classification analysis of the synthetic reports produced by instruction-tuned LLMs, GPT-4o [33], Mixtral-8x7B [1], and MedGemma-27B [11], when prompted to paraphrase the ground-truth radiologist report.

Dataset Analysis. We evaluate generated outputs against reference radiologist reports using BLEU-N and ROUGE-L to capture lexical fidelity and coherence, and benchmark paraphrasing behavior of LLMs on radiology reports, while not directly measuring clinical correctness. Among the models compared, GPT-4o [33] achieves the highest lexical fidelity, with BLEU-4 of 28.50 ± 0.11 and ROUGE-L of 60.17 ± 0.35, followed by MedGemma-27B [11] and Mixtral-8x7B [1]. Overall, we observe a strong overlap compared to the ground-truth across all models, with variations solely reflecting medical terminology choices. Examples of generated samples have been shown in Table 6 (Appendix).

Classification Analysis. We evaluate authorship classification with our style–content disentanglement model. GPT-4o [33] generated reports can be classified with the highest MCC 99.64 ± 0.30, followed by Mixtral-8x7B [1] with 99.55 ± 0.16 and MedGemma-27B-Text [11] with 99.10 ± 0.00. These results suggest that the model captures stable stylistic cues and that **writing style differs between human-authored reports and LLM-generated ones**.

5.2 Image-to-Text (I2T)

In this section, we evaluate models R2Gen [5] and MedGemma-4B [12] on the Image-to-Text (I2T) setting, where the task is to generate radiology reports from chest radiograph images.

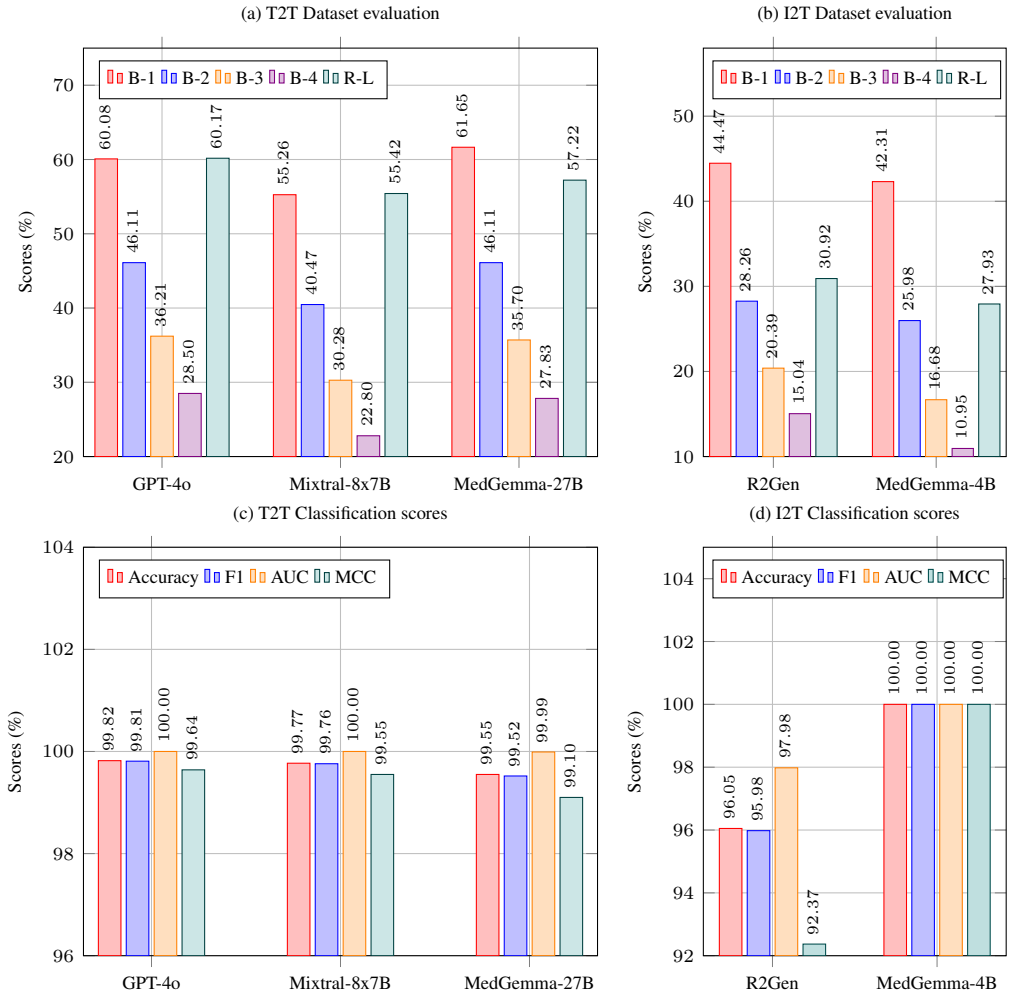


Figure 3: **Bar plot results across categories.** Plots (a) and (b) show BLEU-N and ROUGE-L scores for the text-to-text and image-to-text categories, respectively, while plots (c) and (d) present authorship classification metrics for the same categories. In the text-to-text setting, GPT-4o [33] achieves the highest ROUGE-L score, indicating the strongest lexical similarity to radiologist-written reports, and also obtains the highest MCC score for authorship classification. For the image-to-text setting, R2Gen [5] attains the best ROUGE-L score, while MedGemma-4B achieves the highest MCC score.

Dataset Analysis As shown in Table 2, R2Gen [5] attains a ROUGE-L of 30.92 ± 0.00, while Medgemma-4B achieves 27.93 ± 0.00. Relative to paraphrase-based text-to-text generation, these image-to-text models show lower lexical overlap with the references, indicating reduced coherence in the vision-conditioned setting. It is important to note that the models have already been trained on the IU-XRAY dataset [28], and our generation is zero-shot.

Classification Analysis We observe that Medgemma-4B [12] generated reports can be perfectly classified with an MCC of 100.00 ± 0.00, while R2Gen [5] reaches 92.37 ± 0.00. The strong MCC score for Medgemma-4B indicates complete separability of its style from human reports. The slightly lower MCC for R2Gen suggests that its generation is more human-like. Overall, these results show that the disentanglement features capture consistent stylistic signals and enable reliable authorship identification for I2T-generated reports.

5.3 Cross-LLM Detection Results

To further validate the robustness of our classifier model, we perform out-of-dataset (OOD) evaluations. In OOD experiments, the classifier is trained on generated reports of one LLM and tested on reports of another LLM. The goal is to understand the ability of the model to disentangle and capture unseen styles from similar content. It is important to note that the generalizability and robustness of the model is evaluated on the MCC scores (Table 3) since it includes all elements of the confusion matrix.

In the T2T setting, training on the GPT-4o [33] or Mixtral-8x7B [1] generated dataset yields robust models adept at disentangling style and content. Whereas in the I2T setting, training on Medgemma-4B [12] generated reports provides us with a more generalizable classification model. We attribute this to the Medgemma-4B reports, which exhibit greater stylistic variations, hence exposing the model to difficult samples. The cross-LLM results support the robustness of the style–content disentanglement approach as the encoders generalizes across generators and input modalities, capturing the differentiable stylistic cues between human and LLMs generated reports.

Train Dataset	Test Dataset	Accuracy	F1	AUC	MCC
Text-to-Text (T2T)					
GPT-4o [33]	Mixtral-8x7B [1]	99.41 ± 0.12	99.38 ± 0.13	99.99 ± 0.00	98.83 ± 0.25
GPT-4o [33]	Medgemma-27B [11]	98.31 ± 0.00	98.23 ± 0.00	99.91 ± 0.00	96.64 ± 0.00
Mixtral-8x7B [1]	GPT-4o [33]	99.48 ± 0.15	99.45 ± 0.16	99.94 ± 0.03	98.96 ± 0.30
Mixtral-8x7B [1]	Medgemma-27B [11]	98.42 ± 0.00	98.34 ± 0.00	99.79 ± 0.01	96.84 ± 0.00
Medgemma-27B [11]	GPT-4o [33]	98.58 ± 0.13	98.50 ± 0.14	99.91 ± 0.05	97.15 ± 0.26
Medgemma-27B [11]	Mixtral-8x7B [1]	98.67 ± 0.19	98.60 ± 0.20	99.93 ± 0.02	97.34 ± 0.37
Image-to-Text (I2T)					
Medgemma-4B [12]	R2Gen [5]	99.32 ± 0.00	99.29 ± 0.00	99.95 ± 0.00	98.65 ± 0.00
R2Gen [5]	Medgemma-4B [12]	98.42 ± 0.00	98.35 ± 0.00	99.90 ± 0.00	96.86 ± 0.00

Table 3: **Cross-LLM detection.** We evaluate our style–content disentanglement pipeline by training on reports generated by one LLM and testing on reports from a different, unseen LLM. In the text-to-text category, the model trained on Mixtral-8x7B [1] and evaluated on GPT-4o [33] achieves the highest detection performance. In the image-to-text category, the model trained on MedGemma-4B [12] and evaluated on R2Gen [5] demonstrates the strongest detection performance.

6 Conclusion

In this paper, we present a new 14k samples dataset of synthetic chest radiology reports that is created through two complementary generation pathways: instruction-tuned large language models for text-to-text generation, and vision–language models for image-to-text generation. By building this dataset, we aim to study the problem of discerning AI-generated medical reports from human-authored medical reports. To this end, we develop a BERT-Mamba classifier model that disentangles style and content from the input text. Extensive experiments across three LLM generated reports and two VLM generated reports indicate that text-to-text models can achieve very high overlap with ground-truth radiologist reports, while image-to-text models face a more difficult challenge because they must link visual information with language. At the same time, our authorship attribution experiments reveal that even when the lexical overlap is strong, stylistic signals remain detectable. The ability to classify human versus AI-authored reports with very high MCC values demonstrates that AI-generated text carries distinctive style markers, even in the medical domain. By releasing this dataset and benchmarks, we provide a controlled testbed for future research on the reliability, trustworthiness, and safe deployment of AI-generated clinical narratives.

References

[1] Mistral AI. *Mixtral-8×7B Instruct v0.1 – Hugging Face*. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1. Accessed: 2025-09-02. 2025.

[2] *Bidirectional Long Short-Term Memory Network*. ScienceDirect, Computer Science Topics. (accessed: 2025-09-03). n.d.

[3] Lawrence Camarda et al. "Self-inflicted long bone fractures for insurance fraud". In: *International Journal of Legal Medicine* (2019). Online ahead of print; PMID: 29943089. URL: https://pubmed.ncbi.nlm.nih.gov/29943089/.

[4] Zhihong Chen et al. *A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation*. 2024. arXiv: 2401.12208 [cs.CV]. URL: https://arxiv.org/abs/2401.12208.

[5] Zhihong Chen et al. "Generating Radiology Reports via Memory-driven Transformer". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Nov. 2020.

[6] Pengyu Cheng et al. *CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information*. 2020. arXiv: 2006.12013 [cs.LG]. URL: https://arxiv.org/abs/2006.12013.

[7] Marissa C. Connor, Gregory H. Canal, and Christopher J. Rozell. "Variational Autoencoder with Learned Latent Structure". In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Available at arXiv:2006.10597. 2021. URL: http://proceedings.mlr.press/v130/connor21a/connor21a.pdf.

[8] Tri Dao and Albert Gu. "Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality". In: *International Conference on Machine Learning (ICML)*. 2024.

[9] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[10] *Document Templates Generator – OnlyFake*. Accessed 2025-08-10. OnlyFake. 2024. URL: https://www.onlyfake.org/.

[11] Google. *MedGemma Hugging Face*. https://huggingface.co/google/medgemma-27b-text-it. Accessed: [Insert your access date, e.g., 2025-09-02]. 2025.

[12] Google. *MedGemma Hugging Face*. https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4. Accessed: 2025-09-02. 2025.

[13] Albert Gu and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *arXiv preprint arXiv:2312.00752* (2023).

[14] Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. "A Dataset of Medical Questions Paired with Automatically Generated Answers and Evidence-supported References". In: *Scientific Data* 12.1 (2025), p. 1035. DOI: 10.1038/s41597-025-05233-z. URL: https://doi.org/10.1038/s41597-025-05233-z.

[15] Iryna Hartsock and Ghulam Rasool. "Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review". In: *arXiv preprint arXiv:2403.02469* (2024). URL: https://arxiv.org/abs/2403.02469.

[16] Ari Holtzman et al. *The Curious Case of Neural Text Degeneration*. 2020. arXiv: 1904.09751 [cs.CL]. URL: https://arxiv.org/abs/1904.09751.

[17] Ziyi Huang, Xuemeng Zhang, and Shu Zhang. "KIUT: Knowledge-Injected UTransformer for Radiology Report Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 19809–19818. DOI: 10.1109/CVPR52729.2023.01906.

[18] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[19] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML]. URL: https://arxiv.org/abs/1312.6114.

[20] Hyungyung Lee et al. *Vision-Language Generative Model for View-Specific Chest X-ray Generation*. 2024. arXiv: 2302.12172 [eess.IV]. URL: https://arxiv.org/abs/2302.12172.

[21] Suhyeon Lee et al. *LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation*. 2024. arXiv: 2305.11490 [cs.CV]. URL: https://arxiv.org/abs/2305.11490.

[22] Hongzhao Li et al. "Prompt-Guided Generation of Structured Chest X-Ray Report Using a Pre-trained LLM". In: *2024 IEEE International Conference on Multimedia and Expo (ICME)*. 2024, pp. 1–6. DOI: 10.1109/ICME57554.2024.10687707.

[23] Hui Li et al. *A Multi-Agent Framework with Automated Decision Rule Optimization for Cross-Domain Misinformation Detection*. 2025. arXiv: 2503.23329 [cs.AI]. URL: https://arxiv.org/abs/2503.23329.

[24] Kuntao Li, Yifan Chen, Qiaofeng Wu, et al. "Cross-Domain Fake News Detection based on Dual-Granularity". In: *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. 2025. URL: https://aclanthology.org/2025.coling-main.631/.

[25] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics. 2004, pp. 74–81.

[26] Guan-Yu Lin and Pu-Jen Cheng. "R-TeaFor: Regularized Teacher-Forcing for Abstractive Summarization". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6303–6311. DOI: 10.18653/v1/2022.emnlp-main.423. URL: https://aclanthology.org/2022.emnlp-main.423/.

[27] Bo Liu et al. "SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 475–483. DOI: 10.1145/3474085.3475456.

[28] Weihua Liu et al. *IU X-ray Dataset*. https://doi.org/10.57702/15w6bsxf. Accessed: 2025-08-11. 2025.

[29] Razi Mahmood et al. *Fact-Checking of AI-Generated Reports*. 2025. arXiv: 2307.14634 [cs.AI]. URL: https://arxiv.org/abs/2307.14634.

[30] Deekshith Marla. *Generative AI and Insurance Fraud: The Deepfake Threat Across Insurance Lines*. Accessed 2025-08-10. Arya.ai. June 2025. URL: https://arya.ai/blog/artificial-intelligence-fraud-in-insurance.

[31] P. Müller et al. *MIMIC-Ext-CXR-QBA: A Structured, Tagged, and Localized Visual Question Answering Dataset with Question-Box-Answer Triplets and Scene Graphs for Chest X-ray Images (version 1.0.0)*. RRID:SCR_007345. 2025. DOI: 10.13026/8qmz-da41. URL: https://doi.org/10.13026/8qmz-da41.

[32] Hannah Murphy. "GPT-4 can proofread radiology reports for a penny apiece". In: *Health Imaging* (Jan. 2025). URL: https://healthimaging.com/topics/health-it/enterprise-imaging/imaging-informatics/gpt-4-can-proofread-radiology-reports-penny-apiece.

[33] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[34] Igor V. Pantic and Snezana Mugosa. "Artificial intelligence strategies based on random forests for detection of AI-generated content in public health". In: *Public Health* 242 (2025), pp. 382–387. ISSN: 0033-3506. DOI: https://doi.org/10.1016/j.puhe.2025.03.029. URL: https://www.sciencedirect.com/science/article/pii/S0033350625001489.

[35] Kishore Papineni et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*. Association for Computational Linguistics. 2002, pp. 311–318.

[36] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: 1912.01703 [cs.LG]. URL: https://arxiv.org/abs/1912.01703.

[37] Chantal Pellegrini et al. "RaDialog: Large Vision-Language Models for X-Ray Reporting and Dialog-Driven Assistance". In: *Medical Imaging with Deep Learning*. 2025.

[38] Karol Przystalski et al. "Stylometry recognizes human and LLM-generated texts in short samples". In: *Expert Systems with Applications* 296 (Jan. 2026), p. 129001. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2025.129001. URL: http://dx.doi.org/10.1016/j.eswa.2025.129001.

[39] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. *Learning Disentangled Representations via Mutual Information Estimation*. 2019. arXiv: 1912.03915 [stat.ML]. URL: https://arxiv.org/abs/1912.03915.

[40] Francesco Dalla Serra et al. *Grounding Chest X-Ray Visual Question Answering with Generated Radiology Reports*. 2025. arXiv: 2505.16624 [cs.CV]. URL: https://arxiv.org/abs/2505.16624.

[41] Cong Sun et al. "Generative Large Language Models Trained for Detecting Errors in Radiology Reports". In: *Radiology* 315.2 (2025), e242575. DOI: 10.1148/radiol.242575.

[42] Omkar Chakradhar Thawakar et al. "XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models". In: *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Ed. by Dina Demner-Fushman et al. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 440–448. DOI: 10.18653/v1/2024.bionlp-1.35. URL: https://aclanthology.org/2024.bionlp-1.35/.

[43] Talles Viana Vargas, Hélio Pedrini, and André Santanchè. "LLM-Driven Chest X-Ray Report Generation With a Modular, Reduced-Size Architecture". In: *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*. 2024, pp. 199–211. DOI: 10.1007/978-3-031-79032-4_14.

[44] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[45] Zhanyu Wang et al. "METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. -–-. URL: https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_METransformer_Radiology_Report_Generation_by_Transformer_With_Multiple_Learnable_Expert_CVPR_2023_paper.pdf.

[46] Zhanyu Wang et al. "R2GenGPT: Radiology Report Generation with frozen LLMs". In: *Meta-Radiology* 1.3 (2023), p. 100033. ISSN: 2950-1628. DOI: https://doi.org/10.1016/j.metrad.2023.100033. URL: https://www.sciencedirect.com/science/article/pii/S2950162823000334.

[47] Zilong Wang et al. *LLM-RadJudge: Achieving Radiologist-Level Evaluation for X-Ray Report Generation*. 2024. arXiv: 2404.00998 [cs.CL]. URL: https://arxiv.org/abs/2404.00998.

[48] J. Wu et al. *Chest ImaGenome Dataset (version 1.0.0)*. RRID:SCR_007345. 2021. DOI: 10.13026/wv01-y230. URL: https://doi.org/10.13026/wv01-y230.

[49] Junchao Wu et al. *DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios*. 2025. arXiv: 2410.23746 [cs.CL]. URL: https://arxiv.org/abs/2410.23746.

A Technical Appendices and Supplementary Material

A.1 Ablation Studies

We conduct three ablations of the style–content disentanglement model : (1) *Depth of Mamba blocks:* we vary the number of Mamba [13] blocks to measure how model depth affects performance. (2) *Encoder Architecture:* we replace Mamba blocks with alternative encoders of comparable parameter count such as transformer encoder [44] or BiLSTM [2] to test the architecture specfic gains. (3) *Training Objectives:* we compare the sum of all the training objectives with variants that remove individual loss function to quantify each component's contribution. All models are trained on the same splits and we report Accuracy, F1, AUC, and MCC as mean $\pm$ std over five runs.

Effect of Mamba Layer Depth. In this ablation study, we investigate the effect of encoder depth by comparing Mamba-based architectures with varying numbers of layers (one, two, four, six and eight). The detailed results are presented in Table 4. Overall, our findings indicate that stacking multiple Mamba layers consistently outperforms the single-layer variant. Specifically, the two-layer architecture yields improvements of 1.71, 1.81, 0.21, and 3.44 points in Accuracy, F1, AUC and MCC respectively relative to the single-layer baseline. Moreover, the two-layer configuration also outperforms the four-layer model by 0.11, 0.12, and 0.23 points in Accuracy, F1, and MCC. The six-layer model performs comparably to the two-layer model with a drop of 0.05 points in MCC, suggesting diminishing returns beyond a certain depth. Finally, the eight-layer architecture underperforms the two-layer variant, with reductions of 0.18 and 0.19 and 0.36 points in Accuracy,F-1 and MCC respectively.

Encoder	No. of layers	Accuracy (%)	F1 (%)	AUC (%)	MCC (%)
Transformer encoder [44]	2	99.75 ± 0.09	99.74 ± 0.10	99.99 ± 0.00	99.50 ± 0.19
Bi-LSTM [2]	2	99.41 ± 0.15	99.38 ± 0.16	99.99 ± 0.00	98.83 ± 0.30
Mamba [13]	**2**	$\mathbf{99.82 \pm 0.15}$	$\mathbf{99.81 \pm 0.16}$	$\mathbf{100.00 \pm 0.00}$	$\mathbf{99.64 \pm 0.34}$
Mamba [13]	1	98.11 ± 0.22	98.00 ± 0.23	99.79 ± 0.03	96.20 ± 0.44
Mamba [13]	4	99.71 ± 0.01	99.69 ± 0.11	100.00 ± 0.00	99.41 ± 0.20
Mamba [13]	6	99.80 ± 0.09	99.79 ± 0.10	100.00 ± 0.00	99.59 ± 0.19
Mamba [13]	8	99.64 ± 0.09	99.62 ± 0.10	100.00 ± 0.00	99.28 ± 0.19

Table 4: **Ablation study of encoder architectures.** We assess different encoder designs for our style–content disentanglement pipeline, comparing the original two-block Mamba [13] encoder with a single Mamba block, a two-layer BiLSTM [2], a two-layer Transformer [44], and deeper Mamba variants (four, six, and eight blocks). The two-block Mamba encoder achieves the best performance, suggesting it offers the right balance between capacity and generalization.

Encoder Ablation To evaluate the specific contribution of the encoder, we replace the two-layer Mamba stack with two alternatives of the same depth: (i) a two-layer BiLSTM [2] and (ii) a two-layer Transformer [44] encoder, holding all other components and training settings fixed. Relative to Mamba [13], the BiLSTM [2] variant reduces Accuracy, F1, AUC, and MCC by 0.41, 0.43, 0.01, and 0.81 points, respectively. The Transformer variant shows similar declines of 0.07, 0.07, 0.01, and 0.14 points on the same metrics. These results indicate that the two-layer Mamba encoder is the strongest among the tested options, consistent with better modeling of long-range dependencies.

Effect of Training Objectives. We study the impact of the training objective on detection by training the style–content model with different combinations of the losses defined in Section 3.3: reconstruction ($\mathcal{L}_{\text{rec}}$), classification ($\mathcal{L}_{\text{cls}}$), and mutual-information regularization ($\mathcal{L}_{\text{mi}}$). All other settings (architecture, data splits, optimization) are held fixed, and results are reported as mean $\pm$ standard deviation over five runs (Table 5). The full objective $\mathcal{L} =, \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mi}}$ yields the strongest performance across Accuracy, F_1, AUC, and MCC. Training with $\mathcal{L}_{\text{mi}}$ alone gives the lowest MCC 83.28 ± 0.96, and the combination $\mathcal{L}_{\text{rec}}+\mathcal{L}_{\text{mi}}$ remains weaker than any setting that includes $\mathcal{L}_{\text{cls}}$, underscoring the need for explicit supervision on stylistic labels. These results suggest complementary roles for the three terms: $\mathcal{L}_{\text{cls}}$ aligns the style representation with the authorship label space, $\mathcal{L}_{\text{rec}}$

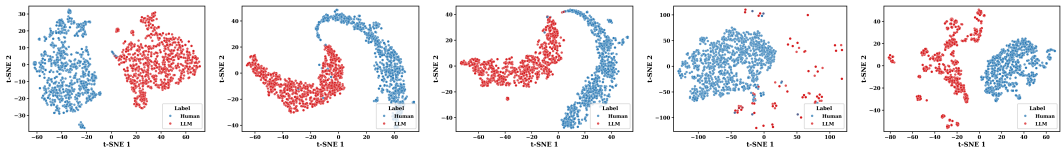
maintains content fidelity and stabilizes training, and $\mathcal{L}_{\text{mi}}$ limits leakage between style and content representations. We therefore adopt the full objective in the main experiments.

L_{recon}	L_{cls}	L_{mi}	Acc. (%)	F1 (%)	AUC (%)	MCC (%)
✓	×	×	98.58 ± 0.39	98.50 ± 0.41	99.87 ± 0.03	97.15 ± 0.78
×	✓	×	99.80 ± 0.15	99.79 ± 0.16	100.00 ± 0.00	99.59 ± 0.29
×	×	✓	91.34 ± 0.52	91.35 ± 0.49	97.22 ± 0.17	83.28 ± 0.96
✓	✓	×	99.73 ± 0.10	99.71 ± 0.11	100.00 ± 0.00	99.46 ± 0.20
✓	×	✓	98.11 ± 0.22	97.99 ± 0.23	99.88 ± 0.05	96.20 ± 0.44
×	✓	✓	99.77 ± 0.14	99.76 ± 0.15	100.00 ± 0.00	99.55 ± 0.28
✓	✓	✓	**99.82 ± 0.15**	**99.81 ± 0.16**	**100.00 ± 0.00**	**99.64 ± 0.34**

Table 5: **Effect of training objectives.** We perform an ablation study on different combinations of training objectives for authorship detection. A ✓ indicates inclusion and a × exclusion of the corresponding loss. Results show that omitting reconstruction and classification losses severely degrades performance, while the full pipeline combining reconstruction, classification, and mutual information losses achieves the best detection accuracy, highlighting the complementary role of these objectives.

A.2 Latent space visualization using t-SNE plots

To illustrate the effect of disentangling style from content, we visualize the learned representations with t-SNE (two-dimensional projection), as shown in Fig. 4. Each point corresponds to a report-level embedding produced by the encoder. In the *style* space, human-generated embeddings (blue) and LLM-generated embeddings (red) form well-separated groups, indicating that authorship-related cues are concentrated in the style representation. We use t-SNE only as a qualitative visualization. The observed separation in the style projection agrees with our quantitative authorship results, and serves as a visual check that the model's style encoder captures stable, source-specific patterns.



(a) T2T – GPT-4o (b) T2T – Mixtral-8x7B (c) T2T – MedGemma-27B (d) I2T – R2Gen (e) I2T - MedGemma-4B

Figure 4: **Latent space visualization.** We show t-SNE projections of the learned style and content latent representations across models. In both the text-to-text (T2T) and image-to-text (I2T) settings, human-authored reports (blue) and LLM-generated reports (red) form distinct clusters, indicating that the framework successfully disentangles stylistic signals. The clear separation of human and LLM clusters across all five models further demonstrates the strength of this disentanglement.

A.3 Dataset Samples

We present representative examples from our generated dataset for both text-to-text (T2T) and image-to-text (I2T) categories in Table 6 and Figure 5. Colored spans (violet, magenta, teal) mark lexical and semantic divergences relative to the reference, allowing a side-by-side view of how models rewrite or summarize the same findings. In the text-to-text cases, models tend to follow stable report templates and differ mainly in word choices and clause order, for example, preferring "cardiac silhouette" versus "heart size," inserting phrases such as "no evidence of," or reordering statements about effusions and pneumothorax. These substitutions are often clinically equivalent but reduce n-gram overlap and slightly shift the perceived level of certainty.

Reference Report	GPT-4o [33]	Mixtral-8x7B [1]	MedGemma-27B [11]
the heart size and pulmonary vascularity appear within normal limits. a large hiatal hernia is noted. the lungs are free of focal airspace disease. no pneumothorax or pleural effusion is seen. degenerative changes are present in the spine.	the cardiac silhouette and pulmonary vasculature are within normal limits. there is a large hiatal hernia observed. the lung fields are clear of any focal airspace disease. no pneumothorax or pleural effusion. degenerative changes are noted in the spine.	the heart size and pulmonary vascularity are within normal ranges. a significant hiatal hernia is detected. the lungs are clear of any localized airspace disease. no pneumothorax or pleural effusion. degenerative alterations are visible in the spine.	the cardiac silhouette is within normal limits. The chest radiograph demonstrates findings consistent with streaky airspace disease. No pneumothorax or pleural effusion is identified.
left lower lobe calcified granuloma.heart size normal. no pleural effusion or pneumothorax. mild medial right atelectasis. mild emphysema.	the chest x-ray reveals a calcified granuloma in the left lower lobe. the heart size appears within normal limits. no pleural effusion or pneumothorax. mild atelectasis in the right medial lung region, with mild emphysematous changes.	there is a calcified granuloma in the left lower lobe. The cardiac silhouette is within normal limits. No pleural effusion or pneumothorax. Mild atelectasis in the medial aspect of the right lung. Findings consistent with mild emphysema.	the chest x-ray demonstrates a calcified granuloma in the left lower lobe, with a normal heart size. no pleural effusion or pneumothorax. mild atelectasis in the medial right portion of the lung, with mild emphysematous changes.
cardiac and mediastinal contours are within normal limits. the lungs are clear. bony structures are intact	the cardiac silhouette and mediastinal contours appear normal. the pulmonary fields are clear, and the osseous structures are intact.	the cardiac and mediastinal borders are observed to be within normal parameters. no abnormalities are detected within the lung fields. the osseous structures are intact, with no evidence of fracture or other trauma.	the cardiac and mediastinal borders are observed to be within normal parameters. no abnormalities are detected within the lung fields. the osseous structures are intact, with no evidence of fracture or other trauma.
the lungs are clear bilaterally. specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion. cardio mediastinal silhouette is unremarkable. visualized osseous structures of the thorax are without acute abnormality.	the bilateral lung fields are clear, with no signs of focal consolidation, pneumothorax, or pleural effusion observed. the cardiomediastinal silhouette appears unremarkable. the visualized bony structures of the thorax do not demonstrate any acute abnormalities.	the lungs show no signs of abnormalities bilaterally, with no focal consolidation, pneumothorax, or pleural effusion observed. the cardiac and mediastinal shadow is unremarkable. the thoracic osseous structures are devoid of any acute changes.	the lungs are clear bilaterally without evidence of consolidation, pneumothorax, or pleural effusion. The cardiomediastinal silhouette is within normal limits. Visualized thoracic osseous structures are unremarkable.
well circumscribed 11 mm right upper lobe nodule, unchanged appearance from previous examination 7 years ago. the trachea is midline. negative for pneumothorax, pleural effusion. the heart size is normal. redemonstrated syndesmophyte.	the chest x-ray reveals a well-defined 11 mm nodule located in the right upper lobe, with no change in appearance since the previous examination conducted seven years ago. the trachea is positioned at the midline. there is no evidence of pneumothorax or pleural effusion. cardiac silhouette appears normal in size. a syndesmophyte is again noted.	the chest x-ray reveals a well-defined 11 millimeter nodule in the right upper lobe, with no alteration in appearance compared to the examination conducted 7 years prior. the trachea is centrally located. no evidence of pneumothorax or pleural effusion is detected. the heart size is within normal limits. a syndesmophyte is once again visualized.	A well-circumscribed 11 mm nodule is identified in the right upper lobe, stable in appearance compared to a prior examination from 7 years ago. The trachea is midline. No pneumothorax or pleural effusion is present. Cardiac silhouette is within normal limits. A syndesmophyte is again noted.

Table 6: **Text-to-text (T2T) examples from our chest radiology dataset.** We present five examples from our generated dataset for different instruction-tuned LLMs. Colored words highlight lexical and semantic differences from the reference reports in the IU-Xray dataset [28]. A homogeneous reporting style is observed in the LLM-generated reports, mainly through paraphrasing. We use violet for GPT-4o [33], magenta for Mixtral-8x7B [1], and teal for MedGemma-27B [11] to show the lexical differences.

By contrast, image-to-text outputs show greater variation in phrasing and brevity, reflecting the added difficulty of linking visual cues to text. Descriptions may generalize findings ("clear lungs" vs. "no focal airspace disease") or omit modifiers such as laterality and degree ("mild," "streaky"), which contributes to lower lexical overlap with the references. Overall, the qualitative patterns in Table 6 reflect the quantitative results.

A.4 Dataset Generation Hyperparameters

We construct our synthetic dataset using a diverse set of large language models (LLMs), including GPT-4o [33], Mixtral-8x7B [1], and MedGemma-27B [11]. To ensure variability and control over text

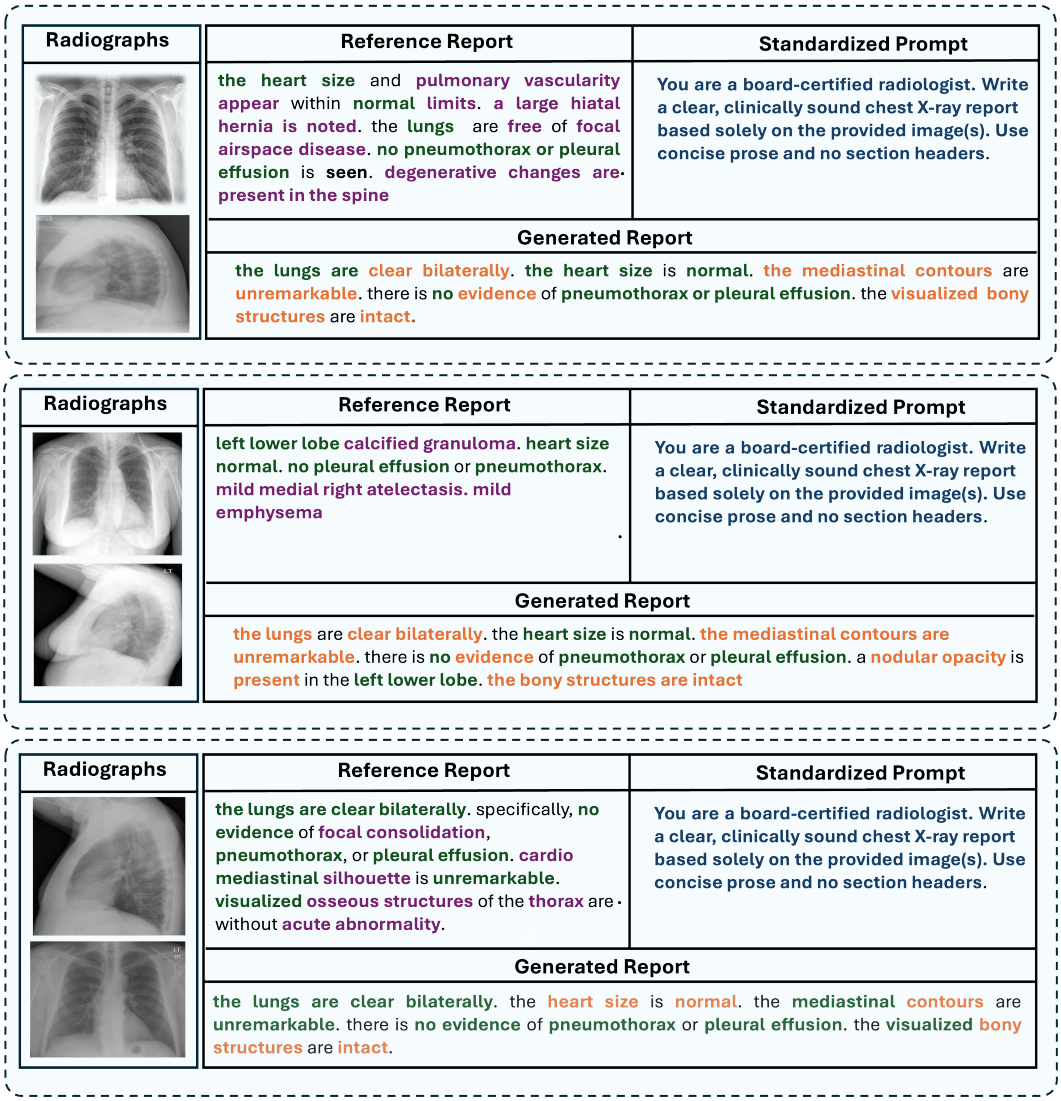


Figure 5: **Dataset samples from Image-to-Text settings.** We showcase three examples from our generated dataset designed using vision-language model (VLM). Each example includes the input image(s), the corresponding ground truth report, and a standardized prompt. Both the reference and generated reports are provided. Words that appear in both reports are highlighted in **green**, words unique to the reference report are shown in **violet**, and words unique to the generated report are shown in **orange**. The prompt text is highlighted in **blue**.

Model	Top-p	Temperature	Max new tokens	Min words / sample	Max words / sample	Avg words / sample
GPT-4o [32]	1.0	0.7	512	8	155	41.10
Mixtral-8x7B [1]	0.90	0.6	256	7	154	43.54
MedGemma-27B [11]	0.90	0.7	256	9	130	36.08
R2Gen [5]	–	–	–	14	46	31.85
Medgemma-4B [12]	–	–	384	17	126	36.02

Table 7: **Data generation hyperparameters.** We report the hyperparameters used to generate synthetic radiology reports across LLMs along with word-level statistics of the outputs. Decoding was performed with temperatures of 0.6–0.7 and top-p values of 0.9-1.0, balancing diversity with clinical consistency. Average word counts are also provided, highlighting differences in verbosity and style across models.

generation, we employ several key hyperparameters, namely *temperature*, *top_p*, and *max_new_tokens* [16].

The *temperature* parameter adjusts the sharpness of the probability distribution over the vocabulary, thereby influencing the degree of randomness in token selection; lower values promote more deterministic outputs, whereas higher values encourage greater diversity. The *top_p* parameter, also referred to as nucleus sampling, restricts token selection to the smallest subset of candidates whose cumulative probability mass exceeds a specified threshold p, balancing quality and diversity in the generated text. Finally, the *max_new_tokens* parameter sets an upper bound on the number of tokens generated, thereby constraining the overall length of each synthetic report. We employ a temperature range of 0.6 to 0.7, which calibrates the LLMs to avoid outputs that are either overly random or excessively deterministic, thereby maintaining both variability and consistency across generations. The values are listed in Table 7.

A.5 Discussion, Limitations, and Future Work

Our novel dataset for chest radiology report generation with large language models (LLMs) and image-to-text models achieves strong lexical performance. The detection pipeline, leveraging style–content disentanglement, yields consistently high MCC scores in the range 92%–100% across both same-and cross-LLM evaluations. Ablation studies show that our proposed BERT–Mamba encoder with two mamba blocks outperforms BiLSTM and Transformer baselines, while combining reconstruction, classification, and mutual information losses achieves the best MCC, underscoring their importance for effective disentanglement. While text-to-text (T2T) systems often match the wording and structure of reference reports, image-to-text (I2T) systems lacks the similar lexical fidelity. The main difficulty is linking visual cues in radiographs to precise language: small or low-contrast findings are easy to miss, and models can struggle with laterality, negation, and uncertainty. As next steps, we plan to fine-tune vision–language models with radiology-specific signals such as section labels and structured findings. We also plan extend the dataset to other categories in radiology and include a broader set of instruction-tuned and vision models.