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 encoder with a single Mamba block, a two-layer Bi-LSTM, a two-layer Transformer, 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 and (ii) a two-layer Transformer encoder, holding all other components and training settings fixed. Relative to Mamba, the BiLSTM variant reduces Accuracy, F_1, 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, $\mathcal{L}_{\text{rec}}$

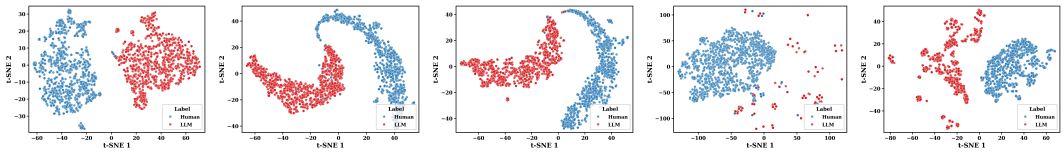
maintains content fidelity and stabilizes training, and $\mathcal{L}_{\text{mi}}$ limits leakage between style and content representations. We therefore adopt the full objective in the main experiments.

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 (blue, red, 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, Mixtral-8x7B, and MedGemma-27B. To ensure variability and control over text generation, we employ several key hyperparameters, namely *temperature*, *top_p*, and *max_new_tokens* [16].

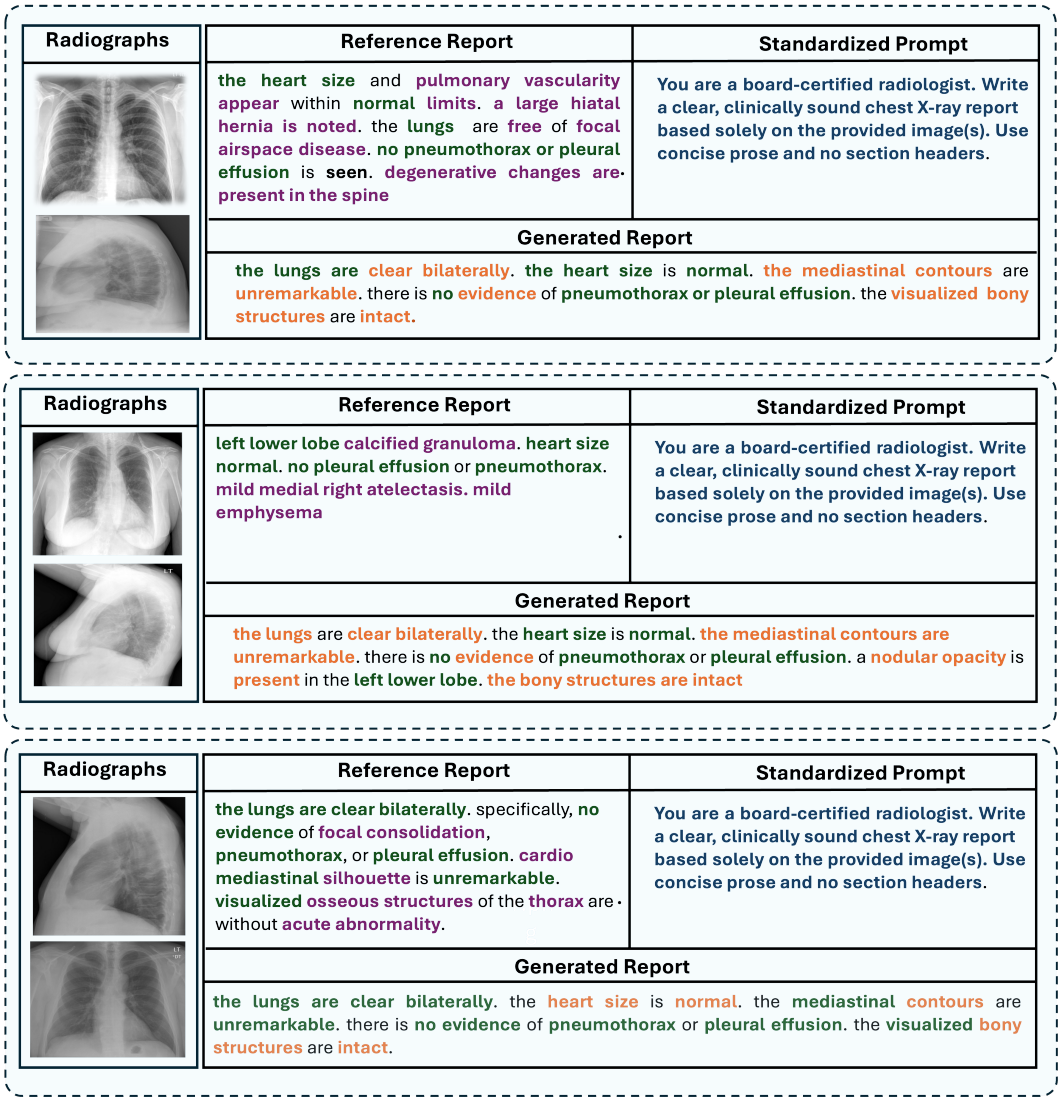


Figure 5: **Dataset samples from Image-to-Text settings.** We showcase three examples from our generated dataset designed using vision-language model (VLM). Each example includes the input image(s), the corresponding ground truth report, and a standardized prompt. Both the reference and generated reports 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	1.0	0.7	512	8	155	41.10
Mixtral-8x7B	0.90	0.6	256	7	154	43.54
MedGemma-27B	0.90	0.7	256	9	130	36.08
R2Gen	–	–	–	14	46	31.85
Medgemma-4B	–	–	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.

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 Bi-LSTM 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.