Clinically Consistent Chest X-ray Report Generation: A Joint Framework Based on Pathological Evidence Constraints and Sequence Optimization

Anonymous ACL submission

Abstract

Chest X-ray report generation aims to translate chest radiographs into clinically coherent diagnostic texts, serving as a crucial technical approach to assist in radiological diagnosis. However, the encoder-decoder architectures adopted by current mainstream methods still face dual limitations. At the encoder, the lack of explicit pathological constraints makes visual features susceptible to statistical priors, leading to semantic hallucinations. At the decoder, word-level cross-entropy optimization tends to induce a normal template bias, weakening the capacity to describe complex lesions. To address these issues, this work proposes a joint framework based on pathological evidence constraints and sequence optimization. Specifically, we design a disease-guided visual representation learning method to enhance the visual encoder's recognition and grounding of critical lesions. Furthermore, we introduce a sequence-level report optimization strategy that models the generation process as sequence decision-making, bridging the gap between word-level training and long-text semantic modeling. Experiments on two mainstream datasets, IU X-Ray and MIMIC-CXR, demonstrate that our method achieves state-of-the-art performance across both natural language generation metrics and clinical efficacy metrics, effectively enhancing the accuracy and reliability of the generated reports. Our code is publicly available at https://anonymous.4open.science/r/CCGen_code-EABACADA

1 Introduction

Chest X-ray (CXR) report generation automatically transforms medical images into clinically compliant diagnostic texts, serving as a vital tool to alleviate radiologists' workload and enhance diagnostic efficiency (Yu et al., 2023). Unlike conventional natural language generation tasks, medical report generation demands both linguistic coherence and exceptionally high standards for factual accuracy and clinical consistency (Sloan et al., 2024; Wang et al., 2025c).

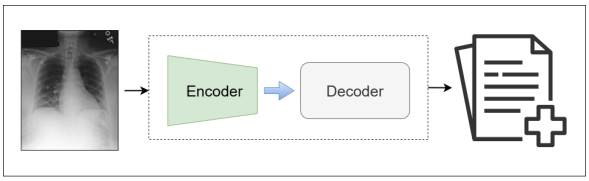

Figure 1: The prevailing encoder-decoder paradigm for CXR report generation.

As illustrated in Figure 1, the prevailing paradigm for automated chest X-ray report generation relies on a standard encoder-decoder architecture, where a visual encoder extracts features from the radiographs and a language decoder subsequently translates these representations into diagnostic texts. Although these architectures have made significant progress in natural language metrics, their underlying paradigm still exhibits dual limitations when confronting complex and variable real-world clinical images (Nicolson et al., 2023). On the one hand, the encoder lacks explicit and effective pathological constraints. Because lesions in CXR images are typically characterized by minute regions, low contrast, and blurred boundaries, existing implicit alignment methods—such as cross-modal contrastive learning or external medical knowledge graphs—fail to fully capture and disentangle crucial features (Qin and Song, 2022; Yin et al., 2025; Mamdouh et al., 2025). Consequently, models remain highly susceptible to background noise or statistical priors dominated by normal samples, thereby inducing semantic hallucinations detached from medical facts at the decoder. On the other hand, mainstream methods heavily rely on word-level cross-entropy loss for autoregressive training at the decoder (Wang et al., 2025a). Since normal or unremarkable descriptions dominate real-world clinical datasets, word-level optimization inevitably triggers a normal template

bias. This bias not only exacerbates exposure bias in long-text generation but also severely impedes the effective modeling of structured semantic flows, substantially weakening the capacity to coherently describe long-range, complex abnormalities.

To address these challenges, we propose a CXR report generation framework based on pathological evidence constraints and sequence optimization, which synergistically regulates the generation process from both pathological visual representation and sequence-level report optimization. At the encoder side, a Disease-Guided Visual Representation Learning (DVRL) method is designed to force the visual encoder to focus on key lesion areas within the feature space, providing reliable and explicit pathological constraints. At the decoder side, we introduce a Sequence-Level Report Optimization (SRO) strategy that transcends the limitations of word-level training by modeling report generation as a global sequential decision-making process. This strategy effectively bridges the gap between local token prediction and long-form semantic modeling, guiding the model to bypass the normal template bias and achieve more accurate and coherent pathological descriptions. The main contributions of this paper are summarized as follows:

- We propose a joint framework based on pathological evidence constraints and sequence optimization, significantly enhancing the semantic consistency and reliability of generated reports.
- We design a Disease-Guided Visual Representation Learning(DVRL) method that leverages auxiliary pathological supervision to strengthen the visual encoder's capability to discern and anchor key pathological features.
- We introduce a Sequence-Level Report Optimization(SRO) strategy that models report generation as a sequential decision-making process, effectively mitigating the issue of high-frequency, templated generation.

2 Related Work

2.1 Radiology Report Generation

Radiology report generation can be viewed as a specialized extension of image captioning in the medical domain (Reale-Nosei et al., 2024). However, compared to general image captioning, radiology report generation exhibits a substantially higher degree of complexity. Early research on radiology report generation mostly adopted the classic architectures of image captioning, utilizing Convolutional Neural Networks (CNNs) to extract visual features from medical images and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to generate diagnostic text word by word (Vinyals et al., 2015). Jing *et al.* (Jing et al., 2018) proposed a generation framework combining multi-task learning with a hierarchical LSTM, which generates topic intents through a high-level network and specific words through a low-level network, thereby preliminarily exploring the structured expression of long medical texts.

Benefiting from the outstanding performance of Transformers in modeling long-range dependencies and cross-modal feature interactions, Transformer-based encoder-decoder architectures have rapidly become the mainstream paradigm in this field, propelling the research focus toward knowledge-driven and cross-modal fusion stages. Since relying solely on visual features makes it difficult to accomplish complex medical semantic reasoning, cross-modal alignment mechanisms, external medical knowledge, and memory-augmented structures have been widely introduced. Chen *et al.* (Chen et al., 2020) proposed a memory-driven Transformer model, R2Gen, which captures potential generation patterns from similar cases via a relational memory module and a conditional layer normalization mechanism. Subsequently, the Cross-Modal Memory Network (CMN) (Chen et al., 2021) was proposed to further share cross-modal memory units, strengthening the semantic associations between imaging regions and medical terminology. Some works have also attempted to integrate medical language priors or construct disease knowledge graphs, leveraging Graph Neural Networks (GNNs) or graph attention mechanisms to enhance the cross-modal feature propagation capability between visual concepts and clinical terms (Zhang et al., 2020; Xu et al., 2023; Yang et al., 2023).

With the development of medical Vision-Language Pre-training (VLP) models and Large Language Models (LLMs), radiology report generation has gradually advanced towards the stages of semantic consistency and fine-grained pathological modeling (Wang et al., 2023). Works such as MedCLIP (Wang et al., 2022) and BioViL (Boecking et al., 2022) construct a unified cross-modal

semantic space through large-scale image-text contrastive learning, substantially boosting the medical visual representation capabilities of models. Building upon this, the research focus has shifted from merely improving language fluency to precisely controlling structured semantic flows and strengthening the perception of pathological regions. Some methods enhance the model's attention to abnormal regions through fine-grained regional alignment, pathological label classification rewards, and medical entity relationship constraints (Wang et al., 2025b; Ding et al., 2026). There are also studies combining reinforcement learning to improve the clinical efficacy of generated reports by directly optimizing clinical evaluation metrics (Miura et al., 2021); however, the majority of these works still rely on implicit alignment.

2.2 Sequence-Level Decision-Making and Optimization in Natural Language Generation

In the field of natural language generation, long-form text generation has long relied on the "teacher forcing" strategy and word-level cross-entropy loss. During the inference phase, this autoregressive training paradigm is highly prone to inducing exposure bias, causing minor single-step prediction errors to continuously accumulate and amplify as the sequence extends (Chiang and Chen, 2021; Nicolson et al., 2024). To overcome the myopic nature of single-step word prediction, the general natural language processing (NLP) domain pioneered the exploration of global sequence decision-making. Self-Critical Sequence Training (SCST), a reinforcement learning-based approach, formulates text generation as a Markov Decision Process (MDP) (Rennie et al., 2017). This empowers the model to transcend local optima and directly optimize non-differentiable global evaluation metrics such as CIDEr or ROUGE. Furthermore, with the rapid advancement of Large Language Model (LLM) alignment technologies, emerging reinforcement learning mechanisms like Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) have further demonstrated that modeling generation tasks as sequence-level preference decisions can significantly enhance a model's performance in complex logical reasoning and long-text planning. Inspired by these advances, we further explore sequence-level optimization for radiology report generation, aiming to improve global semantic coherence and long-form clinical consistency beyond conventional word-level training.

3 Method

As illustrated in Figure 2, we propose a joint framework based on pathological evidence constraints and sequence optimization. Grounded in a standard encoder-decoder topology, the framework extends the standard encoder-decoder paradigm to address the critical pain points of semantic hallucinations and normal template bias inherent in the CXR report generation process.

3.1 Fundamental Architecture

Given an input CXR image X, the objective of the model is to generate the corresponding diagnostic report text $Y = \{y_1, y_2, \ldots, y_t\}$.

At the visual encoding side, the input image X is first fed into a Convolutional Vision Transformer (CvT) (Wu et al., 2021) to extract visual representations, producing a flattened visual patch feature sequence $V \in \mathbb{R}^{N \times d_v}$, where N denotes the number of visual patches and d_v represents the visual feature dimension.

To bridge the modality gap between visual and textual representations, a Multilayer Perceptron (MLP) is employed as a projection function $g(\cdot)$ to map the visual features into the embedding space of the language model, yielding the projected visual prefix features $H_0 \in \mathbb{R}^{N \times d_l}$, where d_l denotes the embedding dimension of the language model. The projection process can be formulated as $H_0 = g(V)$.

At the text generation side, a pre-trained DistilGPT2 (Sanh et al., 2019) is adopted as the core decoder, with an additional multi-head cross-attention module inserted into each decoder layer. During decoding, the projected visual features H_0 are fed into the cross-attention module as conditioning keys and values. Through deep cross-modal interactions, the model predicts the joint conditional probability of the report sequence word by word:

$$p(y_t \mid y_{<t}, X) = f_{\text{dec}}(y_{<t}, H_0) \tag{1}$$

where $y_{<t} = \{y_1, y_2, \ldots, y_{t-1}\}$ denotes the previously generated tokens and $f_{\text{dec}}(\cdot)$ represents the decoding function of the language model.

3.2 Disease-Guided Visual Representation Learning

In the absence of explicit constraints, the visual encoder is highly susceptible to being dominated

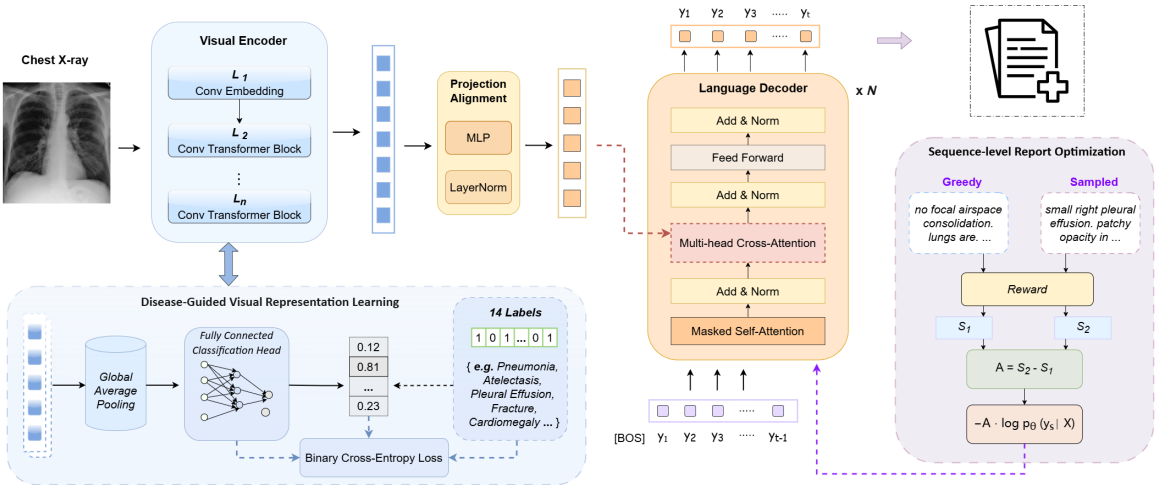

Figure 2: Overall architecture of the proposed joint generation framework based on pathological evidence constraints and sequence-level optimization.

by background noise and statistical data biases. To address this, we design the DVRL method at the encoding side, introducing an auxiliary disease supervision mechanism. Specifically, we append a multi-label pathology prediction branch after the CvT, aiming to classify and predict 14 common chest pathological findings.

Let $l \in \{0,1\}^{14}$ denote the ground-truth pathological label vector corresponding to the input image X, and $\hat{l}$ denote the predicted probability vector output by the model. The multi-label Binary Cross-Entropy (BCE) classification loss for this module is defined as:

$$L_{path} = -\sum_{i=1}^{14} \left[l_i \log(\hat{l}_i) + (1 - l_i) \log(1 - \hat{l}_i) \right] \tag{2}$$

Through the backpropagation of this loss, the visual encoder is explicitly constrained to learn a feature space with strong pathological discriminability. This mechanism not only prompts the model to capture global appearance features but also endows it with a visual grounding capability to focus on critical lesion regions. Such disease-aware representation significantly enhances the model's recognition sensitivity to minute or rare abnormalities, providing reliable factual anchors for the subsequent text generation.

3.3 Sequence-level Report Optimization

In the context of long medical texts, traditional training based on word-level cross-entropy loss not only induces exposure bias but also drives the model to develop a severe normal template bias in pursuit of minimizing single-step prediction loss, thereby weakening its capability to describe complex abnormalities. To address this, we propose the SRO strategy, which models report generation as a sequential decision-making process and introduces a reinforcement learning strategy based on SCST for global optimization.

For the same input image X, at each training step, the decoder generates two output sequences in parallel:

- Greedy baseline sequence $\hat{Y}$: Generated by directly selecting the token with the highest probability at each time step.
- Sampled sequence Y^s: Generated through random sampling based on the model's current output probability distribution.

We utilize CIDEr as the reward function $R(\cdot)$, since it has been widely adopted as a surrogate objective for optimizing global semantic alignment in radiology report generation. The optimization objective function of SCST is defined as follows:

$$L_{SCST} = -(R(Y^s) - R(\hat{Y})) \sum_{t=1}^{T} \log p(y_t^s | y_{<t}^s, X) \tag{3}$$

Driven by the relative reward difference, the model directly aligns with global sequence-level semantic evaluation criteria. When the global matching score of the sampled sequence outperforms the greedy baseline, the probability of its generation trajectory is amplified; otherwise, it is penalized. This mechanism effectively bridges the gap between single-step prediction and the global semantics of

long texts, guiding the model to escape the trap of high-frequency vocabulary.

3.4 Joint Training Objective

To achieve a consistent mapping from visual pathological evidence to global textual expression, the entire framework employs a two-stage, dual-path constraint mechanism for end-to-end optimization:

Warm-up Alignment Stage: This stage utilizes the traditional word-level cross-entropy loss L_{CE} jointly optimized with the pathological visual supervision loss. This enables the network to initially establish cross-modal image-text alignment and fundamental pathological perception capabilities:

$$L_{stage1} = L_{CE} + \lambda L_{path} \tag{4}$$

where λ is a hyperparameter balancing the contribution weights of the two losses.

Sequence Reinforcement Stage: Building upon the convergence of the warm-up model, we transition to a global optimization paradigm. The original word-level cross-entropy loss is replaced by SCST, while the pathological supervision signal is retained to maintain the stability of visual anchoring:

$$L_{stage2} = L_{SCST} + \lambda L_{path} \tag{5}$$

This joint optimization framework establishes a unified optimization mechanism between disease-aware visual representation learning and sequence-level semantic optimization. At the visual side, L_{path} encourages the encoder to learn clinically discriminative disease-aware representations. At the generation side, L_{SCST} directly optimizes global semantic coherence at the sequence level. Their complementary optimization improves the clinical consistency and reliability of the generated reports.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

We evaluate the proposed method on two widely used public CXR report generation benchmark datasets, namely IU X-Ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019). The IU X-Ray dataset contains 7,470 chest X-ray images and 3,955 corresponding radiology reports, with each sample typically comprising both frontal and lateral views. Following previous studies, we adopt the widely used split setting, dividing the dataset into training, validation, and testing sets with a ratio of 7:1:2. MIMIC-CXR is one of the largest publicly available chest radiology datasets, comprising 377,110 chest X-ray images and 227,835 diagnostic reports covering 65,379 patients. We strictly follow the official split scheme, dividing the dataset into 70% for training, 10% for validation, and 20% for testing. The detailed statistics of the datasets are summarized in Table 1.

Dataset	IU X-Ray			MIMIC-CXR		
	Train	Val	Test	Train	Val	Test
Images	5226	748	1496	368960	2991	5159
Reports	2770	395	790	222758	1808	3269
Avg.length	37.6	36.8	33.6	53.0	53.1	66.4

Table 1: Statistics of the IU X-Ray and MIMIC-CXR datasets.

4.1.2 Evaluation Metrics

To comprehensively and objectively evaluate the generation performance of the proposed method, we employ two categories of evaluation metrics: Natural Language Generation (NLG) and Clinical Efficacy (CE) (Irvin et al., 2019). Specifically, NLG metrics focus on assessing the linguistic fluency and lexical overlap of the generated reports, utilizing four mainstream metrics: BLEU-n (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). CE metrics evaluate the clinical relevance and diagnostic accuracy of generated reports by computing Precision, Recall, and F1 score over extracted clinical findings extracted by CheXbert (Smit et al., 2020).

4.1.3 Experimental Details

All experiments are conducted on 4 NVIDIA A100 GPUs. The hyperparameter λ, which controls the contribution of the multi-label disease classification loss, is set to 1.0. Following the warm-up training stage, the Sequence-Level Report Optimization (SRO) strategy is introduced starting from the 15th epoch. The entire framework is optimized using the AdamW optimizer. The initial learning rate for the visual encoder is set to 5×10^{-5}, while the remaining parameters are updated with a learning rate of 1×10^{-4}. An early stopping strategy is adopted with a patience of 15 epochs and a minimum improvement threshold of 1×10^{-4}, where the CIDEr score is used as the monitoring metric. The maximum report length is set to 120 tokens.

During inference, beam search decoding with a beam size of 4 is employed to generate the final radiology reports.

4.2 Main Results

We compare the proposed method with several representative state-of-the-art baselines on both the MIMIC-CXR and IU X-Ray datasets. The relevant experimental results are summarized in Table 2:

- **R2Gen (Chen et al., 2020):** A milestone baseline model in the field of radiology report generation. It proposes a relational memory-driven Transformer architecture that facilitates long-text generation by recording historical states during the decoding phase.
- **R2GenCMN (Chen et al., 2021):** Constructs a shared cross-modal memory network to enhance the alignment between visual features and medical terminology within a common semantic space.
- **DCL (Li et al., 2023):** Addresses the long-tail distribution issue inherent in medical imaging data by introducing dynamic contrastive learning, which pulls positive image-text pairs closer while pushing negative samples apart.
- **R2Gen-Mamba (Sun et al., 2025):** Integrates cutting-edge state space models, leveraging the sequence modeling advantages of the Mamba architecture with linear time complexity. It replaces the standard self-attention mechanism to efficiently capture global dependencies.
- **EMRRG (Zhang et al., 2025):** A recently proposed enhanced generation method that focuses on optimizing deep fusion strategies for multimodal features.
- **ITAdapter (Ding et al., 2026):** Represents the latest trend in parameter-efficient fine-tuning. It utilizes a lightweight Adapter module to bridge the visual encoder and the pre-trained Large Language Model, enabling highly efficient knowledge transfer across modalities.

To further validate the factual accuracy and clinical consistency of the proposed method in real-world medical scenarios, we present the Clinical Efficacy (CE) evaluation results of various methods on the MIMIC-CXR dataset in Table 3. The proposed method achieves the highest scores across all three core clinical metrics, securing a Precision of 0.536, a Recall of 0.479, and an overall F1 score of 0.506.

4.3 Ablation Study

To thoroughly verify the independent effectiveness and synergistic interaction of each core component in our proposed joint framework, we conduct an ablation study on the MIMIC-CXR dataset. As summarized in Table 4, introducing DVRL alone improves all generation metrics, with CIDEr exhibiting a significant improvement from 0.349 to 0.485, which powerfully demonstrates the necessity of implementing explicit pathological constraints. Similarly, introducing SRO in isolation yields distinct performance gains, indicating that SRO can consistently enhance text generation quality even in the absence of enhanced visual features. When both DVRL and SRO are simultaneously integrated into the joint framework, the system achieves the optimal overall performance, with CIDEr further rising to 0.712. This significantly validates the synergistic gains derived from combining pathological evidence constraints with sequence-level optimization.

Specifically, the hyperparameter λ in the DVRL method determines the relative weight of the auxiliary classification loss within the overall objective function. The process of tuning this parameter is essentially finding the optimal equilibrium between local spatial feature retention and global pathological semantic refinement. We conduct a rigorous evaluation across a gradient of values for λ on the MIMIC-CXR dataset, as illustrated in Figure 3.

4.4 Visualization Analysis

To evaluate clinical consistency, we qualitatively compare the generated reports against the ground truth. Figure 4 illustrates two representative examples. We use red text to denote key lesions and clinical findings in the ground truth, while green text indicates correctly generated content consistent with the ground truth. Blue text highlights the complex, highly clinically relevant pathological findings accurately identified by our method, and orange text for hallucinated or templated descriptions driven by statistical priors.

In case 1, the Ground Truth explicitly points out a "Borderline size of the cardiac silhouette." Confronted with this ambiguous abnormality lacking

Table 2: Evaluation results of NLG metrics on MIMIC-CXR and IU X-Ray datasets. Best results are in bold and second-best are underlined.

Dataset	Method	BL-1	BL-2	BL-3	BL-4	METEOR	ROUGE-L	CIDEr
MIMIC-CXR	R2Gen(2020)	0.353	0.218	0.145	0.103	0.142	0.277	-
	R2GenCMN(2021)	0.353	0.218	0.148	0.106	0.142	0.278	-
	DCL(2023)	-	-	-	0.109	0.150	0.284	0.281
	R2Gen-Mamba(2024)	0.352	0.222	0.152	0.110	0.141	0.284	-
	EMRRG(2025)	0.407	0.256	0.175	0.125	0.164	0.288	0.239
	ITAdapter(2026)	0.411	0.260	0.187	0.141	0.152	**0.314**	-
	Ours	**0.416**	**0.273**	**0.200**	**0.159**	**0.167**	0.302	**0.712**
IU X-Ray	R2Gen(2020)	0.470	0.304	0.219	0.165	0.187	0.371	-
	R2GenCMN(2021)	0.475	0.309	0.222	0.170	0.191	0.375	-
	DCL(2023)	-	-	-	0.163	0.193	0.383	0.586
	R2Gen-Mamba(2024)	0.482	0.315	0.228	0.176	0.208	0.382	-
	EMRRG(2025)	0.487	0.325	0.222	0.167	**0.226**	0.385	0.476
	ITAdapter(2026)	**0.536**	**0.377**	**0.274**	**0.206**	0.220	**0.420**	-
	Ours	0.471	0.309	0.227	0.178	0.197	0.372	**0.737**

Table 3: Precision, Recall and F1 scores for clinical efficacy on the MIMIC-CXR dataset. Best results are in bold and second-best are underlined.

Dataset	Method	CE Metrics		
		P	R	F1
MIMIC-CXR	R2Gen	0.333	0.273	0.276
	R2GenCMN	0.334	0.275	0.278
	DCL	0.471	0.352	0.373
	R2Gen-Mamba	0.483	0.325	0.389
	EMRRG	0.421	0.372	0.395
	ITAdapter	0.411	0.459	0.434
	Ours	**0.536**	**0.479**	**0.506**

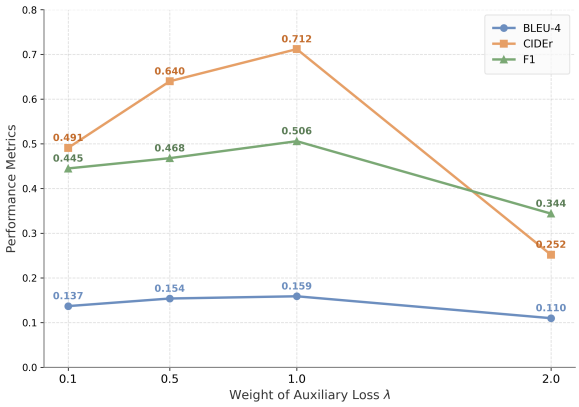

Figure 3: Sensitivity analysis of the classification weight λ. The proposed method achieves the best overall performance when $\lambda = 1.0$.

strong visual features, the baseline models output highly templated and generic "safe" descriptions such as "the lungs are clear. the heart size is normal...". In contrast, our method precisely generates content that is semantically aligned with the Ground Truth. In case 2, compared to the baseline models, our proposed method not only accurately describes the specific distance of the endotracheal tube from the carina but also articulates the complex spatial relationships of the enteric feeding tube passing through the diaphragm and ultimately resting at the proximal stomach in the left upper quadrant. This capability to generate coherent long sentences spanning multiple anatomical regions further demonstrates that our method effectively activates the exploration space of the language decoder, substantially enhancing the richness and spatial logical rigor of long-text sequences.

5 Conclusion

To address the challenges of semantic hallucinations and normal template bias in automated chest X-ray report generation, we propose a joint framework based on pathological evidence constraints and sequence optimization. At the encoding side, we design a Disease-Guided Visual Representation Learning (DVRL) method that introduces explicit auxiliary pathological supervision, effectively enhancing the model's capability to discern and ground complex, minute lesions while mitigating statistical-prior-driven feature shifts at the source.

Table 4: Ablation study results on the MIMIC-CXR dataset. Best results are highlighted in bold.

DVRL	SRO	Metrics						
		BL-1	BL-2	BL-3	BL-4	METEOR	ROUGE-L	CIDEr
×	×	0.385	0.236	0.160	0.117	0.148	0.276	0.349
✓	×	0.405	0.257	0.180	0.136	0.159	0.288	0.485
×	✓	0.397	0.250	0.174	0.130	0.155	0.284	0.426
✓	✓	**0.416**	**0.273**	**0.200**	**0.159**	**0.167**	**0.302**	**0.712**

Image	**Ground Truth**	**Baseline**	**Ours**
	As compared to the previous radiograph, there is no relevant change. **Borderline size of the cardiac silhouette**. No acute process, in particular no pneumonia or pulmonary edema. No pleural effusions. No pneumothorax.	the lungs are clear. the heart size is top normal. the mediastinal contours are normal. there are no pleural effusions. no pneumothorax is seen.	as compared to the previous radiograph there is no relevant change. **borderline size of the ca rdiac silhouette without pulmonary edema**. no pleural effusions. no pneumothorax. no focal parenchymal opacity suggesting pneumonia.
	Endotracheal tube terminates 5 cm above the level of the carina. **Enteric tube courses below the level of the diaphragm, terminating in the left abdomen, in the expected location of the stomach**. No pneumothorax is seen. There is no focal consolidation or pleural effusion. The **cardiac silhouette is top-normal**. The aorta is calcified. No pulmonary edema is seen.	the endotracheal tube ends 4 cm above the carina. the enteric tube ends off the inferior edge of the image. there are no pleural effusions. no pneumothorax is seen. the lungs are clear. the heart size is top normal. the mediastinal contours are normal.	**endotracheal tube terminates approximately 3-8 cm above the level of the carina. enteric tube courses below the diaphragm terminating in the left upper quadrant with side port at the level of the proximal stomach**. there is no pulmonary edema. no new focal consolidation is seen. there is no large pleural effusion or pneumothorax. **the cardiac silhouette is normal**.

Figure 4: Qualitative visualization results of the reports generated by our proposed method and the baseline model.

At the decoding side, we propose a Sequence-Level Report Optimization (SRO) strategy that formulates report generation as a global sequential decision-making process, thereby overcoming the exposure bias inherent in traditional word-level cross-entropy optimization and strengthening the descriptive capacity for long, pathologically complex texts. Extensive experiments on the IU X-Ray and MIMIC-CXR benchmark datasets demonstrate that our joint framework achieves state-of-the-art performance across both natural language generation and clinical efficacy metrics, effectively improving the pathological accuracy and clinical consistency of the generated reports.

Limitations

Although the proposed method demonstrates superior performance, several limitations remain. First, the generated diagnostic reports have not yet been comprehensively evaluated by expert radiologists; future work will involve collaboration with domain experts to further validate the practical value of this approach in real-world clinical workflows. Furthermore, the SRO strategy introduces additional computational overhead due to the parallel generation of multiple sequences and the subsequent reward computation. To address this challenge, we are actively exploring more efficient optimization strategies.

Ethical Considerations

This work utilizes publicly available, de-identified datasets (IU X-Ray and MIMIC-CXR) in strict compliance with their data use agreements. The proposed framework is strictly an assistive tool and must not replace professional medical judgment. Although our method reduces semantic hallucinations, deep learning models may still generate inaccuracies or exhibit demographic biases. Therefore, any real-world clinical deployment necessitates rigorous prospective trials and comprehensive fairness evaluations.

Acknowledgments

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1439–1449.

Ting-Rui Chiang and Yun-Nung Chen. 2021. Relating neural text degeneration to exposure bias. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 228–239.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Shuaipeng Ding, Jianan Shui, Mingyuan Ge, Mengnan Fan, Xin Li, Yijie Zhu, and Mingyong Li. 2026. Itadapter: Image-tag adapter framework with retrieval knowledge enhancer for radiology report generation. *Expert Systems with Applications*, page 131494.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2577–2586.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3334–3343.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Dima Mamdouh, Mariam Attia, Mohamed Osama, Nesma Mohamed, Abdelrahman Lotfy, Tamer Arafa, Essam A Rashed, and Ghada Khoriba. 2025. Advancements in radiology report generation: A comprehensive analysis. *Bioengineering*, 12(7):693.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633.

Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. 2024. e-health csiro at rrg24: entropy-augmented self-critical sequence training for radiology report generation. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 99–104.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Gabriel Reale-Nosei, Elvira Amador-Domínguez, and Emilio Serrano. 2024. From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. *Medical Image Analysis*, 97:103264.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1500–1519.

Yongheng Sun, Yueh Z Lee, Genevieve A Woodard, Hongtu Zhu, Chunfeng Lian, and Mingxia Liu. 2025. R2gen-mamba: A selective state space model for radiology report generation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Pengyu Wang, Shuchang Ye, Usman Naseem, and Jinman Kim. 2025a. Mrg-r1: Reinforcement learning for clinically aligned medical report generation. *arXiv preprint arXiv:2512.16145*.

Xiaochen Wang, Yuan Zhong, Lingwei Zhang, Lisong Dai, Ting Wang, and Fenglong Ma. 2025b. Medmkg: Benchmarking medical knowledge exploitation with multimodal knowledge graph. *arXiv preprint arXiv:2505.17214*.

Xinyi Wang, Grazziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. 2025c. A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*, 103:103627.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31.

Dexuan Xu, Huashi Zhu, Yu Huang, Zhi Jin, Weiping Ding, Hang Li, and Menglong Ran. 2023. Vision-knowledge fusion model for multi-domain medical report generation. *Information Fusion*, 97:101817.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Heng Yin, Shanlin Zhou, Pandong Wang, Zirui Wu, and Yongtao Hao. 2025. KIA: Knowledge-guided implicit vision-language alignment for chest X-ray report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Mingzheng Zhang, Jinfeng Gao, Dan Xu, Jiangrui Yu, Yuhan Qiao, Lan Chen, Jin Tang, and Xiao Wang.

2025. Emrrg: Efficient fine-tuning pre-trained x-ray mamba networks for radiology report generation. *arXiv preprint arXiv:2510.16776*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12910–12917.