RRG-Mamba: Efficient Radiology Report Generation with State Space Model

Xiaodi Hou[1] , **Xiaobo Li**[2] , **Mingyu Lu**[1] , **Simiao Wang**[1] and **Yijia Zhang**[2*]
[1]School of Artificial Intelligence, Dalian Maritime University [2]School of Information Science and Technology, Dalian Maritime University
{houxiaodi, xiaobo.li, lumingyu, wangsimiao}@dlmu.edu.cn, zhangyijia@dlmu.edu.cn

Abstract

Recent advancements in radiology report generation have utilized deep neural networks such as CNNs and Transformers, achieving notable improvements in generating accurate and detailed reports. However, their practical adoption is hindered by the challenge of balancing global dependency modeling with computational efficiency. The state space model, particularly its enhanced variant Mamba, offers promising linear-complexity solutions for long-range dependency modeling. Despite its strengths, Mamba's fixed positional encoding limits its ability to effectively capture complex spatial dependencies. To address this gap, we propose RRG-Mamba, an advanced framework for efficient radiology report generation. Within the RRG-Mamba, we enhance the vanilla Mamba by integrating rotary position encoding (RoPE), enabling dynamic modeling of relative positional information in visual feature sequences. Furthermore, we design a global dependency learning module to optimize long-range visual feature sequence modeling. Extensive experiments on publicly available datasets, including IU X-Ray and MIMIC-CXR, demonstrate that RRG-Mamba achieves a 3.7% improvement in BLEU-4 score over existing models, along with significant gains in computational and memory efficiency. Our code is available at https://github.com/Eleanorhxd/RRG-Mamba.

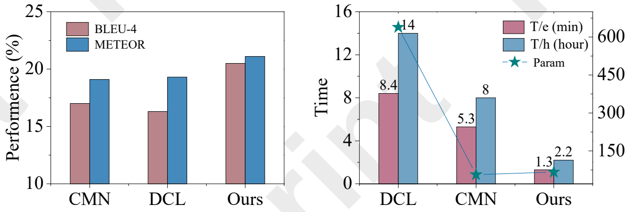
1 Introduction

Automatic radiology report generation (RRG) has emerged as an influential research area in medical imaging, driven by the increasing demand for efficient and accurate diagnostic tools [Hou *et al.*, 2023; Bu *et al.*, 2024]. Traditional radiology report generation relies on radiologists' expertise, making it time-consuming and susceptible to errors such as misdiagnoses or missed findings [Yan and Pei, 2022; Wang *et al.*, 2023]. This challenge has led to growing interest in leveraging deep learning technologies to automate extracting critical information from medical images and generate professional reports, which can significantly enhance the efficiency of medical service delivery, optimize diagnostic accuracy, and alleviate the heavy workload of radiologists [Li *et al.*, 2022b; Tanida *et al.*, 2023; Bu *et al.*, 2024].

Recent advancements in image captioning and encoder-decoder frameworks have driven significant progress in RRG. By leveraging convolutional neural networks (CNNs) for image encoding [Zhang *et al.*, 2020; Huang *et al.*, 2023; Bu *et al.*, 2024] and Transformer models for report decoding [Li *et al.*, 2022a; Li *et al.*, 2023; Wang *et al.*, 2023], these systems [Yan *et al.*, 2021; Wang *et al.*, 2023] aim to provide comprehensive and precise insights for timely and accurate medical diagnosis. This task involves two distinct modalities: visual (image) and textual (report) information. To facilitate cross-modal alignment and semantic integration, several approaches [Chen *et al.*, 2020; Chen *et al.*, 2021; Shen *et al.*, 2024] incorporate a memory matrix mechanism that effectively integrates key features from both modalities, allowing the model to retain critical interactions between the images and the generated text, thereby improving the accuracy and consistency of the reports. Additionally, given the highly technical and medically specialized nature of radiology reports, recent studies [Zhang *et al.*, 2020; Wang *et al.*, 2022a; Yang *et al.*, 2023; Huang *et al.*, 2023] have focused on integrating multi-source medical knowledge, such as posterior-and-prior knowledge [Liu *et al.*, 2021], medical knowledge graphs [Hou *et al.*, 2023], and dynamic knowledge graphs [Li *et al.*, 2023] to better understand medical image content, improving the visual feature representation and enhancing the overall report generation process.

Despite the significant progress in recent RRG methods, they still have several limitations. Traditional methods utilizing pre-trained CNNs [Zhang *et al.*, 2020; Huang *et al.*, 2023; Bu *et al.*, 2024] focus on local feature extraction but struggle to model global context and long-range dependencies due to their limited receptive fields [Zhu *et al.*, 2024; Liu *et al.*, 2024]. Vision Transformers (ViT) address the above limitations by employing self-attention mechanisms to capture global dependencies [Li *et al.*, 2022a; Wang *et al.*, 2022b; Li *et al.*, 2023]. However, this comes at the cost of increased computational resources and storage requirements, particularly on large-scale datasets. The self-attention mechanism requires computing correlations between all positions in the input sequence, leading to a quadratic growth in parameter

[*] Corresponding author.



(a) Accuracy Comparison (b) Time and Parameter Comparison

Figure 1: Accuracy and effiency comparison of RRG-Mamba with different RRG models (i.e. R2GenCMN (CMN), DCL) on the IU X-Ray dataset: (a) accuracy comparision, (b) time and parameter complexity comparision.

count with sequence length. Consequently, the trade-off between global context modeling and computational efficiency in the RRG task remains an unresolved challenge.

To fill the gap, we propose RRG-Mamba, a novel framework that leverages Mamba's capability for global information modeling in radiology report generation. Furthermore, RRG-Mamba introduces rotary position encoding (RoPE) to optimize the vanilla Mamba, improving its flexibility and efficiency in capturing relative positional information and long-range dependencies from the visual feature sequences. Building on this, we design a global dependency learning module that dynamically captures relative positional information while effectively modeling complex spatial relationships. Finally, comprehensive experiments have demonstrated the effectiveness of RRG-Mamba. As shown in Figure 1, this approach not only alleviates the model's computational overhead but also preserves its accuracy, achieving a superior trade-off between efficiency and performance.

The main contributions in the paper are as follows:

- We identify a key challenge in radiology report generation: the inherent trade-off between effectively capturing global dependencies and maintaining computational efficiency. To our knowledge, this work marks the pioneering attempt to apply Mamba for addressing this critical issue within the task.
- We propose RRG-Mamba, a novel framework for efficient visual representation. Additionally, RRG-Mamba designs a global dependency learning module that integrates rotary position encoding, enhancing the vanilla Mamba for effectively modeling of long-sequence visual feature dependencies.
- We conduct comprehensive experiments across two public datasets, meticulously evaluating the performance of our proposed RRG-Mamba. The results demonstrate RRG-Mamba's superior efficacy over multiple baselines and establish new performance benchmarks, while also improving computational and memory efficiency.

2 Related Work

2.1 Radiology Report Generation

RRG is a critical task in medical artificial intelligence, which aims to automatically generate descriptive and clinically relevant reports from medical images. In recent years, researchers [Huang *et al.*, 2023; Xue *et al.*, 2024; Shen *et al.*, 2024; Hou *et al.*, 2025] have made significant progress in RRG. [Chen *et al.*, 2021] used a shared memory matrix in the encoder to fully explore the association between medical images and texts, thereby promoting the interaction between cross-modal information. [Li *et al.*, 2023] proposed a dynamic knowledge graph-enhanced model for radiology report generation that integrates medical knowledge to improve visual feature representations and optimizes dynamic graph retrieval through contrastive learning, thereby enhancing report accuracy. [Wang *et al.*, 2023] employed a Vision Transformer as an encoder to extract visual features. They introduced multiple learnable "expert" tokens in both the encoder and decoder to interact with visual tokens, thereby enhancing the model's attention to fine-grained lesion areas. Although these methods have made significant progress in RRG, they still face limitations in extracting fine-grained pathological features and enhancing the representation of key lesion areas.

2.2 State Space Models

Recently, state space models (SSMs) [Hui *et al.*, 2019; Gu *et al.*, 2021; Gu *et al.*, 2022] have received widespread attention due to their potential in sequence modeling. In particular, the enhanced Mamba model, leveraging the SSM, significantly accelerates inference speed while effectively modeling long-range dependencies in sequential data through a hardware-aware parallelization strategy [Gu and Dao, 2023]. Inspired by Mamba, multiple models have demonstrated remarkable advantages across diverse applications. For example, [Liu *et al.*, 2024] proposed a state-space model with a global receptive field, incorporating multi-directional scanning and hierarchical networks to comprehensively capture information at every position within the input sequence. [Yue and Li, 2024] utilized grouped convolutions and channel shuffling to achieve efficient and generalized medical image classification while significantly reducing computational overhead. Similarly, [Zhu *et al.*, 2024] employed a bidirectional state-space model to effectively capture global information, complemented by a position embedding module for local semantic feature perception. The successful application of SSM to complex sequence modeling tasks underscores their efficacy and offers valuable insights that pave the way for further advancements in medical image analysis.

3 Preliminary

State Space Models. The SSM-based models, e.g., Mamba, map the one-dimensional input sequence $x(t){\in}\mathbb{R}$ to the output sequence $y(t){\in}\mathbb{R}$, capturing their relationship through a hidden state $h(t){\in}\mathbb{R}^N$ for sequence modeling and prediction. The calculation process is as follows:

$$\begin{aligned} h^{'}(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t), \end{aligned} \tag{1}$$

where $A{\in}\mathbb{R}^{N\times N}$ is evolution parameter, $B{\in}\mathbb{R}^{N\times 1}$ and $C{\in}\mathbb{R}^{N\times 1}$ refer to the projection parameters.

To adapt SSM for deep learning, it is necessary to convert it from a continuous-time model to a discrete-time model by

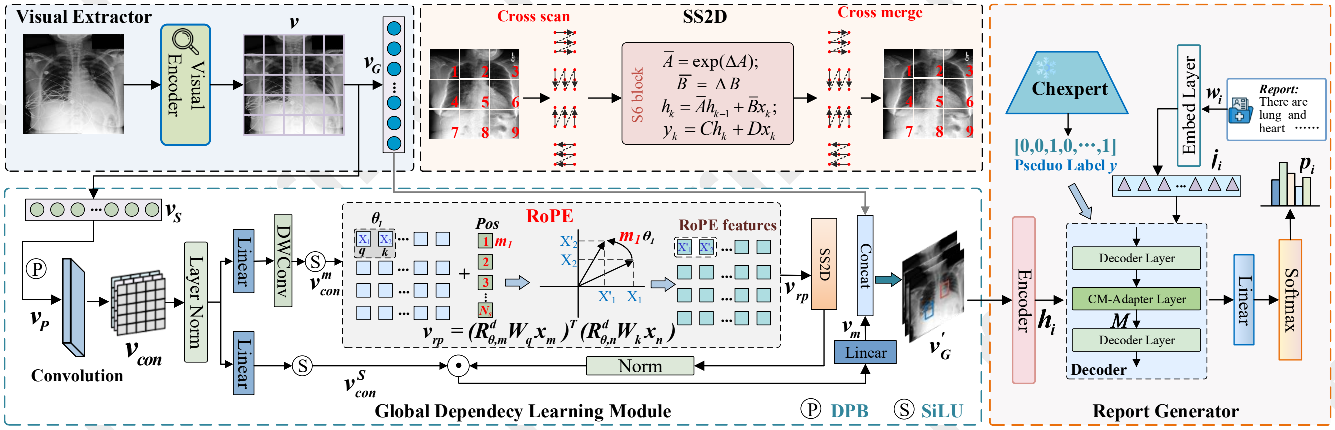


Figure 2: RRG-Mamba overall architecture, including visual extractor, global dependency learning module (GDLM) and report generator. GDLM combines rotary position encoding (RoPE) with 2D selective scan (SS2D) layer to capture relative position information and model long-distance dependencies. The CM-adapter is cross-modal adapter and $\odot$ is element-wise multiplication. The DPB is dynamic position bia and SiLU is an activation function.

introducing a time scale parameter Δ and applying the zero-order hold (ZOH) discretization rule, as follows:

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}), \tag{2}$$

$$\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}, \tag{3}$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ denote the discrete parameters.

After discretization, the Eq.(1) is written as follows:

$$\begin{aligned} h_k &= \bar{\mathbf{A}}h_{k-1} + \bar{\mathbf{B}}x_k, \\ y_k &= \mathbf{C}h_k. \end{aligned} \tag{4}$$

Finally, the SSM model utilizes a global convolution to compute the output:

$$\begin{aligned} \bar{K} &= (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, ..., \mathbf{C}\bar{\mathbf{A}}^{\mathcal{L}-1}\bar{\mathbf{B}}), \\ y &= x * \bar{K}, \end{aligned} \tag{5}$$

where $\bar{K} \in \mathbb{R}^{\mathcal{L}}$ and $*$ denote a structured convolutional kernel and the convolution operation, respectively. $\mathcal{L}$ represents the length of the input sequence $x(t)$.

Radiology Report Generation. RRG involves automatically generating descriptive and accurate reports $R = \{w_1, w_2, ..., w_{N_R}\}$ based on medical imaging $I = \{i_1, i_2, ..., i_{N_I}\}$, where N_R and N_I are the number of tokens in reports and the number of images, respectively.

4 Method

This section details RRG-Mamba, comprising the visual extractor, global dependency learning module, and report generator, as illustrated in Figure 2.

4.1 Visual Extractor

For each image i, the corresponding visual features $v \in \mathbb{R}^{H \times W \times C} = \text{Vencoder}(i)$ are extracted through the visual encoder, where H and W represent the spatial dimensions of the feature map, and C is the number of channels. Vencoder$(\cdot)$ denotes the visual encoder.

Subsequently, the visual feature v generates two distinct visual representations: the global visual feature representation $v_G \in \mathbb{R}^{HW \times C}$ and the serialized token feature representation $v_S \in \mathbb{R}^{HW \times C}$. The detailed computation procedure is outlined as follows:

$$\begin{aligned} v_G &= \text{AvgPool}(v), \\ v_S &= \text{LN}(\text{Proj}(\text{Flat}(v))), \end{aligned} \tag{6}$$

where 'AvgPool', 'LN', 'Proj' and 'Flat' represent global average pooling, layer normalization, projection and flattening operations, respectively.

4.2 Global Dependecy Learning Module

Due to the limitations of CNNs' local receptive field, we design the global dependecy learning module (GDLM), leveraging the SSM backbone to effectively model long-range dependencies within medical image sequences and capture critical lesion features. Additionally, we integrate rotary position encoding (RoPE) [Su *et al.*, 2023] to enhance the vanilla Mamba, improving its efficiency and flexibility in capturing relative positional information from visual feature sequences.

Specifically, inspired by dynamic position bias (DPB) [Chu *et al.*, 2023], we first use DPB to adjust the spatial information of the serialized tokens v_S, thereby enhancing RRG-Mamba's sensitivity to the spatial regions of visual features and capturing spatial structural information. The calculation process is as follows:

$$v_P = v_S + \text{DPB}(v_S), \tag{7}$$

where $v_P \in \mathbb{R}^{HW \times C}$ is the visual feature sequence adjusted by DPB.

Then, to enhance the correlation between various positions in medical images, we input the feature v_P adjusted by DPB into the convolution operation Conv$(\cdot)$ to extract richer local spatial features and strengthen global dependency modeling:

$$v_{con} = \text{Conv}(v_P), \tag{8}$$

where v_{con} is the feature representation obtained after the convolution operation.

To further capture global long-term dependencies and contextual relationships within the visual feature sequence, we enhance Mamba by integrating rotary positional encoding (RoPE), which improves the model's ability to preserve spatial positions and rotational invariance. The specific calculation process is as follows:

$$\begin{aligned} v^m_{con} &= \mathrm{SiLU}(\mathrm{DWConv}(\mathrm{Linear}(\mathrm{LN}(v_{con})))), \\ v^s_{con} &= \mathrm{SiLU}(\mathrm{Linear}(\mathrm{LN}(v_{con}))), \end{aligned} \tag{9}$$

$$v_{rp} = (R^d_{\Theta,m} W_q v^m_{con,m})^\top (R^d_{\Theta,n} W_k v^m_{con,n}), \tag{10}$$

$$v_M = \mathrm{Linear}(v^s_{con} \odot \mathrm{Norm}(\mathrm{SS2D}(v_{rp}))), \tag{11}$$

where v^m_{con} and v^s_{con} represent the intermediate features obtained by projecting to the hidden space. $v^m_{con,m}$, $v^m_{con,n}$ are the m-th and n-th visual embedding vectors in the input sequence v^m_{con}, respectively. $R^d_{\Theta,m}$ and $R^d_{\Theta,n}$ are orthogonal matrices. W_q and W_k are weight matrices. v_{rp} is the feature sequence obtained by the RoPE, and v_M represents the output of the improved Mamba. 'SiLU', 'DWconv', 'Linear', 'Norm'and 'SS2D' represent activation function, depthwith convolution, linear layer, normalization layer and 2D selective scan, respectively. The $\odot$ is element-wise multiplication.

Finally, we concatenate v_G and v_M to obtain the output of the GDLM:

$$v^{'}_G = \mathrm{Concat}(v_G, v_M), \tag{12}$$

where $v^{'}_G \in \mathbb{R}^{HW \times C}$ is the global visual features obtained by GDLM and 'Concat' denotes the concatenate operation.

In Eq.(11), SS2D is a pivotal component of the Mamba framework, with its structure depicted in Figure 2. In this process, visual features are initially partitioned into non-overlapping patches, then scanned along four distinct paths, generating four independent sequences. Each sequence is subsequently processed by the selective scan space state sequential model (S6) [Gu and Dao, 2023], which extracts spatial information from various directions while effectively preserving critical contextual features. Finally, the four sequences are merged to form the consolidated 2D visual feature representation. This multi-path scanning and selective spatial processing enable RRG-Mamba to effectively capture richer spatial dependencies, which are essential for accurately modeling the complex visual patterns in medical images.

4.3 Report Generator

Cross-modal Adapter. We introduce cross-modal adapter to enhance the interaction and fusion of features across modalities. Specifically, inspired by [Wang *et al.*, 2024], we employ Chexpert [Irvin *et al.*, 2019] to generate pseudo labels y for each visual feature $v^{'}_G$, automatically annotating the presence of 14 prevalent diseases in medical images. This process strengthens the semantic alignment between visual and textual features, improving cross-modal consistency. The label generation is formalized as:

$$\{y_1, y_2, \ldots, y_{N_C}\} = \Phi(\mathrm{softmax}(W_c \cdot v^{'}_G)), \tag{13}$$

where N_C is the number of disease categories and W_c is weight parameter. $\Phi(\cdot)$ is the label generation function.

Then, we apply a projection layer to projects the visual feature sequence v^l_G fused with the pseudo labels and the hidden state $\mathcal{H}$ from the decoder layer into a shared feature space, obtaining the cross-modal feature $\mathcal{M}$, as follows:

$$\mathcal{M} = \mathrm{Proj}(v^l_G, \mathcal{H}). \tag{14}$$

Encoder-Decoder. We employ a Transformer-based model to generate radiology report. Specifically, at time step T, the encoder maps the visual features $v^{'}_G$ into intermediate feature representations $h_i \in \mathbb{R}^{1 \times d}$. Next, we use the embedding layer to obtain the word embedding $j_i \in \mathbb{R}^{1 \times d}$ of each word token w_i in the report R. d denotes the dimensionality of the hidden states. Then, the decoder generates the output of the current time step. In this cycle, a complete radiology report is gradually generated. The process is as follows:

$$\{h_1, h_2, ..., h_{N_S}\} = \mathrm{Encoder}(v^{'}_G), \tag{15}$$

$$\{j_1, j_2, ..., j_{T-1}\} = \mathrm{Embed}(w_1, w_2, ..., w_{T-1}), \tag{16}$$

$$p_T = \mathrm{Decoder}(h_1, h_2, ..., h_{N_S}; \mathcal{M}; j_1, j_2, ..., j_{T-1}), \tag{17}$$

where w_i and p_T are the i-th word in the ground truth and the predicted word at time step T, respectively. N_S denotes the number of the intermediate features. The 'Embed' is embedding layer.

4.4 Objective Function

The cross-entropy loss function $\mathcal{L}_{CE}$ is utilized to measure the divergence between the predicted report $\{p_i\}_{i=1}^{N_R}$ and the corresponding ground truth $\{w_i\}_{i=1}^{N_R}$, thereby enhancing the model's ability to predict accurately, as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N_R} \sum_{i=1}^{N_R} w_i \cdot log(p_i). \tag{18}$$

5 Experiments and Analysis

5.1 Experiment Settings

Datasets. We evaluate our model on two publicly available RRG datasets: IU X-Ray [Shin *et al.*, 2016] and MIMIC-CXR [Johnson *et al.*, 2019]. IU X-Ray is a public radiology report generation dataset released by Indiana University. MIMIC-CXR is a large-scale chest X-ray dataset, which is widely used in tasks such as medical image processing.

Metrics. We use natural language generation (NLG) metrics to evaluate the quality of generated medical reports, including BLEU [Papineni *et al.*, 2002], METEOR [Denkowski and Lavie, 2011], and ROUGE-L [Lin, 2004]. These metrics can measure the fluency and accuracy of the generated report and evaluate its similarity with the reference report, thereby effectively reflecting the generated report's language quality and content consistency. To assess the clinical utility of generated reports, we employ clinical efficacy (CE) metrics, including precision, recall, and F1-score. The evaluation focuses on disease-specific keywords derived from radiology reports, where we convert unstructured radiologist narratives into 14 structured labels.

Implementation Details. Following [Yan and Pei, 2022; Shen *et al.*, 2024], we adopt the pre-trained DenseNet-121 [Huang *et al.*, 2017] as the visual encoder. We provide several other variants of visual encoders, including ResNet-101

Type	**Models**	**IU X-Ray**						**MIMIC-CXR**					
		BL-1	BL-2	BL-3	BL-4	M	R	BL-1	BL-2	BL-3	BL-4	M	R
ResNet	R2Gen	0.470	0.304	0.219	0.165	0.187	0.371	0.353	0.218	0.145	0.103	0.142	0.277
	R2GenCMN	0.475	0.309	0.222	0.170	0.191	0.375	0.353	0.218	0.148	0.106	0.142	0.278
	PromptMRG	0.401	-	-	0.098	0.160	0.281	0.398	-	-	0.112	**0.157**	0.268
	R2Gen-Mamba	0.482	0.315	0.228	0.176	0.208	0.382	0.352	0.222	0.152	0.110	0.141	0.284
DenseNet	PPKED	0.483	0.315	0.224	0.168	-	0.376	0.360	0.224	0.149	0.106	0.149	0.284
	MAN	0.501	0.328	0.230	0.170	0.213	0.386	0.396	0.244	0.162	0.115	0.151	0.274
	Clinical-BERT	0.495	0.330	0.231	0.170	0.209	0.376	0.383	0.230	0.151	0.106	0.144	0.270
ViT	BLIP	0.492	0.314	0.222	0.169	0.193	0.381	0.350	0.219	0.152	0.109	0.151	0.283
	DCL	-	-	-	0.163	0.193	0.383	-	-	-	0.109	0.150	0.284
	METransformer	0.483	0.322	0.228	0.172	0.192	0.380	0.386	0.250	**0.169**	**0.124**	0.152	0.291
Ours	ViT-B/32	0.506	0.322	0.228	0.165	0.207	0.371	0.387	0.234	0.149	0.104	0.145	0.277
	ViT-B/16	0.505	0.327	0.237	0.179	0.194	0.395	0.389	0.239	0.150	0.110	0.144	0.284
	ResNet-101	0.507	0.334	0.248	0.192	0.198	0.401	0.392	0.245	0.156	0.113	0.146	0.285
	DenseNet-121	**0.528**	**0.368**	**0.271**	**0.207**	**0.215**	**0.408**	**0.406**	**0.253**	**0.169**	0.121	0.154	**0.293**

Table 1: Comparing the performance of our proposed RRG-Mamba with other competitive models on the publicly available IU X-Ray and MIMIC-CXR datasets, with the best performing scores highlighted in bold. BL, M, and R refer BLEU, METEOR, and ROUGE-L, respectively, and ViT refers to Vision Transformer.

Models	Precision	Recall	F1
R2Gen	0.333	0.273	0.276
R2GenCMN	0.334	0.275	0.278
METansformer	0.364	0.309	0.311
DCL	0.471	0.352	0.373
MAN	0.411	0.398	0.389
Ours	**0.498**	**0.453**	**0.475**

Table 2: Clinical efficacy metrics comparison of RRG-Mamba and other models on the MIMIC-CXR dataset.

RRG-Mamba	#Param(MB)	T/e(*min*)	T(*h*)	B/e
ViT-B/32	134.33	2.29	3.82	35
ViT-B/16	126.43	2.53	4.21	42
ResNet-101	176.76	2.04	3.40	18
DenseNet-121	66.56	1.33	2.22	16

Table 3: Analysis of RRG-Mamba on different visual encoder complexity on IU X-Ray. The #Param, T/e, T and B/e represent the number of training parameters, the time for one training epoch, the total training time and the number of epochs with the best result.

[He *et al.*, 2016], ViT (ViT/B-16 and ViT/B-32) [Dosovitskiy, 2020] to further explore the performance of different visual encoders in RRG tasks. We configure the word identification ratio as k=0.5, thereby controlling the proportion of significant words. According to [Gu *et al.*, 2022], we design three versions of GDLM with different structures (tiny, samll and base) to explore the impact of model capacity on RRG-Mamba.

5.2 Main Experiment

Overall Performance. We evaluate the performance of our proposed model by comparing it with state-of-the-art (SOTA) methods on both datasets. These methods are categorized based on their visual encoders: ResNet-101 (**R2Gen** [Chen *et al.*, 2020], **R2GenCMN** [Chen *et al.*, 2021], **PromptMRG** [Jin *et al.*, 2024], **R2Gen-Mamba** [Sun *et al.*, 2025]), DenseNet-121 (**PPKED** [Liu *et al.*, 2021], **Clinical-BERT** [Yan and Pei, 2022], **MAN** [Shen *et al.*, 2024]), and Vision Transformer (ViT) (**BLIP** [Li *et al.*, 2022a], **DCL** [Li *et al.*, 2023], **METransformer** [Wang *et al.*, 2023]).

Table 1 summarizes the experimental results of our model on both datasets. The findings indicate that our model consistently outperforms most SOTA methods for the radiology report generation task across various evaluation metrics. Specifically, on the IU X-Ray dataset, compared to the MAN model (DenseNet-121), RRG-Mamba achieves notable improvements in BLEU{1-4} scores by 2.7%, 4.0%, 4.1%, and 3.7%, respectively, as well as enhancements in METEOR and ROUGE-L scores by 0.2% and 2.2%, respectively. Similarly, on the MIMIC-CXR dataset, RRG-Mamba (DenseNet-121) demonstrates superior performance, further validating its efficacy. Additionally, R2Gen-Mamba directly combines the vanilla Mamba with ResNet-101 to improve training and inference efficiency. In contrast, our proposed RRG-Mamba extends Mamba further by introducing RoPE to enhance relative position representation and incorporating a DPB to strengthen the model's ability to capture spatial positional information. These technical advancements lead to significant performance gains, with our method outperforming R2Gen-Mamba by 4.6% and 5.4% in BLEU-1 scores on two benchmark datasets. Next, we analyze the superiority of our proposed method from the following three perspectives.

Analysis on Visual Encoders. Table 1 shows the experimental results of our model RRG-Mamba using different visual encoders (ViT-B/32, ViT-B/16, ResNet-101, and DenseNet-121) on both datasets. Table 1 presents RRG-

Dataset	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-Ray	w/o CMA	0.502	0.340	0.248	0.188	0.207	0.404
	w/o GDLM	0.491	0.328	0.236	0.182	0.193	0.395
	w/o RoPE	0.496	0.325	0.241	0.180	0.195	0.369
	w/o CMA+GDLM	0.451	0.289	0.209	0.159	0.175	0.365
	RRG-Mamba	**0.528**	**0.368**	**0.271**	**0.207**	**0.215**	**0.408**
MIMIC -CXR	w/o CMA	0.386	0.241	0.163	0.115	0.148	0.281
	w/o GDLM	0.375	0.226	0.155	0.108	0.142	0.275
	w/o RoPE	0.382	0.235	0.153	0.105	0.138	0.278
	w/o CMA+GDLM	0.324	0.203	0.138	0.100	0.135	0.276
	RRG-Mamba	**0.406**	**0.253**	**0.169**	**0.121**	**0.154**	**0.293**

Table 4: Ablation study results of RRG-Mamba on the IU X-Ray and MIMIC-CXR datasets.

Model	BL-1	BL-4	METEOR	ROUGE-L
GDLM-T	0.496	0.182	0.205	0.385
GDLM-S	0.516	0.198	0.208	0.396
GDLM-B	**0.528**	**0.207**	**0.215**	**0.408**

Table 5: Results of different versions of global dependecy learning module (GDLM) on IU X-Ray. The T, S, B denotes Tiny, Small, Base, respectively.

Mamba outperforms the Vision Transformer when using pretrained CNNs, particularly with DenseNet-121. This suggests that RRG-Mamba effectively extracts local pathological features from medical images through multi-level convolutional operations and captures global context information via global dependency learning module. Consequently, RRG-Mamba acquires rich semantic representations that significantly enhance the discriminative power of medical image features.

Analysis on Clinical Efficacy Metrics. We evaluate the clinical accuracy and effectiveness of the generated medical reports using the CE metrics. Table 2 compares the performance of RRG-Mamba with existing models on the MIMIC-CXR dataset. The results show that RRG-Mamba achieves notable improvements of 8.7%, 5.5%, and 8.6% in precision, recall, and F1-score, respectively, compared to the MAN. These enhancements may be attributed to RRG-Mamba's ability to effectively capture fine-grained lesion features, resulting in reports with higher clinical accuracy and relevance.

Analysis of Model Complexity. Table 3 presents the model complexity experimental results of RRG-Mamba using different visual encoders on the IU X-Ray. A comparative analysis of the performance across various encoders reveals that the pre-trained CNN-based encoder offers superior reasoning capabilities and training efficiency compared to the ViT-based encoder. Specifically, the CNN-based encoder requires fewer training parameters, considerably reduces training time, and shows faster convergence during training.

Although the ViT excels in global context modeling, the SSM-based design of RRG-Mamba effectively mitigates the computational challenges inherent in ViT. Our approach not only efficiently captures long-range dependencies but also significantly enhances inference efficiency and accelerates training speed. This highlights the advantage of combining global dependency learning with optimized visual extraction, leading to both improved performance and reduced computational overhead.

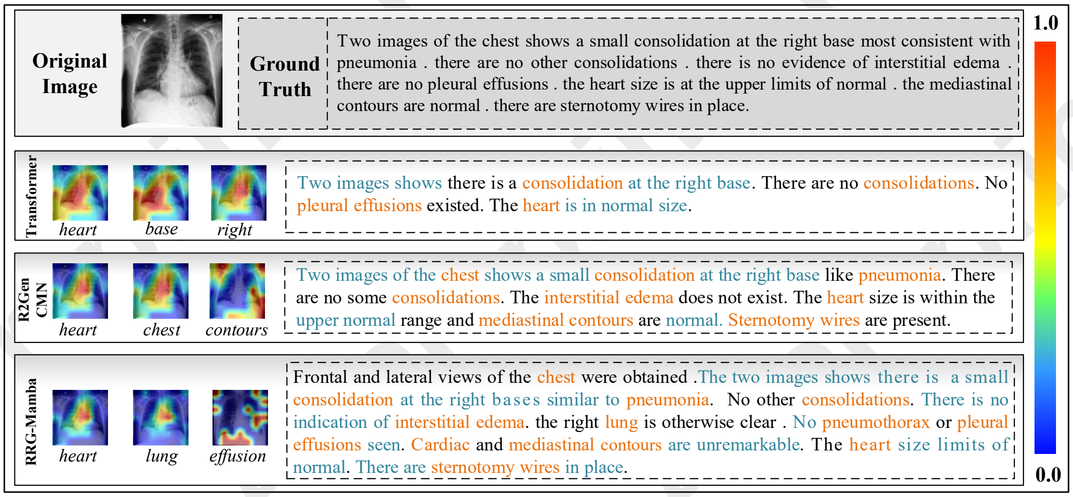
5.3 Ablation Study

We perform an ablation study to validate the effectiveness of RRG-Mamba's core components on both datasets. Four variants are tested: w/o GDLM, which removes the global dependency learning module (GDLM) and omits long-range visual feature modeling; w/o RoPE, which excludes rotary position encoding within the GDLM, thereby neglecting relative positional information; w/o CMA, which eliminates the cross-modal adapter and adopts a simplistic approach for cross-modal feature fusion; and w/o CMA+GDLM, which ablates both the GDLM and CMA, relying solely on the model's basic structure for report generation.

Table 4 presents the results of the ablation study. The removal of the GDLM leads to a marked performance drop, underscoring its role in capturing long-range dependencies in visual features. Excluding RoPE diminishes performance, highlighting the importance of modeling relative positional information. The ablation of the CMA results in a significant decline, emphasizing the critical role of effective cross-modal interaction in improving overall performance. Finally, the simultaneous removal of both modules yields the lowest performance, further reinforcing the essential contributions of these two core components.

Analysis of Different Versions of Global Dependency Learning Module. Table 5 shows the experimental results of different versions of the GDLM (tiny, small and base) with varying model capacities on the IU X-Ray. Among these, GDLM-B achieves the best performance, with BLEU-1, BLEU-4, METEOR, and ROUGE-L scores of 0.528, 0.207, 0.215, and 0.408, respectively.

These results underscore the critical role of model capacity in effective global dependency learning for medical images. GDLM-B, with its more sophisticated network architecture, is better equipped to model long-range dependencies, capturing both complex global and local features within medical images. These findings demonstrate that increasing model capacity can significantly improve the performance of global context modeling, enabling precise and comprehensive analysis of medical images.



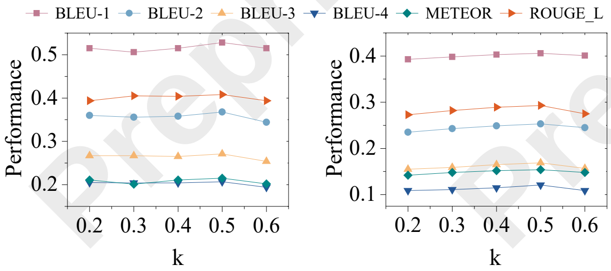
Two images of the chest shows a small consolidation at the right base most consistent with pneumonia . there are no other consolidations . there is no evidence of interstitial edema . there are no pleural effusions . the heart size is at the upper limits of normal . the mediastinal contours are normal . there are sternotomy wires in place.

Two images shows there is a consolidation at the right base. There are no consolidations. No pleural effusions existed. The heart is in normal size.

Two images of the chest shows a small consolidation at the right base like pneumonia. There are no some consolidations. The interstitial edema does not exist. The heart size is within the upper normal range and mediastinal contours are normal. Sternotomy wires are present.

Frontal and lateral views of the chest were obtained .The two images shows there is a small consolidation at the right bases similar to pneumonia. No other consolidations. There is no indication of interstitial edema. the right lung is otherwise clear . No pneumothorax or pleural effusions seen. Cardiac and mediastinal contours are unremarkable. The heart size limits of normal. There are sternotomy wires in place.

Figure 3: Visualize the reports and attention heatmaps generated by RRG-Mamba and different models (Transformer, R2GenCMN) on MIMIC-CXR. The orange font represents the organ or related disease, the blue denotes the semantic description similar to the ground truth.



(a) IU X-Ray

(b) MIMIC-CXR

Figure 4: Performance for varying hyperparameter k on IU X-Ray and MIMIC-CXR datasets.

5.4 Hyperparameter Study

We analyze the impact of the hyperparameter k (representing the word identification ratio) on RRG-Mamba's performance across both datasets. k quantifies the proportion of key terms identified in the generated text, optimizing the model's ability to effectively capture critical terms and control information redundancy. As depicted in Figure 4, RRG-Mamba achieves peak performance at k=0.5. However, increasing k beyond this value leads to performance degradation, likely due to the inclusion of redundant high-frequency terms (e.g., "the" "there"), which introduce noise and disrupt the cross-modal integration process, thereby impairing the accuracy of the generated reports. Conversely, excessively low k values may omit critical medical terms (e.g., "heart"), undermining the comprehensiveness and clinical relevance of the generated reports. Thus, the selection of k is crucial for optimizing report quality and enhance model performance.

6 Case Study

To explore the efficacy of RRG-Mamba, we randomly select a case from the MIMIC-CXR for detailed analysis. Figure 3 shows the medical reports and image-to-text attention heatmaps generated by the RRG-Mamba model and other comparison models (Transformer and R2GenCMN). We use orange fonts to mark organs or diseases and blue fonts to mark semantic descriptions similar to ground truth. It can be intuitively observed from Figure 3, the radiology reports generated by the RRG-Mamba can accurately identify the key feature areas in medical images and generate descriptions that contain professional terms and are in line with clinical reality.

Specifically, compared with the Transformer model, the RRG-Mamba model can capture subtle lesions (such as *"a small consolidation"* and *"interstitial edema"*) and generate more accurate semantic descriptions. In contrast, the Transformer model has a weaker recognition ability for these lesions, and the generated reports are more general and lack detailed support. Compared with the RRG-Mamba, the R2GenCMN model lacks detailed descriptions and semantic richness, making it less effective in comprehensively analyzing medical images. Differently, RRG-Mamba model generates more comprehensive results, which not only identifies specific lesions (*"a small consolidation"*) but also captures subtle findings in medical images and exclude other potential pathological features (*"the right lung is otherwise clear. No pneumothorax or pleural effusions seen"*).

7 Conclusion

We propose RRG-Mamba, a novel framework that leverages Mamba's capability for global information modeling in radiology report generation. Additionally, RRG-Mamba designs a global dependency learning module that integrates rotary position encoding, enhancing the vanilla Mamba for effective modeling of long-sequence visual feature dependencies. At last, we conduct extensive experiments on two publicly available datasets, demonstrating RRG-Mamba's superior effectiveness compared to representative baselines and establishing new performance benchmarks. Furthermore, RRG-Mamba exhibits significant computational and memory efficiency advantages over prevailing neural network architectures, such as CNNs and Transformers.

Acknowledgments

This work is supported by grant from the Natural Science Foundation of China (No.62372077).

Contribution Statement

Xiaodi Hou and Xiaobo Li made equal contribution.

References

[Bu *et al.*, 2024] Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204, 2024.

[Chen *et al.*, 2020] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449. Association for Computational Linguistics, 2020.

[Chen *et al.*, 2021] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 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. Association for Computational Linguistics, 2021.

[Chu *et al.*, 2023] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Denkowski and Lavie, 2011] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91, 2011.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Gu *et al.*, 2021] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[Gu *et al.*, 2022] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hou *et al.*, 2023] Xiaodi Hou, Zhi Liu, Xiaobo Li, Xingwang Li, Shengtian Sang, and Yijia Zhang. MKCL: Medical Knowledge with Contrastive Learning model for radiology report generation. *Journal of Biomedical Informatics*, 146:104496, 2023.

[Hou *et al.*, 2025] Xiaodi Hou, Xiaobo Li, Zhi Liu, Shengtian Sang, Mingyu Lu, and Yijia Zhang. Recalibrated cross-modal alignment network for radiology report generation with weakly supervised contrastive learning. *Expert Systems with Applications*, page 126394, 2025.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[Huang *et al.*, 2023] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023.

[Hui *et al.*, 2019] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019.

[Irvin *et al.*, 2019] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 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, 2019.

[Jin *et al.*, 2024] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 2607–2615, 2024.

[Johnson *et al.*, 2019] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[Li *et al.*, 2022a] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2022b] Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20656–20665, 2022.

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

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

[Liu *et al.*, 2021] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 13753–13762, 2021.

[Liu *et al.*, 2024] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *CoRR*, abs/2401.10166, 2024.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Shen *et al.*, 2024] Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4776–4783, 2024.

[Shin *et al.*, 2016] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.

[Su *et al.*, 2023] J Su, Y Lu, S Pan, A Murtadha, B Wen, and Y Liu Roformer. Enhanced transformer with rotary position embedding., 2021. *DOI: https://doi. org/10.1016/j. neucom*, 2023.

[Sun *et al.*, 2025] Yongheng Sun, Yueh Z Lee, Genevieve A Woodard, Hongtu Zhu, Chunfeng Lian, and Mingxia Liu. 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, 2025.

[Tanida *et al.*, 2023] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023.

[Wang *et al.*, 2022a] Song Wang, Liyan Tang, Mingquan Lin, George Shih, Ying Ding, and Yifan Peng. Prior knowledge enhances radiology report generation. In *AMIA Annual Symposium Proceedings*, volume 2022, page 486. American Medical Informatics Association, 2022.

[Wang *et al.*, 2022b] Tao Wang, Junlin Lan, Zixin Han, Ziwei Hu, Yuxiu Huang, Yanglin Deng, Hejun Zhang, Jianchao Wang, Musheng Chen, Haiyan Jiang, et al. O-net: a novel framework with deep fusion of cnn and transformer for simultaneous segmentation and classification. *Frontiers in neuroscience*, 16:876065, 2022.

[Wang *et al.*, 2023] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023.

[Wang *et al.*, 2024] Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. Camanet: class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[Xue *et al.*, 2024] Youyuan Xue, Yun Tan, Ling Tan, Jiaohua Qin, and Xuyu Xiang. Generating radiology reports via auxiliary signal guidance and a memory-driven network. *Expert Systems with Applications*, 237:121260, 2024.

[Yan and Pei, 2022] Bin Yan and Mingtao Pei. Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990, 2022.

[Yan *et al.*, 2021] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4009–4015. Association for Computational Linguistics, 2021.

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

[Yue and Li, 2024] Yubiao Yue and Zhenzhang Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.

[Zhang *et al.*, 2020] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917, 2020.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.