CXPMRG-Bench: Pre-training and Benchmarking for X-ray Medical Report Generation on CheXpert Plus Dataset
– Supplementary Material –

Xiao Wang[1], Fuling Wang[1], Yuehang Li[1], Qingchuan Ma[1], Shiao Wang[1], Bo Jiang[1*], Jin Tang[1]
[1] School of Computer Science and Technology, Anhui University, Hefei, China
{*xiaowang, jiangbo, tangjin*}*@ahu.edu.cn*
{*e23201049, e23201112, e02114334*}*@stu.ahu.edu.cn, wsa1943230570@126.com*

1. Related Works

1.1. State Space Model

Since its introduction in 2017, Transformer [56] has quickly become the preferred model framework for researchers due to its strong performance. However, as the model scales and sequences become longer, its limitations have surfaced. One major drawback is the quadratic growth in computational complexity of the self-attention mechanism with increased context length. Mamba [21] addresses these issues by using Selective State Space Models (SSMs) to improve traditional state space models and incorporating a hardware-aware parallel algorithm for recurrent operations. Vim [85] (Vision Mamba) is the first SSM model adapted for vision tasks. It uses positional embeddings and bidirectional state space models to achieve high performance, particularly on high-resolution images. VMamba [40] extends Mamba by providing a global receptive field with linear complexity. MambaMLP [51] is a new architectural component based on Mamba, designed to enhance feature mixing and representation learning by combining Mamba with an MLP, thereby improving performance on visual tasks. The new SSD (State Space Duality) algorithm proposed by Mamba-2 [15] can fully utilize matrix multiplication units on modern hardware, making it 2-8 times faster than the vanilla Mamba. The successful applications of the Mamba in many computer vision tasks [28, 62, 65] inspired us to adapt it to the pre-trained X-ray large model for medical report generation.

2. Dataset and Evaluation Metric

• **IU X-ray Dataset** [17] [1] published in 2016 is one of the most frequently used publicly available medical image datasets for medical report generation. It contains 7,470 images and 3,955 radiology reports, with each report associated with either frontal or both frontal and lateral view images. Each report is divided into four sections: Indication, Comparison, *Findings*, and *Impression*. For a fair comparison, we used the same dataset split protocol as R2GenGPT [68], dividing the dataset into training, testing, and validation sets with a ratio of 7:1:2.

• **MIMIC-CXR Dataset** [32] [2] is one of the largest publicly available chest X-ray datasets, containing free-text radiology reports. These records from 2011-2016 include 377,110 radiographic images and 227,835 radiology reports collected from 65,379 patients at the Beth Israel Deaconess Medical Center Emergency Department in Boston, Massachusetts. For fair comparison, we used the same dataset split protocol as R2GenGPT, with 270,790 samples for training the model, and 2,130 and 3,858 samples for validation and testing sets, respectively.

• **CheXpert Plus Dataset** [7] [3] is a new radiology dataset designed to enhance the scale, performance, robustness, and fairness of deep learning models in the field of radiology. This dataset includes 223,228 chest X-rays (in DICOM and PNG formats), 187,711 corresponding radiology reports (de-identified and parsed into 11 sections), de-identified demographic data from 64,725 patients, 14 chest pathology labels, and RadGraph [30] annotations. For a fair comparison, we followed the dataset split protocol used in R2GenCSR [63] which adopted *Findings* as the ground truth and split the training/validation/testing subset based on the ratio 7:1:2. The training subset with 40,463 samples, the validation subset with 5,780 samples, and the testing subset with 11,562 samples. Given that current researchers tend to focus on the *Findings* section of the dataset rather than the *Impressions* section, and considering that the *Impressions* often contains a significant amount of irrelevant information that could negatively impact the model's performance,

*✉ Corresponding Author: Bo Jiang

[1] https://iuhealth.org/find-medical-services/x-rays

[2] https://physionet.org/content/mimic-cxr/2.0.0/

[3] https://github.com/Stanford-AIMI/chexpert-plus

we have chosen to use the *Findings* section, as it contains precise and relevant medical report information.

More in detail, CIDEr [57] evaluates text through TF-IDF weighted n-gram matching, placing greater emphasis on the importance of words; BLEU [47] evaluates text quality through n-gram matching; ROUGE-L [35] evaluates text using the longest common subsequence; METEOR [2] improves upon BLEU by considering synonyms and word order.

3. Implementation Details

• **Pre-training Stage.** Both MambaXray-VL-Base and MambaXray-VL-Large were pre-trained for 100 epochs, with batch sizes set at 256 and 128, respectively. The base learning rate, based on a batch size of 256, was set to 1.5e-4. We adopted a cosine decay schedule with a warm-up for 5 epochs and used the AdamW [41] optimizer with a weight decay of 0.05. The resolution of input images is resized to 192×192 in the pre-training phase.

In the second stage, we utilized a vision-text contrastive learning pre-training method to train MambaXray-VL, enabling alignment to the text feature space. Specifically, we used a dataset of 480,000 image-text pairs, composed of publicly available datasets from MIMIC-CXR [32], CheXpert Plus [7], and IU-Xray [17]. Inspired by ARM [51], we used a unidirectional scanning approach in the first stage that fits the autoregressive generation to achieve more efficient pre-training. In the second stage, we extend the scanning block to four copies in order to improve the performance of the model. During this stage, we chose to pre-training for 50 epochs, with a batch size set to 192. The visual encoder was Vim [85], loaded with weights from the first stage of pre-training, while the text encoder was Bio_ClinicalBERT [1], both encoders were set to be trainable. We employed the same optimizer as in the first stage, but the input image size was changed to 224×224.

4. Experiment

4.1. Comparison on Public Benchmark Datasets

• **Results on IU X-ray Dataset.** As shown in Table 1, it can be seen that both our MambaXray-VL-Base and MambaXray-VL-Large exhibit excellent performance on the IU X-ray dataset. Among them, the MambaXray-VL-Large model is at the SOTA level on BLEU-2, BLEU-3, and BLEU-4 metrics with scores of 0.330, 0.241, and 0.185, respectively. This result indicates the superiority of our method over other report generation methods. However, on some other metrics such as BLEU-1, ROUGE-L, METEOR, and CIDEr, our method does not achieve optimal performance. This reflects the need to improve the generalization of our method on other datasets.

• **Results on MIMIC-CXR Dataset.** As shown in Table 1, our method also demonstrates outstanding performance on the MIMIC-CXR dataset, surpasses all other advanced report generation methods, and achieves the most advanced level in several common indicators (e.g., BLEU-1, BLEU-2, BLEU-3, and BLEU-4). Specifically, our method improves the BLEU-4 metric by 6% compared to R2GenGPT. Encouragingly, we achieved favorable results for two of the three remaining metrics, ROUGE-L and METEOR, further demonstrating the superior performance of our model. Moreover, compared to other vision-language pretraining models like PTUnifier [13] and PhenotypeCLIP [61], our method also leads in all metrics, especially in BLEU-4. This further highlights the robustness and superiority of our model.

4.2. Clinical Efficacy Metrics

Clinical Efficacy (CE) metrics have significant practical value, as they can assess report quality to ensure usability and reliability in real medical scenarios, thereby improving the quality of healthcare services and patient safety. According to R2Gen [9], unless otherwise specified, this study adopts macro-average for CE metrics. As shown in Table 2, our model also reports CE metrics on the Mimic-CXR dataset. Our model surpasses all existing methods in terms of Recall and F1 score, and achieves commendable performance in Precision, only slightly trailing behind HERGen [58]. Overall, our model demonstrates strong performance in CE metrics, reflecting its robustness and efficiency.

We provide results calculated using both *macro-average* and *micro-average* based on **14** key categories. Macro-average scores tend to be lower because they treat all categories equally, assigning the same weight to both high-frequency and low-frequency classes. In contrast, some prior studies, such as the RGRG [54] and DCL [33], have reported CE metrics using micro-average.

Notably, if we adopt the same micro-average approach, as shown in Table 2, our model achieves a precision of 0.561, a recall of 0.460, and an F1-score of 0.505. These results are competitive with state-of-the-art methods and even outperform them in certain aspects.

4.3. Visualization

As shown in Fig. 1, we give some examples to illustrate the effectiveness of our proposed MambaXray-VL model for the X-ray image based report generation. For specific X-ray images, we compared ground truth with the report generated by the MambaXray-VL model and the report generated by the R2GenGPT model. The X-ray images we chose contain both front and side views, normal images, and images containing lesion areas, enabling a more comprehensive and rational visualization. For a more intuitive visual-

Table 1. Comparison of our model's performance on the IU X-ray and MIMIC-CXR datasets. The symbol † indicates that we follow the R2Gen annotation using *Findings* and evaluate with our method, as their report modifies the ground truth to an *Impression* concatenated with *Findings*. The best result is highlighted in bold, and the second-best result is underlined. This symbol ⋆ indicates that the algorithm is a visual language pre-trained model like ours.

Dataset	Methods	Publication	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
IU X-Ray	R2Gen [9]	EMNLP 2020	0.470	0.304	0.219	0.165	0.371	0.187	-
	R2GenCMN [10]	ACL-IJCNLP 2021	0.475	0.309	0.222	0.170	0.375	0.191	-
	PPKED [39]	CVPR 2021	0.483	0.315	0.224	0.168	0.376	0.187	0.351
	AlignTrans [80]	MICCAI 2021	0.484	0.313	0.225	0.173	0.379	0.204	-
	CMCL [38]	ACL 2021	0.473	0.305	0.217	0.162	0.378	0.186	-
	Clinical-BERT [75]	AAAI 2022	0.495	**0.330**	0.231	0.170	0.376	0.209	0.432
	METransformer [67]	CVPR 2023	0.483	0.322	0.228	0.172	0.380	0.192	0.435
	DCL [33]	CVPR 2023	-	-	-	0.163	0.383	0.193	**0.586**
	R2GenGPT† [68]	Meta Radiology 2023	0.465	0.299	0.214	0.161	0.376	0.219	0.542
	PromptMRG [31]	AAAI 2024	0.401	-	-	0.098	0.160	**0.281**	-
	BootstrappingLLM [37]	AAAI 2024	**0.499**	0.323	0.238	0.184	**0.390**	0.208	-
	MambaXray-VL-Base	Ours	0.479	0.322	0.236	0.179	0.388	0.215	0.508
	MambaXray-VL-Large	Ours	0.491	**0.330**	**0.241**	**0.185**	0.371	0.216	0.524
MIMIC-CXR	R2Gen [9]	EMNLP 2020	0.353	0.218	0.145	0.103	0.277	0.142	-
	R2GenCMN [10]	ACL-IJCNLP 2021	0.353	0.218	0.148	0.106	0.278	0.142	-
	PPKED [39]	CVPR 2021	0.360	0.224	0.149	0.106	0.284	0.149	0.237
	AlignTrans [80]	MICCAI 2021	0.378	0.235	0.156	0.112	0.283	0.158	-
	CMCL [38]	ACL 2021	0.344	0.217	0.140	0.097	0.281	0.133	-
	Clinical-BERT [75]	AAAI 2022	0.383	0.230	0.151	0.106	0.275	0.144	0.151
	METransformer [67]	CVPR 2023	0.386	0.250	0.169	0.124	**0.291**	0.152	**0.362**
	DCL [33]	CVPR 2023	-	-	-	0.109	0.284	0.150	0.281
	R2GenGPT† [68]	Meta Radiology 2023	0.408	0.256	0.174	0.125	0.285	0.167	0.244
	PromptMRG [31]	AAAI 2024	0.398	-	-	0.112	0.268	0.157	-
	BootstrappingLLM [37]	AAAI 2024	0.402	0.262	0.180	0.128	**0.291**	**0.175**	-
	PTUnifer⋆ [13]	ICCV 2023	-	-	-	0.107	-	-	0.210
	PhenotypeCLIP⋆ [61]	ACL 2023	-	-	-	0.119	0.286	0.158	0.259
	MambaXray-VL-Base	Ours	0.420	0.264	0.180	0.129	0.283	0.162	0.206
	MambaXray-VL-Large	Ours	**0.422**	**0.268**	**0.184**	**0.133**	0.289	0.167	0.241

Table 2. Comparing the Clinical Efficacy (CE) metrics of different models on the Mimic-CXR dataset.

Method	Publication	Average	Precision	Recall	F1
R2Gen [9]	EMNLP 2020	Macro	0.333	0.273	0.276
METransformer [67]	CVPR 2023	Unclear	0.364	0.309	0.311
KiUT [29]	CVPR 2023	Unclear	0.371	0.318	0.321
MedRAT [24]	ECCV 2024	Unclear	0.285	0.265	0.227
CXR-IRGen [52]	WACV 2024	Unclear	-	-	0.293
HERGen [58]	ECCV 2024	Unclear	**0.415**	0.301	0.317
MambaXray-VL-L	Ours	Macro	0.371	**0.321**	**0.340**
DCL [33]	CVPR 2023	Micro	0.471	0.352	0.373
RGRG [54]	CVPR 2023	Micro	0.524	**0.474**	0.498
MambaXray-VL-L	Ours	Micro	**0.561**	0.460	**0.505**

ization, we have highlighted the parts that match the ground truth. The yellow highlighted area is the part of the report generated by our model that matches the ground truth, and the blue highlighted area is the part of the report generated by the R2GenGPT model that matches the ground truth. The pink highlighted area is the portion of the report generated by both our model and the R2GenGPT model that matches the ground truth. It is clear that the report generated by our model is closer to the real report than the report generated by the R2GenGPT model, which indicates that our model is effective.

As shown in Fig. 2, to present the experimental results more intuitively, we visualized the Clinical Efficacy (CE) metrics of all mainstream algorithms on the CheXpert Plus dataset using bar charts. The bar charts clearly show that our proposed model, MambaXray-VL-L, achieved the best results in all three metrics: Precision, Recall, and F1.

• **Does VLMs Pre-trained using Natural Image-Text Samples Ready for the X-ray Report Generation?** In this paper, we also conduct supervised fine-tuning on the CheXpert Plus dataset using Vision-Language Models (VLMs), including InternVL-2 [14] and MiniCPM V2.5 [79]. We replace R2Gen-GPT's vision and language backbone with VLMs to adapt them for X-ray image-based report generation. As illustrated in Table 3, we can find that the performance of the two models is not as good as the compared models. These experiments demonstrate a large gap between pre-training on the natural and X-ray images. In our future works, we consider further adapting the pre-trained VLMs using natural images to the X-ray image domain to achieve a better performance.

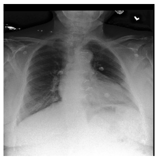
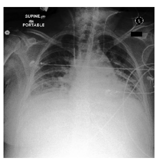
Image	Ground Truth	Ours	R2GenGPT
	Pa and lateral views of the chest provided. Midline sternotomy wires and mediastinal clips are again noted. The previously noted port-a-cath has been removed. The lungs are clear bilaterally without focal consolidation effusion or pneumothorax. Cardiomediastinal silhouette is stable. Bony structures are intact. No free air below the right hemidiaphragm is seen.	Ap upright and lateral views of the chest provided. Midline sternotomy wires and mediastinal clips are again noted. There is no focal consolidation large effusion or pneumothorax. The cardiomediastinal silhouette is stable. Bony structures are intact. No free air below the right hemidiaphragm is seen.	Frontal and lateral views of the chest were obtained. The patient is status post median sternotomy and cabg. The cardiac and mediastinal silhouettes are stable. There is no focal consolidation pleural effusion or pneumothorax. Mild pulmonary vascular congestion is noted. Degenerative changes are seen in the thoracic spine.
	Pa and lateral views of the chest provided. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen. Elevation of the right hemidiaphragm is unchanged from chest radiograph.	Pa and lateral views of the chest provided. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen. Clips are noted in the right upper quadrant of the abdomen.	Pa and lateral views of the chest provided. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen. Surgical clips in the right upper quadrant suggest prior cholecystectomy .
	Frontal and lateral views of the chest were obtained. Dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and right ventricle. The lungs are hyperinflated with flattening of the diaphragms suggesting chronic obstructive pulmonary disease. No pleural effusion or pneumothorax is seen. Slight increased opacity at the right lung base best seen on the fron.	Frontal and lateral views of the chest were obtained. Dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and right ventricle. There is no evidence of pneumothorax or pleural effusion. The lungs are hyperinflated with flattening of the diaphragms consistent with chronic obstructive pulmonary disease. Cardiomediastinal silhouette is stable. Bony structures are intact.	Frontal and lateral chest radiographs demonstrate hyperexpanded lungs with flattening of the diaphragms consistent with chronic obstructive pulmonary disease. There is no focal consolidation pleural effusion or pneumothorax. The cardiac mediastinal and hilar contours are unremarkable. A left-sided pacemaker device is noted with leads terminating in the right atrium and right ventricle.
	As compared to the previous radiograph there is no relevant change. The monitoring and support devices are constant. Low lung volumes borderline size of the cardiac silhouette. Mild pulmonary edema. Moderate retrocardiac atelectasis. No evidence of pneumonia.	As compared to the previous radiograph there is no relevant change. The monitoring and support devices are in unchanged position. Low lung volumes with minimal atelectasis at both lung bases. No larger pleural effusions or pneumothorax. Borderline size of the cardiac silhouette. No pulmonary edema. No other parenchymal abnormalities.	In comparison with the study of the monitoring and support devices remain in place. Continued enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases. No evidence of acute focal pneumonia or pneumothorax. Central catheters remain in place.

Figure 1. X-ray images and their corresponding ground-truths, along with the output of our model and R2GenGPT model generation reports on the MIMIC-CXR dataset. Matching sentences in our report are highlighted in yellow, R2GenGPT matching sentences are highlighted in cyan, and sentences matching by both models are highlighted in pink.

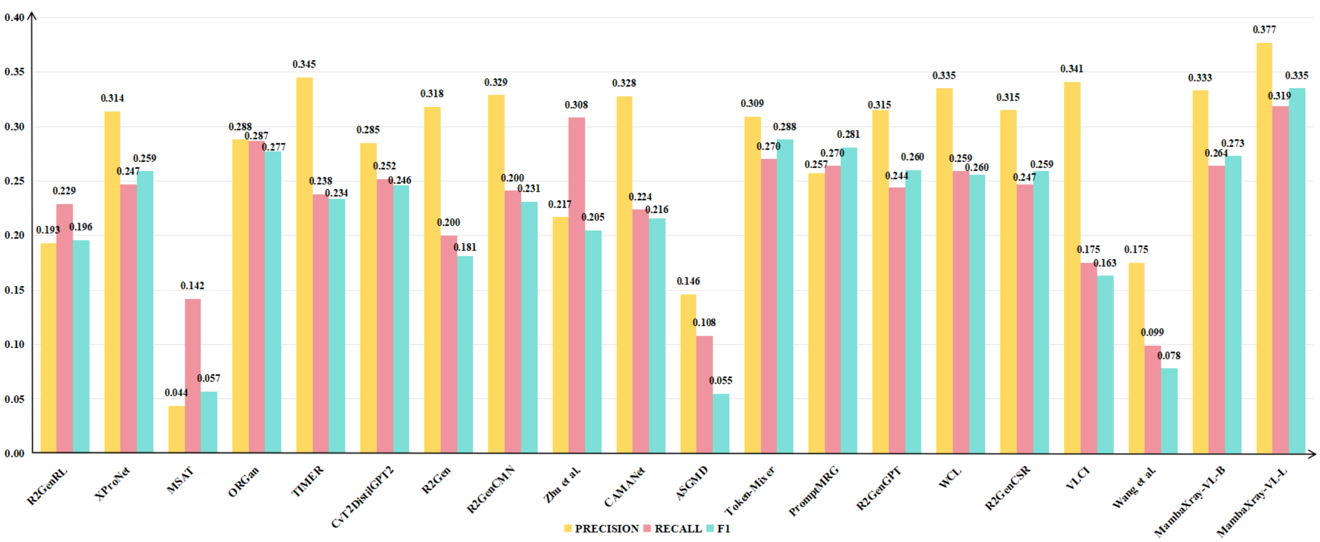


Figure 2. A bar chart visualizing the Clinical Efficacy (CE) metrics of all mainstream algorithms on the CheXpert Plus dataset. *Orange*, *Pink*, and *Cornflower blue* represent the Precision, Recall, and F1 metrics in CE, respectively.

4.4. Limitation Analysis

This paper provides a comprehensive benchmark for the X-ray image based medical report generation, which covers the mainstream MRG models and LLMs. The LLMs evaluated in this work focus on 7B and 13B which is hardware friendly, and the LLMs with more parameters are not discussed due to the limited computational resources. On the other hand, there are still many Vision-Language Models (VLMs) developed for natural images that are not benchmarked, due to the limited performance of the X-ray image-based medical report generation.

Table 3. Experimental Results of Medical Report Generation on the CheXpert Plus Dataset using different **LLMs and VLMs based on R2Gen-GPT**. The symbol † indicates that the model is a VLM. The Param listed in this table denotes the parameters of LLM/VLM.

Index	LLM/VLM	Year	B4	R-L	M	C	P	R	F1	Time (*min*)	Param	Code
#01	Vicuna-V1.5 [83]	2023	**0.104**	**0.272**	0.160	0.202	0.334	0.258	0.276	72.00	6.7B	URL
#02	Qwen-1.5 [20]	2024	0.098	0.262	0.139	0.139	0.303	0.233	0.241	154.25	7.7B	URL
#03	Qwen-2 [20]	2024	0.100	0.270	0.142	0.159	0.313	0.269	0.261	103.33	7.6B	URL
#04	InternLM [6]	2024	0.063	0.207	0.136	0.104	0.307	**0.274**	**0.284**	294.00	7.3B	URL
#05	Llama-2 [55]	2023	0.102	0.267	0.157	0.179	0.315	0.244	0.260	77.78	6.7B	URL
#06	Llama-2 [55]	2023	0.101	0.269	0.160	**0.214**	0.321	0.254	0.267	116.42	13.0B	URL
#07	Llama-3 [18]	2024	0.077	0.220	0.121	0.134	0.306	0.232	0.222	130.00	8.0B	URL
#08	Llama-3.1 [18]	2024	0.075	0.221	0.121	0.136	0.295	0.251	0.242	110.00	8.0B	URL
#09	GPT2-Medium [50]	2019	0.063	0.198	0.104	0.067	**0.358**	0.186	0.165	57.33	354M	URL
#10	Orca-2 [42]	2023	0.103	0.270	**0.161**	0.199	0.330	0.251	0.271	177.33	6.7B	URL
#11	Orca-2 [42]	2023	0.100	0.266	0.159	0.187	0.317	0.242	0.257	108.66	13.0B	URL
#12	Deepseek-LLM [4]	2024	0.096	0.268	0.137	0.150	0.336	0.256	0.253	201.30	6.9B	URL
#13	Yi-1.5 [82]	2024	0.091	0.263	0.131	0.136	0.322	0.229	0.226	43.66	6.1B	URL
#14	Yi-1.5 [82]	2024	0.096	0.269	0.138	0.155	0.336	0.241	0.243	48.50	8.8B	URL
#15	InternVL-2† [14]	2023	0.058	0.188	0.112	0.085	0.196	0.127	0.132	108.50	8.0B	URL
#16	MiniCPM-V2.5† [79]	2024	0.046	0.177	0.085	0.076	0.254	0.152	0.122	51.50	8.4B	URL

5. Discussion

• **We have attempted to replicate the mainstream algorithms on the CheXpert Plus dataset.** In this paper, we initially attempted to replicate the accuracy of 42 mainstream algorithms on the CheXpert Plus dataset. However, we successfully replicated only 19 algorithms in their entirety. The remaining 23 algorithms could not be replicated successfully due to various reasons. For instance, COMG [22] requires additional configuration files, DeltaNet [70] employs its own method for splitting the training and test sets, leading to unfair results, and CoFE [34] has not yet released its complete code. Table 4 shows the mainstream algorithms we specifically tried to replicate.

• **Why choose Mamba as the backbone?** Firstly, we fully acknowledge the computational efficiency of CNNs. However, our experiments and literature review indicate that while CNNs are computationally lightweight, they often fall short in performance compared to Transformer-based models on complex tasks. Transformers are renowned for their superior performance, particularly due to their ability to capture global context, although this comes at the cost of high computational complexity ($O(N^2)$). Mamba strikes an effective balance between these two extremes. With a reduced computational complexity ($O(N)$) and the ability to retain a global receptive field, Mamba is well-suited for tasks like report generation that benefit from a global context.

Secondly, while the input resolution in our experiments is 192×192, the original resolution of X-ray images is often very high, such as 3000×3000. Such high-resolution images generate a large number of input sequences during feature extraction. Efficiently handling these long sequences poses a significant challenge for traditional Transformer models due to their computational demands. In contrast, Mamba, with its optimized state-space model design, can process these sequences more efficiently.

Finally, although the current Mamba model demonstrates excellent performance in our experiments, we believe there is significant untapped potential in its application to medical image analysis. Further research into optimizing Mamba-based X-ray visual encoders can not only improve the trade-off between accuracy and efficiency for report generation but also provide valuable insights for other medical imaging tasks.

• **From a theoretical perspective, why does ARG perform better than MAE?** In the theoretical analysis, ARG is suitable for tasks that require progressively generating high-quality images, as it can capture fine-grained details of the image. However, it is computationally inefficient and training is complex. MAE [23] offers high training efficiency and is well-suited for handling large-scale data. [3, 5, 48, 76] points out that chest X-ray images have high contrast, rich details, and high similarity, with abnormal lesions typically occupying only a small portion of the image. The surrounding details of these areas also require special attention. ARG, through its step-by-step generation approach, can precisely capture the image details, making it particularly suitable for handling complex image structures like X-rays. Since each generation step depends on the previous one, it generally ensures high quality and consistency. When combined with Mamba's efficient computation capabilities, integrating ARG, chest X-rays, and Mamba can theoretically yield excellent results. On the other hand, MAE, which relies on large-scale masking and reconstruction, may struggle to effectively focus on the detailed ab-

Table 4. The mainstream algorithms we have attempted. ✓indicates successful replication on the CheXpert Plus dataset, while ✗indicates unsuccessful replication.

Index	Algorithm	Publish	Encoder	Decoder	Success
#01	R2GenRL [49]	ACL 2022	Transformer	Transformer	✓
#02	XProNet [59]	ECCV 2022	Transformer	Transformer	✓
#03	MSAT [66]	MICCAI 2022	ViT-B/16	Transformer	✓
#04	DeltaNet [70]	ICCL 2022	CNN	LSTM	✗
#05	RECAP [26]	EMNLP 2023	ViT	Transformer	✗
#06	RGRG [54]	CVPR 2023	ResNet-50	GPT-2	✗
#07	ORGen [27]	ACL 2023	CNN	Transformer	✓
#08	M2KT [77]	MIA 2021	CNN	Transformer	✓
#09	Delbrouck et al. [16]	EMNLP 2022	CNN	Bert	✗
#10	DCL [33]	CVPR 2023	ViT	Transformer	✗
#11	TIMER [71]	CHIL 2023	Transformer	Transformer	✓
#12	CvT2DistilGPT2 [44]	AIM 2023	Transformer	GPT-2	✓
#13	R2Gen [9]	EMNLP 2020	Transformer	Transformer	✓
#14	CheXbert [53]	EMNLP 2020	Bert	Bert	✗
#15	R2GenCMN [10]	ACL 2021	Transformer	Transformer	✓
#16	Zhu et al. [86]	MICCAI 2023	Transformer	Transformer	✓
#17	COMG [22]	WACV 2024	ResNet	Transformer	✗
#18	CAMANet [60]	IEEE JBH 2023	Swin-Former	Transformer	✓
#19	ASGMD [73]	ESWA 2024	ResNet-101 Transformer	Transformer	✓
#20	HERGen [58]	ECCV 2024	CvT	GPT-2	✗
#21	CoFE [34]	ECCV 2024	ViT-S+PubMedBERT	GPT-2	✗
#22	Token-Mixer [78]	IEEE TMI 2023	ResNet-50	Transformer	✓
#23	CXR-IRGen [52]	WACV 2024	CNN+ViT	Transformer	✗
#24	EKAGen [5]	CVPR 2024	ResNet+ViT	Transformer	✗
#25	PromptMRG [31]	AAAI 2024	ResNet-101	Bert	✓
#26	R2GenGPT [68]	Meta Radiology 2023	Swin-Transformer	Llama2-7B	✓
#27	R2-LLM [36]	AAAI 2024	ViT	Vicuna	✗
#28	WCL [74]	EMNLP 2021	Transformer	Transformer	✓
#29	RATCHET [25]	MICCAI 2021	DenseNet-121	Transformer	✗
#30	IFCC [43]	ACL 2021	M2Trans	Transformer	✗
#31	CXRMate-RRG24 [46]	arXiv 2024	UniFormer	Llama	✗
#32	ARL [12]	ACMMM 2022	CLIP-ViT-B+RoBERTa-base	Transformer	✗
#33	M3AE [11]	MICCAI 2022	CLIP-ViT-B+RoBERTa-base	Transformer	✗
#34	MedKLIP [69]	ICCV 2023	ResNet-50+ClinicalBERT	Transformer	✗
#35	MedicalMAE [72]	WACV 2023	ViT-S	Transformer	✗
#36	MRM [84]	ICLR 2023	ViT	Transformer	✗
#37	CXR-CLIP [81]	MICCAI 2023	ResNet-50	None	✗
#38	PTUnifier [13]	ICCV 2023	CLIP-ViT-B+RoBERTa-base	Transformer	✗
#39	CXRMate [45]	arXiv 2024	Transformer	Transformer	✗
#40	VLCI [8]	arXiv 2024	Transformer	Transformer	✓
#41	R2GenCSR [63]	arXiv 2024	VMamba	Llama2-7B	✓
#42	Wang et al. [64]	arXiv 2024	ViT	Llama2-7B	✓
#43	MambaXray-VL-B	Ours	MambaXray-VL	Llama2-7B	✓
#44	MambaXray-VL-L	Ours	MambaXray-VL	Llama2-7B	✓

normal lesion areas in X-ray images, leading to compatibility issues in downstream tasks, especially in medical report generation.

Based on [19], under the same pretraining settings, ARG models with autoregressive objectives outperform MAE models with masking objectives in terms of frozen backbone performance on ImageNet-1k. Ren et al. [51] also discovered that by combining ARG and Mamba, they could compensate for each other's shortcomings and achieve state-of-the-art performance on ImageNet-1k.

• **Explain why multi-stage training is chosen. What are the advantages of multi-stage training compared to joint training?** Models trained with multi-stage training perform better than those with joint training, and we use different datasets at each stage. Multi-stage training allows us to use more data. Specifically, at first, through the self-supervised autoregressive generation stage, the model can focus on extracting effective features from X-ray images and learning the basic structure of the images. In the contrastive learning stage, the model can further align the feature spaces of images and text, thereby improving the matching relationship between images and text. This phased training approach avoids the risk of conflicting objectives that might occur in joint training.

Second, multi-stage training can gradually optimize the model, enhancing the quality of image understanding and text generation. Compared to joint training, which simultaneously optimizes all objectives from the beginning, phased training allows for an initial focus on image encoding, followed by optimization of text generation and image-text alignment in later stages. This helps the model learn and generalize more effectively, improving training efficiency and stability.

Third, considering that different datasets are used in the three stages, in the first stage, ARG uses only X-ray images without corresponding reports, resulting in a dataset of over one million images. In the second stage, image-text contrastive learning requires image-report pairs, which are more limited in quantity. Since precise image-text alignment is not crucial in this stage, we use the Impressions section from the CheXpert-plus dataset, which is more abundant but less accurate than the Findings section, resulting in a dataset of around 500,000 pairs. In the third stage, downstream task fine-tuning involves refining the model on each specific dataset, using the most accurate parts of each dataset. If joint training were used, the available data would be very limited, making it difficult to fully utilize the potential of LLMs. Therefore, we chose multi-stage training.

As shown in Table 5, **Base** represents the base model trained without using the image-text contrastive learning strategy; **Joint** represents the model trained by combining image-text contrastive learning and supervised fine-tuning in a single stage; **Multi-Stage** represents the model trained using a multi-stage approach. It can be observed that the model trained with joint training performs significantly worse than the model trained with the multi-stage approach on all accuracy metrics, and even performs worse than the model without using the image-text contrastive learning strategy. We speculate that this is likely due to the conflicting objectives in joint training, leading to a decline in performance. This also empirically validates the effectiveness and robustness of our multi-stage training approach.

• **Details about the truncation operation.** When replicating different mainstream algorithms, the lack of a unified standard has led researchers to adopt varying levels of truncation for ground-truth reports. This discrepancy makes it challenging to fairly compare the performance of different algorithms. Therefore, we made every effort to apply a consistent no-truncation strategy across all algorithms, ensuring that the resulting accuracy is meaningful. Specifically, we modified the code of all mainstream algorithms so that the models output their predicted reports on the test set. We then directly compared these predicted reports with the complete ground-truth reports to calculate accuracy. This approach maximizes fairness in comparing different algorithms.

• **Other details.** we outline the steps we took to address reproducibility and ensure fairness in benchmarking: **Model Reproduction on MIMIC-CXR**: Our first step was to identify representative open-source works from recent years and attempt to reproduce their results on the MIMIC-CXR dataset. Since the CheXpert Plus dataset shares many similarities with MIMIC-CXR in terms of structure and task objectives, we hypothesized that any model successfully reproduced on MIMIC-CXR could also be effectively fine-tuned and evaluated on CheXpert Plus. **Dataset Preparation for CheXpert Plus**: To facilitate this process, we preprocessed the CheXpert Plus dataset to match the format of the MIMIC-CXR dataset, especially the configuration files. Specifically, the dataset was structured as follows:

```
{
    'train': [{'id': ..., 'image_path': ..., '
        report': ..., ...}, ...],
    'val': [{'id': ..., 'image_path': ..., '
        report': ..., ...}, ...],
    'test': [{'id': ..., 'image_path': ..., '
        report': ..., ...}, ...]
}
```

Fine-Tuning and Benchmarking on CheXpert Plus. Once the models were successfully reproduced on MIMIC-CXR, we fine-tuned and evaluated them on the CheXpert Plus dataset. The following measures were taken to ensure fairness and reproducibility: *Dataset Splits*: We used identical data splits for all models to maintain consistency across experiments. *Hyperparameter Settings*: While keeping most hyperparameters at their default values, we

Table 5. Comparing the performance of multi-stage training strategy and joint training on the Mimic-CXR dataset.

Strategy	NLG Metrics							CE Metrics		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Meteor	CIDEr	Precision	Recall	F1
Base	0.416	0.262	0.180	0.130	0.286	0.162	0.224	0.329	0.243	0.255
Joint	0.419	0.262	0.178	0.128	0.281	0.161	0.212	0.330	0.231	0.251
Multi-Stage	**0.422**	**0.268**	**0.184**	**0.133**	**0.289**	**0.167**	**0.241**	**0.371**	**0.321**	**0.340**

adjusted the batch size to maximize GPU memory usage on a single A800 GPU. Correspondingly, the learning rate was scaled to align with the new batch size. *Testing Process*: To ensure fair comparisons, we modified the evaluation code of certain models to output the generated reports during testing. These reports were then re-evaluated using a unified methodology for computing Natural Language Generation (NLG) metrics, eliminating inconsistencies caused by differing ground truth preprocessing methods. These steps were implemented to address the challenges of reproducibility and fairness in evaluating multiple models on a unified dataset. We hope these clarifications provide a comprehensive understanding of our efforts.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.

[2] Satanjeev Banerjee and Alon Lavie. 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, 2005.

[3] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449, 2024.

[4] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954, 2024.

[5] 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 (CVPR), pages 14194–14204, 2024.

[6] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. arXiv preprint arXiv:2403.17297, 2024.

[7] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538, 2024.

[8] Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Shen Zhao, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal intervention for medical report generation. arXiv preprint arXiv:2303.09117, 2023.

[9] 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, 2020.

[10] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Generating radiology reports via memory-driven transformer. In Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.

[11] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2022.

[12] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In Proceedings of the 30th ACM International Conference on Multimedia, 2022.

[13] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 23403–23413, 2023.

[14] Zhe Chen, Jiannan Wu, and Wenhai et al. Wang. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238, 2023.

[15] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In International Conference on Machine Learning (ICML), 2024.

[16] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4348–4360. Association for Computational Linguistics, 2022.

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

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[19] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. arXiv preprint arXiv:2401.08541, 2024.

[20] Jinze Bai et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.

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

[22] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 7995–8004, 2024.

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15979–15988, 2022.

[24] Elad Hirsch, Gefen Dawidowicz, and Ayellet Tal. Medrat: Unpaired medical report generation via auxiliary tasks. In European Conference on Computer Vision, pages 18–35. Springer, 2025.

[25] Benjamin Hou, Georgios Kaissis, Ronald Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest x-ray diagnosis and reporting, 2021.

[26] Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2134–2147, Singapore, 2023. Association for Computational Linguistics.

[27] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. ORGAN: Observation-guided radiology report generation via tree reasoning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8108–8122, Toronto, Canada, 2023. Association for Computational Linguistics.

[28] Ju Huang, Shiao Wang, Shuai Wang, Zhe Wu, Xiao Wang, and Bo Jiang. Mamba-fetrack: Frame-event tracking via state space model. arXiv preprint arXiv:2404.18174, 2024.

[29] 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.

[30] Saahil Jain, Ashwin Agrawal, and Adriel et al. Saporta. Radgraph: Extracting clinical entities and relations from radiology reports. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021.

[31] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(3):2607–2615, 2024.

[32] 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.

[33] 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 (CVPR), pages 3334–3343, 2023.

[34] Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. Contrastive learning with counterfactual explanations for radiology report generation. In European Conference on Computer Vision, pages 162–180. Springer, 2025.

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

[36] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In AAAI, pages 18635–18643, 2024.

[37] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 18635–18643, 2024.

[38] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical 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 3001–3012, Online, 2021. Association for Computational Linguistics.

[39] 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/CVF conference on computer vision and pattern recognition, pages 13753–13762, 2021.

[40] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166, 2024.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR, 2019.

[42] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, et al. Orca 2: Teaching small language models how to reason. arXiv preprint arXiv:2311.11045, 2023.

[43] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 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, pages 5288–5304, Online, 2021. Association for Computational Linguistics.

[44] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest X-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine, 144:102633, 2023.

[45] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation. arXiv preprint arXiv:2307.09758, 2023.

[46] Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. 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. Association for Computational Linguistics, 2024.

[47] 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.

[48] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. arXiv preprint arXiv:2311.18681, 2023.

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

[50] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[51] Sucheng Ren, Xianhang Li, Haoqin Tu, Feng Wang, Fangxun Shu, Lei Zhang, Jieru Mei, Linjie Yang, Peng Wang, Heng Wang, et al. Autoregressive pretraining with mamba in vision. arXiv preprint arXiv:2406.07537, 2024.

[52] Junjie Shentu and Noura Al Moubayed. Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 5212–5221, 2024.

[53] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 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, Online, 2020. Association for Computational Linguistics.

[54] 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 (CVPR), pages 7433–7442, 2023.

[55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017.

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

[58] Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal data. In Computer Vision–ECCV 2024: 19th European Conference, 2024.

[59] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV, pages 563–579. Springer, 2022.

[60] 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.

[61] Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. Fine-grained medical vision-language representation learning for radiology report generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15949–15956, 2023.

[62] Xiao Wang, Weizhe Kong, Jiandong Jin, Shiao Wang, Ruichong Gao, Qingchuan Ma, Chenglong Li, and Jin Tang. An empirical study of mamba-based pedestrian attribute recognition. arXiv preprint arXiv:2407.10374, 2024.

[63] Xiao Wang, Yuehang Li, Fuling Wang, Shiao Wang, Chuanfu Li, and Bo Jiang. R2gencsr: Retrieving context samples for large language model based x-ray medical report generation. arXiv preprint arXiv:2408.09743, 2024.

[64] Xiao Wang, Yuehang Li, Wentao Wu, Jiandong Jin, Yao Rong, Bo Jiang, Chuanfu Li, and Jin Tang. Pre-training on high definition x-ray images: An experimental study. arXiv preprint arXiv:2404.17926, 2024.

[65] Xiao Wang, Shiao Wang, Xixi Wang, Zhicheng Zhao, Lin Zhu, Bo Jiang, et al. Mambaevt: Event stream based visual object tracking using state space model. arXiv preprint arXiv:2408.10487, 2024.

[66] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 655–664. Springer, 2022.

[67] 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 (CVPR), pages 11558–11567, 2023.

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

[69] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[70] Xian Wu, Shuxin Yang, Zhaopeng Qiu, Shen Ge, Yangtian Yan, Xingwang Wu, Yefeng Zheng, S Kevin Zhou, and Li Xiao. Deltanet: Conditional medical report generation for covid-19 diagnosis. arXiv preprint arXiv:2211.13229, 2022.

[71] Yuexin Wu, I-Chan Huang, and Xiaolei Huang. Token imbalance adaptation for radiology report generation. CHIL-2023, 209, 2023.

[72] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3588–3600, 2023.

[73] 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.

[74] 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, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[75] 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, pages 2982–2990, 2022.

[76] Sixing Yan, William K Cheung, Ivor W Tsang, Keith Chiu, Terence M Tong, Ka Chun Cheung, and Simon See. Ahive: Anatomy-aware hierarchical vision encoding for interactive radiology report retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14324–14333, 2024.

[77] 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.

[78] Yan Yang, Jun Yu, Zhenqi Fu, Ke Zhang, Ting Yu, Xianyun Wang, Hanliang Jiang, Junhui Lv, Qingming Huang, and Weidong Han. Token-mixer: Bind image and text in one embedding space for medical image reporting. IEEE Transactions on Medical Imaging, pages 1–1, 2024.

[79] Yuan Yao, Tianyu Yu, and Ao et al. Zhang. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint 2408.01800, 2024.

[80] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, pages 72–82. Springer, 2021.

[81] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K. Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, pages 101–111. Springer Nature Switzerland, 2023.

[82] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652, 2024.

[83] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36:46595–46623, 2023.

[84] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling, 2023.

[85] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In Forty-first International Conference on Machine Learning, 2024.

[86] Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M. Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, pages 189–198, Cham, 2023. Springer Nature Switzerland.