MambaXray-CTL: Multi-Stage Contrastive Training for Medical Report Generation with a Mamba-Based Multi-Modal Large Model

Wenbin Feng[1], Yu Lu[1,2,*], Xiaoqing Li[3], Shijie Shi[1], Yingjian Qi[1]

Abstract— **The rapid advancement of artificial intelligence (AI) is revolutionizing radiology, particularly in the automation of medical report generation. AI-driven systems offer the potential to alleviate the increasing workload of radiologists while improving diagnostic accuracy, consistency, and overall workflow efficiency. Despite significant progress in multi-modal medical image captioning, existing approaches often suffer from high computational costs and limitations in modeling long-range dependencies between visual and textual features. To address these challenges, we propose MambaXray-CTL, a novel framework for medical report generation that integrates a Mamba-based vision backbone with a large language model (LLM) text decoder. By leveraging a multi-stage training strategy—including autoregressive pretraining, image-text contrastive learning, and supervised fine-tuning with contrastive regularization—our model achieves precise visual-textual alignment while maintaining computational efficiency. Experimental results on the IU X-ray and CheXpertPlus datasets demonstrate that MambaXray-CTL achieves performance comparable to or surpassing state-of-the-art methods in key metrics, particularly BLEU-4 and CIDEr, while significantly reducing inference cost compared to Vision Transformer (ViT)-based architectures. These findings highlight the promise of state space models and contrastive learning in building scalable and effective vision-language systems for real-world clinical deployment.**

I. INTRODUCTION

The integration of artificial intelligence (AI) into radiology report generation marks a major advancement in healthcare, enabling automated, clinically accurate reporting from medical images such as chest X-rays. This innovation alleviates radiologists' increasing workloads and enhances diagnostic accuracy, improving patient outcomes. With the growing volume of medical imaging data, automated report generation enhances efficiency, allowing physicians to focus on complex cases while ensuring consistency and reducing human errors.

Deep learning, particularly contrastive learning and Large Language Models (LLMs), has driven progress in multimodal medical report generation. The Dynamic Graph Enhanced Contrastive Learning (DCL) framework [10] improves visual-textual alignment, while the Token-Mixer model [26] unifies image-text representations. The Hierarchical Semantic Alignment (HSA) framework [29] uses game-theoretic interactions for multi-level semantic alignment. Other approaches, such as ECRG [8], GMoD [23], and KARGEN [11], incorporate energy functions, knowledge distillation, and medical knowledge graphs to enhance report accuracy. However, most rely on Transformer-based architectures with quadratic complexity $O(N^2)$, limiting scalability.

Mamba, a State Space Model (SSM)-based alternative [6], reduces computational costs while maintaining efficiency in long-sequence modeling [28]. Vision Mamba and VMamba [15] have demonstrated strong performance in high-resolution medical imaging. MambaXray-VL [20] replaces Vision Transformers with Vision Mamba, reducing computational costs in X-ray report generation. Collectively, these innovations advance the quality and clinical utility of automated medical report generation systems. However, the primary challenge remains in generating accurate reports, which requires precise alignment between visual tokens and textual tokens.

In this paper, we propose MambaXray-CTL, a novel medical report generation framework with multi-scale contrastive learning that leverages the Mamba vision backbone to address the high computational cost and achieves precise alignment between multi-modality features. The model consists of three key components: (1) unsupervised pre-training, where the Auto-regressive pre-training with Mamba in Vision (ARM) module [17] is used to pre-train the vision backbone, enabling the model to learn robust visual features from large-scale unlabeled image data; (2) image-text contrastive learning, which aligns image and report embeddings using the Mamba vision backbone and an LLM-based text decoder to ensure effective interaction between visual and textual representations; and (3) supervised fine-tuning with integrated contrastive learning, where text labels are embedded back into the latent space to compute similarity between paired and unpaired samples, further reinforcing alignment and improving report accuracy. Experiments on the IU X-ray and CheXpertPlus datasets demonstrate that **MambaXray-CTL** achieves state-of-the-art performance, effectively balancing computational efficiency and clinical relevance.

In summary, the contributions of this paper are as follows:

1) We propose MambaXray-CTL, a novel framework that integrates the Mamba vision backbone and an LLM-based decoder to reduce computational costs while achieving high-quality medical report generation. The model attains state-of-the-art results on the IU X-ray and CheXpertPlus datasets.
2) We design a multi-stage training strategy that includes unsupervised pretraining, image-text contrastive learning, and supervised fine-tuning with contrastive loss. This pipeline enhances the alignment between visual and textual features, boosting clinical accuracy.

[1] Wenbin Feng, Yu Lu, Shijie Shi, and Yingjian Qi are with College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China

[2] Yu Lu is also with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China

[3] Xiaoqing Li is with the Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[*] Corresponding author. Email: `lvyu@sztu.edu.cn`

3) We introduce a contrastive fine-tuning mechanism that embeds textual labels into the latent space, aligning them with image features. This reinforces vision-language consistency and improves the faithfulness of generated reports.

II. METHOD

In this section, we first introduce the Mamba networks and then provide a detailed explanation of each stage of our proposed MambaXray-CTL framework.

A. Mamba-based Vision Backbone

Recent advancements in Mamba networks have been developed based on the continuous **State Space Model (SSM)**, which maps a 1-D function or sequence $x(t) \in \mathbb{R}^p$ to $y(t) \in \mathbb{R}^q$ through a hidden state $h(t) \in \mathbb{R}^n$. The continuous SSM is represented by the following equations:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \tag{1}$$

$$y(t) = \mathbf{C}h(t) \tag{2}$$

Since the data we process—images and text—are discrete, it is necessary to discretize the continuous SSM. For this, both the **S4** model [6] and the Mamba model employ the **Zero-Order Hold (ZOH)** transformation. The equations for this discretization are as follows:

$$\bar{A} = \exp(\Delta\mathbf{A}), \quad (1) \tag{3}$$

$$\bar{B} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - I) \cdot \Delta\mathbf{B}, \quad (2) \tag{4}$$

After discretizing $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using a step size Δ, we can reformulate the discrete version of the SSM as:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \quad (3) \tag{5}$$

$$y_t = \mathbf{C}h_t, \quad (4) \tag{6}$$

Finally, the output computation is performed through a global convolution:

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}\overline{\mathbf{B}}, \dots, \mathbf{C}\overline{\mathbf{A}}^{M-1}\overline{\mathbf{B}}), \quad (5) \tag{7}$$

$$\mathbf{y} = \mathbf{x} * \overline{\mathbf{K}}, \quad (6) \tag{8}$$

To further enhance the modeling and computational efficiency of the State Space Model (SSM), Gu et al. introduced the **Mamba** model [5], which transitions the model from being time-invariant to time-dependent. This adaptation not only accelerates the training and inference processes but also incorporates several hardware-aware algorithms. After its success in natural language processing (NLP), the Mamba model has been successfully adapted for use in computer vision tasks. For example, the **Vision Mamba** [28] and **VMamba** [15] models are notable adaptations of the Mamba framework in the domain of computer vision.

B. Multi-Stage Contrastive Training Framework

Our approach consists of three key stages: (1) **Unsupervised Pretraining**, where an Autoregressive Pretraining with Mamba in Vision framework (ARM) is employed to pretrain the Mamba-based vision backbone, enhancing feature extraction from chest X-ray images. (2) **Image-text Contrastive Learning**. This stage aligns the visual and textual feature spaces by optimizing a contrastive loss function, which maps embeddings of paired image-report data generated by the Mamba backbone and LLM encoder, respectively. (3) **Supervised Fine-tuning with Contrastive Learning**, where textual labels are embedded back into the latent space to optimize similarity metrics between paired and unpaired samples, ensuring the generation of accurate and clinically relevant reports. A diagram illustrating the overall process of the MambaXray-CTL model can be seen in Fig. 1.

Stage 1: Auto-regressive Pretraining. We use the pre-trained model from MambaXray-VL [20] for its strong visual feature extraction capability. X-ray images $\mathcal{I} \in \mathbb{R}^{192\times192\times3}$ are divided into patches $\mathcal{P}_i \in \mathbb{R}^{16\times16\times3}$, which are projected into visual tokens $\mathcal{T}_i \in \mathbb{R}^{1024}$ and processed by the Vim backbone encoder [28]. This process has a computational complexity of $\mathcal{O}(N)$, lower than the Transformer framework's $\mathcal{O}(N^2)$.

The visual tokens are normalized and processed by SSM and scan branches, with outputs combined using skip connections and further processed by SwiGLU [18]. Finally, an MLP layer is used for token reconstruction with autoregressive generation loss. The goal is to predict the next patch sequentially, with the loss function defined as:

$$\mathcal{L}_{\text{ARM}} = \sum_{i=1}^{n-1} l(Vim([P_1, ..., P_i]), P_{i+1})$$
$$l(\hat{y}, y) = |\hat{y} - y|^2$$

where $Vim(\cdot)$ is the Vim encoder, and P_i represents the i^{th} image patch.

Stage 2: Contrastive learning of image text. In this stage, we adopt the Vim backbone encoder from the first stage to perform image-text contrastive learning, aiming to align the visual and textual features. Specifically, we randomly sample a batch and embed the images and corresponding reports using the pre-trained Vim backbone encoder and the language model (Bio ClinicalBERT [1]). Then, we compute the cosine similarity of paired and unpaired samples within the same batch, which is defined as:

$$\mathcal{L}_{\text{CTL}} = \text{Similarity}(\text{Vim}(\mathcal{I}_i), \text{LM}(\mathcal{R}_j)) \tag{9}$$

where i and j denote the indices of the X-ray image and report, respectively, and $LM(\cdot)$ denotes the language model. The function Similarity$(\cdot)$ quantifies the degree of similarity between these two representations, with a lower value indicating greater similarity.

Stage 3: Supervised Fine-tuning with Contrastive Learning. Differently from [20], we perform supervised fine-tuning integrated with contrastive learning to further align

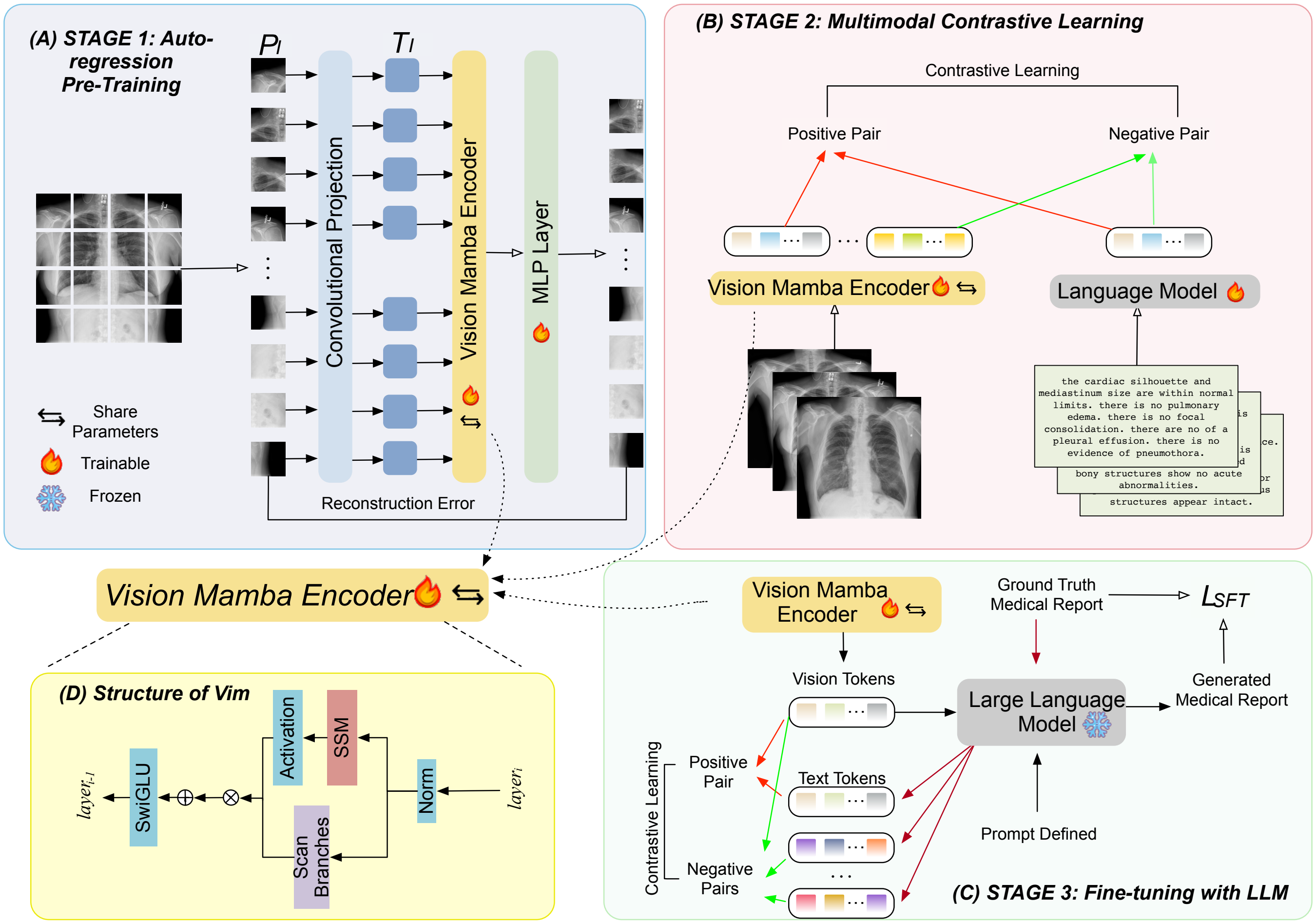


Fig. 1: Overview of the MambaXray-CTL framework, which consists of three stages: (1) autoregressive pretraining with Mamba to enhance visual feature learning, (2) image-text contrastive learning to align multi-modal embeddings, and (3) supervised fine-tuning with contrastive loss to further improve report generation accuracy.

text-image features (as shown in Fig. 3). First, we divide the X-ray image into non-overlapping patches and project them into visual tokens, which are then passed into the pre-trained Vim backbone encoder for feature extraction. Next, we wrap the visual embeddings with an instruction prompt: "Human: visual embeddings. Generate a comprehensive and detailed diagnosis report for this chest X-ray image. Assistant:", and process them through a Large Language Model (Qwen1.5-1.8B-Chat) to generate high-quality medical reports.

The loss function for supervised fine-tuning is:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{T} \log p_\theta \left(y_i \mid \text{Prompt},\, y_{<i} \right) \tag{10}$$

While fine-tuning the Vim backbone encoder, we embed the label text back into feature tokens to perform contrastive learning with visual tokens. This further aligns textual and visual features, ensuring the generation of accurate and clinically relevant reports. The loss function is the same as the second stage:

$$\mathcal{L}_{\text{CTL-FT}} = \text{Similarity}(\text{Vim}(\mathcal{I}_i), \text{LM}(\mathcal{R}_j)) \tag{11}$$

Finally, the total loss function for this stage is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SFT}} + \lambda \cdot \mathcal{L}_{\text{CTL-FT}} \tag{12}$$

where the parameter λ is a hyperparameter that controls the relative contribution of the contrastive learning loss to the total loss.

III. EXPERIMENT

A. Dataset

For the supervised image-text contrastive learning and fine-tuning with an LLM, IU-Xray [4] and CheXpert Plus [2] datasets are used in our work.

IU-Xray Dataset: consists of 7,470 chest radiographs (both frontal and lateral views) paired with 3,955 reports. Each report includes four standardized sections: *Indication*, *Comparison*, *Findings*, and *Impression*. Consistent with the established protocol in R2GenGPT [22], we adopted a 7:1:2 split ratio for training, validation, and testing.

CheXpert Plus: contains 223,462 pairs of chest X-ray images and radiology reports from 64,725 patients, with 14 pathology labels, and a large corpus of text data (36 million tokens). Following [20], we partitioned the dataset into training, validation, and testing subsets, following a 7:1:2 split.

B. Implementation Details

All models were trained on 8 Nvidia 4090 GPUs. In stage two, we trained for 50 epochs (batch size 192) using Bio ClinicalBERT [1] as the text encoder, with both vision and text encoders optimized using AdamW (weight decay 0.05). Input images were resized to 224×224.

In the fine-tuning stage, we trained for 30 epochs (batch size 20) on IU-Xray and 6 epochs (batch size 2) on CheXpert Plus. The Vim encoder [28] was initialized from stage two, while frozen LLMs (Qwen-1.5-1.8B and LLaMA2-7B) were used with sequence lengths of 60 and 100, respectively. Only the vision encoder and mapper layers were updated. The contrastive loss weight λ was set to 0.4.

C. Evaluation Metrics

To evaluate the quality of generated medical reports, we adopt three widely used metrics: BLEU, ROUGE-L, and CIDEr.

BLEU measures n-gram overlap between the generated and reference reports, with BLEU-1 to BLEU-4 capturing increasingly longer phrase-level matches. It primarily assesses lexical precision.

ROUGE-L evaluates the longest common subsequence between generated and reference texts, reflecting fluency and structural similarity.

CIDEr assigns higher weights to clinically relevant terms by incorporating TF-IDF statistics over consensus references, making it especially suitable for medical report generation tasks where semantic fidelity is critical.

Together, these metrics provide a comprehensive assessment of both syntactic accuracy and clinical relevance.

D. Experimental Results

The performance of our methods on the IU X-ray and CheXpert Plus datasets is summarized in Table I, with the best result in bold and the second-best underlined. Both base and large models excel in generating accurate medical reports, evaluated by CIDEr, BLEU, and ROUGE-L. CIDEr emphasizes clinically relevant phrases, BLEU assesses n-gram matching, and ROUGE-L evaluates the longest common subsequence. These metrics demonstrate the lexical precision, semantic relevance, and overall quality of our models' outputs.

On the IU X-Ray dataset, the MambaXray-CTL-Large model achieves state-of-the-art (SOTA) performance in BLEU-2, BLEU-3, and BLEU-4 with scores of 0.334, 0.252, and 0.199, respectively, demonstrating a significant improvement over other models. It also performs well in CIDEr with a score of 0.544, though BLEU-1 and ROUGE-L scores are 0.491 and 0.374, indicating room for further improvement. On the CheXpert Plus dataset, MambaXray-CTL-Large excels across all metrics, achieving a BLEU-1 score of 0.381, BLEU-2 of 0.240, BLEU-3 of 0.162, BLEU-4 of 0.114, ROUGE-L of 0.283, and CIDEr of 0.242. These results surpass the second-best models in each category, with improvements of up to 6.6%. The MambaXray-CTL-Base model also performs competitively with scores of 0.376 (BLEU-1), 0.232 (BLEU-2), 0.153 (BLEU-3), 0.104 (BLEU-4), 0.272 (ROUGE-L), and 0.227 (CIDEr), further highlighting the robustness of the MambaXray-CTL models.

In general, the MambaXray-CTL-Large and MambaXray-CTL-Base models consistently demonstrate their effectiveness in generating accurate, comprehensive, and clinically relevant medical reports in both datasets.

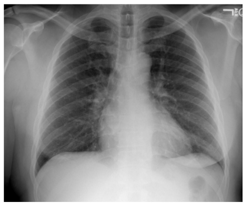
Original Image	Ours
	**MambaXray-CTL-Large:** the cardiomediastinal silhouette is normal in size and contour . pulmonary vascularity is within normal limits . no focal airspace consolidation pleural effusion or pneumothorax . xxxx xxxx are unremarkable . the osseous structures of the thorax are unremarkable .
Ground Truth the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .	**MambaXray-CTL-base:** heart size and pulmonary vasculature are within normal limits . lungs are clear . no pneumothorax or pleural effusion . osseous structures are normal .

Fig. 2: An illustration of the reports generated by our model with their corresponding ground truth. Matching sentences are highlighted in blue and green, respectively. Sentences matching by both Large and Base versions are highlighted in red.

Fig. 2 provides a qualitative comparison of the reports generated by the MambaXray-CTL-Large and Base models against the ground truth. Both models effectively capture critical clinical observations, with matching sentences highlighted in blue and green, reflecting alignment with the ground truth.

Notably, the Large model demonstrates greater detail, capturing nuanced observations such as "cardiomediastinal silhouette is normal in size and contour" and "no focal airspace consolidation, pleural effusion, or pneumothorax," showcasing its ability to generate more comprehensive X-ray interpretations. The Base model, while accurate, generates more concise reports, occasionally omitting finer details noted by the Large model. Sentences matching across both versions and the ground truth, highlighted in red, underscore the consistency of key findings between the models. These results clearly demonstrate that both of our proposed models can generate accurate and clinically relevant reports.

E. Ablation Study

The ablation study on the IU X-ray dataset using our Large model (Table II) highlights the contribution of each module in the proposed framework. Among the unsupervised pretraining strategies, the Autoregressive Pretraining with Mamba in Vision (ARM) module yields the most significant performance improvement. In this study, SIMCLR denotes Simple Contrastive Learning of Representations, MAE refers to Masked Autoencoders, ARM represents autoregressive

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
IU X-Ray	R2Gen [3]	0.470	0.304	0.219	0.165	0.371	-
	PPKED [14]	0.483	0.315	0.224	0.168	0.376	0.351
	AlignTrans [27]	0.484	0.313	0.225	0.173	0.379	-
	CMCL [13]	0.473	0.305	0.217	0.162	0.378	-
	Clinical-BERT [25]	0.495	0.330	0.231	0.170	0.376	0.432
	METransformer [21]	0.483	0.322	0.228	0.172	0.380	0.435
	DCL [10]	-	-	-	0.163	0.383	**0.586**
	R2GenGPT [22]	0.465	0.299	0.214	0.161	0.376	0.542
	PromptMRG [9]	0.401	-	-	0.098	0.160	-
	BootstrappingLLM [12]	**0.499**	0.323	0.238	0.184	**0.390**	-
	MambaXray-VL [20]	0.491	0.330	0.241	0.185	0.371	0.524
	MambaXray-CTL-Base	0.478	0.323	0.240	0.185	0.364	0.501
	MambaXray-CTL-Large	0.491	**0.334**	**0.252**	**0.199**	0.374	0.544
Chexpert Plus	R2Gen [3]	0.301	0.179	0.118	0.081	0.246	0.077
	R2GenCMN [3]	0.321	0.195	0.128	0.087	0.256	0.102
	XProNet [19]	0.364	0.225	0.148	0.100	0.265	0.121
	ORGan [7]	0.320	0.196	0.128	0.086	0.261	0.107
	R2GenGPT [22]	0.361	0.224	0.149	0.101	0.266	0.123
	ASGMD [24]	0.267	0.149	0.094	0.063	0.220	0.044
	Token-Mixer [26]	0.378	0.231	0.153	0.091	0.262	0.098
	PromptMRG [9]	0.326	0.174	-	0.095	0.222	0.044
	MambaXray-VL [20]	-	-	-	0.112	0.276	0.139
	MambaXray-CTL-base	0.376	0.232	0.153	0.104	0.272	0.227
	MambaXray-CTL-Large	**0.381**	**0.240**	**0.162**	**0.114**	**0.283**	**0.242**

TABLE I: Performance comparison of recent methods on IU X-ray and Chexpert Plus datasets using BLEU, ROUGE-L, and CIDEr metrics.

visual pretraining, CTL indicates contrastive learning in the second stage, CTL-FT refers to contrastive learning during fine-tuning, and SFT stands for supervised fine-tuning.

As shown in Table II, it is evident from indices #1, #2, and #3, where the inclusion of ARM results in higher scores, particularly a substantial $+40.8\%$ increase in CIDEr (calculated as $(0.607 - 0.431)/0.431$) from index #2 to #3. The importance of Contrastive Learning in the second stage (CTL) is underscored by the performance drop when it is removed, as shown by the BLEU-4 decrease of -4.3% between indices #3 and #4. The necessity of the ARM pretraining step is further highlighted by the performance gap between indices #4 and #5, with a -28.1% reduction in CIDEr when ARM is excluded. Additionally, the role of Contrastive Learning in the fine-tuning stage (CTL-FT) is evident, as it results in a $+6.3\%$ increase in BLEU-4 from index #7 to #4. These findings collectively emphasize the synergistic effects of these components in achieving superior performance.

F. Visualizations of Feature Space

Fig. 3 visualizes the embeddings from MambaXray-CTL with and without contrastive learning in Stage 3 (CTL-FT). In the left figure (only CTL in stage 2), image embeddings are misaligned with their positive-pair text, mixing with negative pairs and causing poor multimodal alignment. In contrast, the right figure (with CTL-FT) shows a well-structured space, where image and positive-pair text embeddings merge, while negative pairs are distinctly separated. This demonstrates that contrastive learning eliminates the modality gap, ensuring robust alignment and enhancing report generation quality.

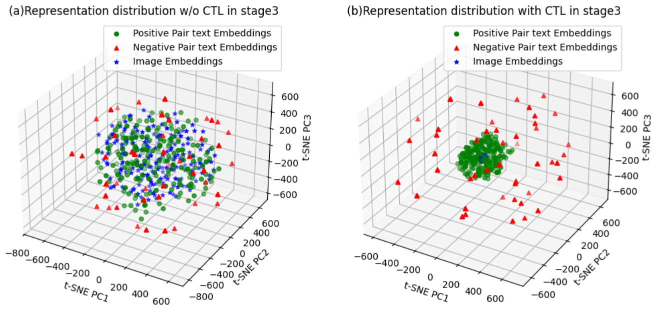


Fig. 3: t-SNE [16] visualization of image and text embeddings with and without contrastive learning in Stage 3.

IV. DISCUSSION

The proposed MambaXray-CTL framework demonstrates strong performance and computational efficiency by leveraging a Mamba-based vision backbone and multi-stage contrastive learning. The autoregressive pretraining enhances visual feature extraction, while contrastive learning significantly improves the alignment between visual and textual modalities. This design contributes to high-quality report generation with reduced resource demands, making it suitable for deployment in real-world clinical scenarios, especially in edge or fog computing environments.

However, several limitations remain. First, the current framework relies on pre-aligned image-report pairs, which may not fully capture the variability of clinical language across institutions. Second, while the model performs well on static frontal chest X-rays, its generalizability to other modalities (e.g., CT, MRI) or multi-view images has yet to

Index	SIMCLR	MAE	ARM	CTL	CTL-FT	SFT	BLEU-1	BLEU-4	ROUGE-L	CIDEr
#1	✓	×	×	✓	✓	✓	**0.488**	0.175	0.365	0.456
#2	×	✓	×	✓	✓	✓	0.471	0.177	0.358	0.431
#3	×	×	✓	✓	✓	✓	0.479	**0.193**	**0.375**	**0.607**
#4	×	×	✓	×	✓	✓	0.463	0.185	0.364	0.451
#5	×	×	×	×	✓	✓	0.444	0.175	0.358	0.352
#6	×	×	✓	✓	×	✓	0.482	0.178	0.362	0.461
#7	×	×	✓	×	×	✓	0.457	0.174	0.353	0.467

TABLE II: Ablation study on the IU X-ray dataset showing the impact of each module. SIMCLR, MAE, ARM, CTL, CTL-FT, and SFT denote different pretraining and fine-tuning strategies.

be explored. Last, the interpretability of the learned visual-textual alignments warrants further investigation to enhance clinical trust and adoption.

Future work will focus on extending the framework to multi-modal and multi-view medical imaging scenarios, incorporating domain adaptation techniques, and improving model explainability through attention visualization or saliency analysis.

V. CONCLUSIONS

This paper presents MambaXray-CTL, a novel framework for medical report generation that integrates a Mamba-based vision backbone with a large language model (LLM) decoder. A central contribution of our approach lies in the use of contrastive learning, which enhances the alignment between visual features and textual representations, thereby improving the quality of generated reports. By combining autoregressive pretraining, image-text contrastive learning, and supervised fine-tuning, MambaXray-CTL achieves state-of-the-art performance on the IU X-ray dataset, with notable gains in the BLEU-4 metric. The ablation study further highlights the importance of each module, particularly the impact of contrastive learning in refining cross-modal alignment.

ACKNOWLEDGMENT

This work was supported by the Key Project of Guangdong Province General Higher Education Institutions in Specific Research Areas (Grant No. 2023ZDZX2055), and the Open Project of the National Engineering Laboratory for Big Data System Computing Technology at Shenzhen University, China (Grant No. SZU-BDSC-OF2024-11).

REFERENCES

[1] Alsentzer, E., Murphy, J., Boag, W., et al.: Publicly Available Clinical BERT Embeddings. In: ClinicalNLP. pp. 72–78 (2019)

[2] Chambon, P., Delbrouck, J.B., Sounack, T., et al.: 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)

[3] Chen, Z., et al.: Generating Radiology Reports via Memory-driven Transformer. In: EMNLP. pp. 1439–1449 (2020)

[4] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., et al.: Preparing a collection of radiology examinations for distribution and retrieval. JAMIA **23**(2), 304–310 (2016)

[5] Gu, A., Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752 (2023)

[6] Gu, A., Goel, K., Re, C.: Efficiently Modeling Long Sequences with Structured State Spaces. In: ICLR (2022)

[7] Hou, W., et al.: ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning. In: ACL. pp. 8108–8122 (2023)

[8] Hou, Z., Yan, R., Yan, Z., et al.: Energy-Based Controllable Radiology Report Generation with Medical Knowledge. In: MICCAI. pp. 240–250. Springer (2024)

[9] Jin, H., Che, H., Lin, Y., Chen, H.: PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation. In: AAAI. pp. 2607–2615 (2024)

[10] Li, M., Lin, B., Chen, Z., et al.: Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation. In: CVPR. pp. 3334–3343 (2023)

[11] Li, Y., et al.: Kargen: Knowledge-enhanced automated radiology report generation using large language models. In: MICCAI. pp. 382–392. Springer (2024)

[12] Liu, C., Tian, Y., Chen, W.: Bootstrapping Large Language Models for Radiology Report Generation. In: AAAI. pp. 18635–18643 (2024)

[13] Liu, F., Ge, S., Zou, Y., Wu, X.: Competence-based Multimodal Curriculum Learning for Medical Report Generation. arXiv preprint arXiv:2206.14579 (2022)

[14] Liu, F., Wu, X., Ge, S., et al.: Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In: CVPR. pp. 13753–13762 (2021)

[15] Liu, Y., Tian, Y., Zhao, Y., et al.: Vmamba: Visual state space model. NeurIPS **37**, 103031–103063 (2025)

[16] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)

[17] Ren, S., Li, X., Tu, H., et al.: Autoregressive Pretraining with Mamba in Vision. arXiv preprint arXiv:2406.07537 (2024)

[18] Shazeer, N.: GLU Variants Improve Transformer. arXiv preprint arXiv:2002.05202 (2020)

[19] Wang, J., Bhalerao, A., He, Y.: Cross-Modal Prototype Driven Network for Radiology Report Generation. In: ECCV. pp. 563–579. Springer (2022)

[20] Wang, X., Wang, F., Li, Y., et al.: CXPMRG-Bench: Pre-training and Benchmarking for X-ray Medical Report Generation on CheXpert Plus Dataset. In: CVPR. pp. 5123–5133 (2025)

[21] Wang, Z., Liu, L., Wang, L., Zhou, L.: METransformer: Radiology Report Generation by Transformer With Multiple Learnable Expert Tokens. In: CVPR. pp. 11558–11567 (2023)

[22] Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology Report Generation with frozen LLMs. Meta-Radiology **1**(3), 100033 (2023)

[23] Xiang, Z., et al.: GMoD: Graph-Driven Momentum Distillation Framework with Active Perception of Disease Severity for Radiology Report Generation. In: MICCAI. pp. 295–305. Springer (2024)

[24] Xue, Y., et al.: Generating radiology reports via auxiliary signal guidance and a memory-driven network. Expert Systems with Applications **237**, 121260 (2024)

[25] Yan, B., Pei, M.: Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation. In: AAAI. pp. 2982–2990 (2022)

[26] Yang, Y., Yu, J., Fu, Z., et al.: Token-Mixer: Bind Image and Text in One Embedding Space for Medical Image Reporting. IEEE Transactions on Medical Imaging (2024)

[27] You, D., et al.: AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation. In: MICCAI. pp. 72–82 (2021)

[28] Zhu, L., Liao, B., Zhang, Q., et al.: Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In: ICML (2024)

[29] Zhu, Z., Cheng, X., Zhang, Y., et al.: Multivariate Cooperative Game for Image-Report Pairs: Hierarchical Semantic Alignment for Medical Report Generation. In: MICCAI. pp. 303–313. Springer (2024)