SwinMambaT5: Swin-Mamba Visual Encoding with T5 for Chest X-Ray Reporting

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Creating radiology reports is an essential clinical duty that typically demands significant time and specialized knowledge from experienced radiologists. Manual report writing requires significant effort and is prone to errors. These limitations have driven the development of automated systems for generating radiology reports, aiming to reduce the workload of radiologists and enhance the uniformity of diagnoses. However, existing automatic approaches rely on text decoders trained from scratch with limited medical vocabularies, constraining the fluency and clinical richness of the generated reports. In this study, we propose SwinMambaT5, a sequential pipeline that combines a Swin-Tiny visual encoder, Mamba sequence encoding, and a pretrained Flan-T5 text decoder. Swin-Tiny is fine-tuned efficiently using low-rank adaptation (LoRA) adapters. Mamba serves as a state-space model for sequential visual feature encoding. Flan-T5 is fine-tuned exclusively through cross-attention layers, bringing pre-trained clinical language knowledge to the decoder. This architecture addresses the clinical language quality limitation of scratch-trained decoder. We evaluated our model on two benchmark datasets: MIMIC-CXR and IU X-Ray. The experimental results achieved METEOR improvements of 87% and 94% over the existing baselines for MIMIC-CXR and IU X-Ray, respectively. The proposed approach trained only 14% of the total model parameters. Studies on ablation have verified that each component of the architecture contributes independently.

Keywords: :Radiology report generation · Mamba SSM · Flan-T5 · Cross-attention · Medical image captioning

1 Introduction

Generating radiology reports is an essential clinical task that demands significant time [5]. The manual report generation traditionally required the expertise of radiologists with extensive domain knowledge [5]. Composing manual reports requires significant labor and is susceptible to errors, particularly among less experienced practitioners. This highlights the need for automated solutions to reduce the workload of radiologists and enhance clinical automation [11]. The increasing volume of medical imaging studies has exacerbated the workload of radiologists. The substantial research interest in automating the reporting process aims to enhance efficiency and diagnostic consistency [9][14].

Li et al. [12] propose HRGR-Agent, a Hybrid Retrieval-Generation Reinforced Agent that employs a hierarchical decision-making procedure to generate radiology reports. The HRGR-Agent model works via reinforcement learning with sentence- and word-level rewards. The model assesses each sentence to decide whether to extract a template sentence from a pre-existing database or to engage a generation module to formulate a new sentence. HRGR-Agent suffers from limited generalization to unobserved clinical findings. Chen et al. [3] proposed R2Gen, which introduces a three-component architecture. The R2Gen components are the visual extractor, encoder, and text decoder. The text decoder was upgraded using relational memory and memory-conditional layer normalization. However, R2Gen trains its text decoder from scratch using a limited, domain-specific vocabulary. Chen et al. [2] extended R2Gen by proposing R2GenCMN. Similar to R2Gen, R2GenCMN has three components. The R2GenCMN components are the visual extractor, encoder, and text decoder. Chen et al. [2] proposed a cross-modal memory network (CMN). CMN are enabled by bidirectional interactions between visual and textual representations. Sun et al. [17] followed the same 3-component architecture as R2Gen. Sun et al. [17] replaced the encoder with a Mamba selective state space model (SSM) [6]. The use of SSM reduces the quadratic $\mathcal{O}(n^2)$ complexity of self-attention to a linear $\mathcal{O}(n)$. The use of Mamba also improves computational efficiency while maintaining competitive performance.

R2Gen-Mamba, however, continues to rely on a scratch-trained text decoder. Wang et al. [18] proposed the R2GenGPT pipeline with components such as a visual encoder, visual mapper, and Large Language Model (LLM). R2GenGPT incorporates a frozen LLM GPT for radiology report generation.Nicolson et al.[15] provide CvT2DistilGPT2, which integrates a convolutional vision transformer (CvT) encoder with a pretrained DistilGPT2 decoder. CvT2DistilGPT2 fine-tunes the entire DistilGPT2 decoder. Jin et al. [8] introduced PromptMRG, a model that integrates diagnosis-driven prompts obtained from a disease classification branch. The prompt of PromptMRG explicitly guided the report generation.

The literature review provides two research gaps. First, CNN-based visual encoders lack the capacity to model long-range anatomical dependencies across the full image. Second, scratch-trained text decoders produce reports with limited linguistic quality issues. SwinMambaT5 addresses both limitations using a hierarchical Swin Transformer visual encoder with LoRA adaptation. Swin is a visual feature extractor capturing long-range anatomical dependencies via shifted-window self-attention. Inspired by [17] we incorporated Mamba for sequential feature encoding. Unlike Wang et al. [18], we fine-tuned only the cross-attention layers of a pretrained Flan-T5 decoder. Fine-tuning cross-attention enables visual grounding while preserving the pretrained language knowledge.

In the next section, a discussion on the three components of the SwimMambaT5 architecture is presented. In the Experiments section, the dataset and implementation details are discussed. In the Results section, the experimental results are discussed. The final section presents the conclusion of the study.

2 Method

2.1 SwinMambaT5 Architecture

We propose SwinMambaT5, a sequential three-stage pipeline for the automated generation of chest X-ray reports. The architecture has three components: Swin as a visual decoder, Mamba as an efficient sequential feature encoder, and a pretrained language decoder, Flan-T5. Figure 1 illustrates our complete pipeline.

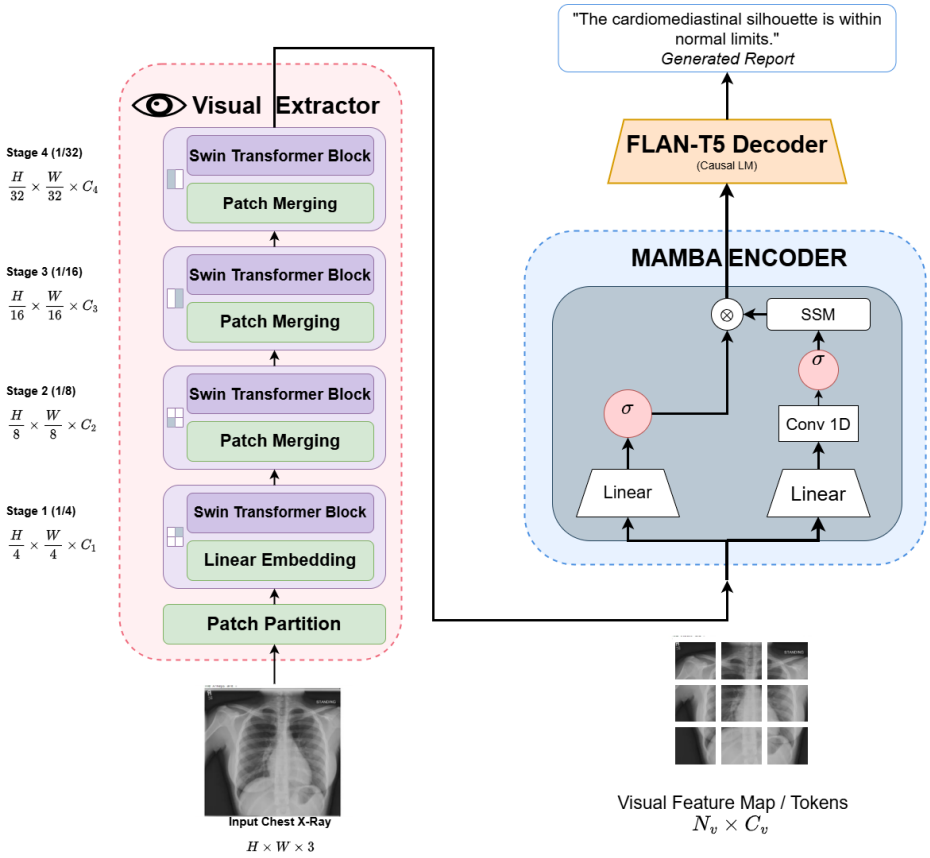


Fig. 1. SWINMambaT5 Architecture.

Visual Extractor The visual extractor component takes a radiology image as input for feature extraction using Swin transformers. Unlike convolutional extractors, such as VGG or ResNet, which are limited to local receptive fields. The visual extractor employs shifted window self-attention (SWSA). The SWSA

captures both local patch-level features and long-range anatomical dependencies. The SWSA can produce anatomically dependent spatial features across the full image. These spatial tokens are flattened to obtain a sequence as the input for the Mamba encoder. To adapt the pretrained Swin encoder to the medical imaging domain, LoRA [7] adapters were applied.

$$\{I_1, I_2, \ldots, I_S\} = f_S(\text{Img}) \tag{1}$$

The process is formulated as 1 ,f_S is visual extraction. The Img is the input, and the $S = 49$ with each spatial token $I_i \in \mathbb{R}^d$.

Mamba Encoder The visual feature sequence $\{I_1, I_2, \ldots, I_S\}$ output from the visual extractor is input to the Mamba encoder[6]. The Mamba encoder model performs sequential contextual dependencies among spatial tokens. To overcome the quadratic $\mathcal{O}(n^2)$ computational complexity of the Transformer self-attention and capture the sequential contextual dependencies among visual tokens, Mamba was employed. Mamba utilizes a selective state space mechanism [6] that operates with linear $\mathcal{O}(n)$ complexity. The state space model (SSM) layer of Mamba allows the encoder to retain clinically relevant spatial relationships while selectively suppressing irrelevant ones. Equation 2 formulates the contextual feature encoding process.

$$\{h_1, h_2, \ldots, h_S\} = f_m(\{I_1, I_2, \ldots, I_S\}) \tag{2}$$

In equation 2, $f_m(\cdot)$ is the Mamba encoder, $I_i \in \mathbb{R}^d$ is the i-th input visual token, and $h_i \in \mathbb{R}^d$ is the corresponding contextually enriched output token. The output sequence $\{h_1, h_2, \ldots, h_S\}$ serves as the input to the T5 decoder.

Transformer Decoder The output sequence $\{h_1, h_2, \ldots, h_S\}$ from the Mamba encoder serves as the input to the text decoder via a trainable projection layer. Chen et al. [3] [2] choose the transformer text decoder to be trained from scratch. Chen et al. [3], Chen et al. [2] text decoder suffers from the problem of limited medical vocabularies. In our proposed approach, we employed the pretrained Flan-T5 [4] model as the text decoder. Flan-T5 provides broad linguistic and domain-agnostic knowledge acquired from large-scale pre-training and instruction tuning. Prior knowledge included exposure to general medical terminology. This preserves the pretrained language generation capability while adapting the decoder to visual inputs. All Flan-T5 parameters were frozen, except for the cross-attention layers. The cross-attention mechanism allows the decoder to attend to the visual feature sequence $\{h_1, h_2, \ldots, h_S\}$ at each decoding step. Equation 3 illustrates the text-decoder formula.

$$\{y_1, y_2, \ldots, y_T\} = f_{T5}(\{h_1, h_2, \ldots, h_S\}) \tag{3}$$

In equation 3, $f_{T5}(\cdot)$ is the T5 decoder, $h_i \in \mathbb{R}^d$ is the i-th contextual visual token from the Mamba encoder, and y_t is the t-th generated report token. The complete generated report $\{y_1, y_2, \ldots, y_T\}$ constitutes the final radiology report output of the SwinMambaT5.

3 Experiments

3.1 Dataset

SwinMambaT5 was experimentally verified on its pipeline using two benchmark datasets: IU X-Ray[5] and MIMIC-CXR [10].The chest X-ray dataset known as MIMIC-CXR is the most extensive dataset currently accessible. The experiment utilized the official division of the MIMIC-CXR dataset. The IU X-Ray dataset is the oldest dataset for generating chest X-ray reports. For the IU X-ray dataset, we followed the official split, as in the MIMIC-CXR dataset. To maintain consistency with previous experiments [3] [2] [12], we input two patient radiographs for the IU X-ray dataset and only one for the MIMIC-CXR dataset.

3.2 Implementation Details

All experiments were performed using a single NVIDIA RTX 3080 Ti GPU equipped with 12 GB of VRAM.. The visual extractor uses Swin-Tiny with a patch size $p = 4$, window size $w = 7$, and input resolution $H \times W = 224 \times 224$. The visual extractor produces $S = 49$ spatial tokens with dimensions $d = 768$. LoRA adapters are applied to the query, key, value, and proj modules. For LoRA, the rank $r = 16$ and the scaling factor $\alpha = 32$. The Mamba encoder consists of $n = 3$ MambaBlocks with hidden dimension $d = 768$, state dimension $d_{\text{state}} = 16$, convolutional kernel $d_{\text{conv}} = 4$, and expansion factor $e = 2$. The scale projection layer is initialized with Xavier uniform initialization with gain $= 0.1$ and zero bias. The Flan-T5-base decoder contains 12 decoder blocks with a vocabulary size of 32,211. In the vocabulary, 40 medical terms were augmented with domain-specific terms. All models were effectively trained using the AdamW optimizer, applying a consistent learning rate of $\eta = 5 \times 10^{-5}$ across all parameter groups. We utilized a ReduceLROnPlateau scheduler to decrease the learning rate by a factor of 0.8 after a patience period of three epochs. The minimum learning rate was firmly established at 10^{-6}. The gradient norms were clipped at 0.5. The training was run for a maximum of 30 epochs. In addition, early stopping, which was triggered after eight consistent epochs, was implemented. The batch size was set to 16, with a maximum target sequence length of 100 tokens for IU X-ray and 150 tokens for MIMIC-CXR. The model evaluation is similar to R2Gen [3], with the best validation BLEU-4 score. The total number of model parameters is $287M$, of which $40.1M$ (14.0%) were trainable. The total parameters comprise the Swin LoRA adapters $0.43M$, Mamba encoder $11.3M$, scale projection layer $0.59M$, and T5 cross-attention layers $28.3M$.

4 Results

We conducted a comparative analysis to assess the effectiveness of SwinMambaT5 against existing HRGR-Agents [12], R2Gen [3], R2Gen-CMN [2], and R2GenMamba [17]. For the evaluation of the text germinated reports, we follow

Table 1. Comparison of the SwimMambaT5 results with baselines on both datasets.

(a) MIMIC-CXR

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
R2Gen [3]	0.353	0.218	0.145	0.103	0.142	0.277
R2Gen-CMN [2]	0.353	0.218	0.148	0.106	0.142	0.278
R2Gen-Mamba [17]	0.352	0.222	0.152	**0.110**	0.141	**0.284**
SwinMambaT5	0.355	**0.223**	**0.153**	0.100	**0.264**	0.247

(b) IU X-Ray

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
R2Gen [3]	0.470	0.304	0.219	0.165	0.187	0.371
R2Gen-Mamba [17]	0.482	0.315	0.228	**0.176**	0.208	0.382
R2Gen-CMN [2]	0.475	0.309	0.222	0.170	0.191	0.375
SwinMambaT5	**0.483**	**0.318**	**0.228**	0.161	**0.404**	0.339

the same measure used by SOTA (State of the Art), BLEU [16], with the n-grams where n=1,2,3,4. We also evaluated using METEOR [1] and ROUGE-L [13]. Table 1 shows a comparison of the results of the SWinMambaT5 proposed models' results with those of the SOTA paper. Table 1 presents the results for both the MIMIC-CXR and IU X-Ray datasets. The first part of the table shows a comparison of SwinMambaT5 with SOTA models on the MIMIC-CXR dataset. The comparison revealed that SwinMambaT5 achieved the highest BLEU-1 with a value of 0.355. Similarly, SwinMabaT5 achieved the highest BLEU-2, BLEU-3, and METEOR with values 0.223,0.153 and 0.264 respectively. The SwimMambaT5 model, utilizing the IU X-Ray dataset, recorded the highest values for BLEU-1, BLEU-2, BLEU-3, and METEOR, which were 0.483, 0.318, 0.228, and 0.404, respectively. Nevertheless, as indicated in Table 1 the BLEU-4 and ROUGE-L scores for both datasets are still comparatively low. However, the higher METEORs scores strongly support that the generated text is not an exact match for n-grams and longest common subsequences, whereas they are synonymously similar. This concept is further strengthened by the output shown in Figure 2. The target text has the value "the cardiac and mediastinal silhouettes are unremarkable", the predicted text has the value "The cardiomediastinal silhouette was within normal limits." When comparing these two strings, it is observed that both are equal and paraphrased texts; such paraphrasing receives no credit under BLEU-4 or ROUGE-L but is captured by METEOR. Figure 2 shows the output of the quantitative on the MIMIC-CRX test sample. Table 2 shows the results of the ablation of the three components: the visual extractor, the Mamba encoder, and the Flan-T5 encoder. In the first ablation, we removed the Mamba encoder and replaced it with a linear encoder. The results show that for both datasets, BLEU-4 and ROUGE-L are less than SwimMambaT5. The result was

Table 2. Ablation study of SwinMambaT5 with both test sets.

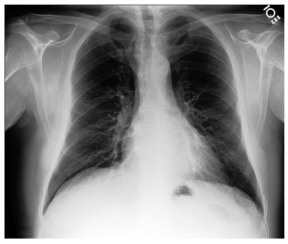
(a) MIMIC-CXR

Model / Variant	BL-4	RG-L
SwinMambaT5 (Proposed)	**0.100**	**0.247**
w/o Mamba (linear)	0.076	0.230
w/o Flan-T5 (customized decoder)	0.069	0.228
W/o Swin (ResNet-50)	0.099	0.224

(b) IU X-Ray

Model / Variant	BL-4	RG-L
SwinMambaT5 (Proposed)	**0.161**	**0.339**
w/o Mamba (linear)	0.160	0.333
w/o Flan-T5 (customized decoder)	0.157	0.331
w/o Swin (ResNet-50)	0.159	0.319

also lower when we ablated Swin with ResNet50 and when we ablated Flan-T5 with a customized decoder.



Predicted Report:
The lungs are clear without focal consolidation. No pleural Effusion or pneumothorax was observed. The cardiac and mediastinal silhouettes are unremarkable.

Ground Truth Report:
The lungs are clear. The cardiomediastinal silhouette is within normal limits. Hypertrophic changes seen the spine, no acute osseous abnormalities.

Fig. 2. Qualitative output of SwinMambaT5 original chest X-ray, ground truth text and generated text .

5 Conclusion

In this study, we propose SwinMambaT5, a sequential pipeline for the automated generation of chest X-ray reports. SwinMambaT5 combines a Swin Transformer visual encoder with LoRA adaptation, a Mamba state space model for sequential visual feature encoding, and a pretrained Flan-T5 decoder. SwinMambaT5 was trained on only 14% of the total model parameters. While achieving METEOR improvements of 87% and 94% over R2Gen-Mamba on MIMIC-CXR and IU

X-Ray, respectively. Ablation studies confirmed the independent contributions of the Swin, Mamba encoder, and Flan-T5 models. Despite these improvements, the BLEU-4 and ROUGE-L scores remained below those of R2Gen-Mamba. The results likely reflect clinically equivalent but lexically diverse report generations. Future studies should incorporate such clinical efficacy metrics. We also extended SwinMambaT5 to larger pretrained decoders using parameter-efficient techniques such as QLoRA. In the future, we will test SwinMambaT5 on a real dataset.

References

1. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
2. Chen, Z., Shen, Y., Song, Y., Wan, X.: 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). pp. 5904–5914 (2021). https://doi.org/10.18653/v1/2021.acl-long.459
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 1439–1449 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.112
4. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024). https://doi.org/10.48550/arXiv.2210.11416
5. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (03 2016). https://doi.org/10.1093/jamia/ocv080
6. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022). https://doi.org/10.48550/arXiv.2106.09685
8. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 2607–2615 (2024). https://doi.org/10.48550/arXiv.2308.12604
9. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586 (2018)
10. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR-JPG - chest radiographs with structured labels. PhysioNet (Mar 2024). https://doi.org/10.13026/jsn5-t979, version 2.1.0

11. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019). https://doi.org/10.6084/m9.figshare.10303823
12. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems **31** (2018). https://doi.org/10.48550/arXiv.1805.08298
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: Machine Learning for Healthcare Conference. pp. 249–269. PMLR (2019)
15. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023). https://doi.org//10.1016/j.artmed.2023.102633, https://www.sciencedirect.com/science/article/pii/S0933365723001471
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135
17. Sun, Y., Lee, Y.Z., Woodard, G.A., Zhu, H., Lian, C., Liu, M.: R2gen-mamba: A selective state space model for radiology report generation. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2025). https://doi.org/10.1109/ISBI60581.2025.10980814
18. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023). https://doi.org//10.1016/j.metrad.2023.100033, https://www.sciencedirect.com/science/article/pii/S2950162823000334