Diff-RRG: Longitudinal Disease-wise Patch Difference as Guidance for LLM-based Radiology Report Generation

Hannah Yun[1][0009−0006−3053−5294], Junyeong Maeng[1][0009−0001−2939−8564], Eunsong Kang[2][0009−0007−3010−5144], and Heung-Il Suk[1⋆][0000−0001−7019−8962]

[1] Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2] Department of Data Science, Kangwon National University, Gangwon State, Republic of Korea
{yunhh20, mjy8086, hisuk}@korea.ac.kr; eskang@kangwon.ac.kr

Abstract. Radiology report generation (RRG) is an emerging field that aims to automatically generate free-text clinical descriptions of radiographic images, incorporating temporal disease progression. However, existing methods rely on coarse-grained image representations and lack explicit mechanisms to integrate patients' historical information. To address these limitations, we propose a novel framework **Diff-RRG** that introduces longitudinal disease-wise patch **Diff**erence as guidance for large language model (LLM)-based **R**adiology **R**eport **G**eneration, aligning with the real-world diagnostic process. Our approach extracts disease-wise difference maps to identify fine-grained patches associated with specific diseases and to capture the difference between consecutive radiographs. Such information is fed into the LLM to provide direct guidance on disease progression. Accordingly, the resulting generated reports can be explained by pinpointing the related regions in the image, thereby enhancing explainability. In the extensive experiments, we have achieved state-of-the-art performance in most of the natural language generation and clinical efficacy metrics on the Longitudinal-MIMIC dataset. Our code is available at https://github.com/ku-milab/Diff-RRG.

Keywords: Radiology Report Generation · Longitudinal Data · LLMs

1 Introduction

Automated radiology report generation (RRG) refers to synthesizing free-text clinical descriptions of radiographic images, aiming to streamline radiologists' workflow by reducing labor-intensive, repetitive, and error-prone tasks. In clinical practice, radiologists document disease presence and progression by comparing current findings with previous records, often utilizing comparative language such as "improving atelectasis and decreasing pleural effusion." Despite the inherently comparative nature of radiology reports, most existing RRG models [2,3,10,15,16] rely solely on single input images, disregarding the patient's

[⋆] Corresponding author.

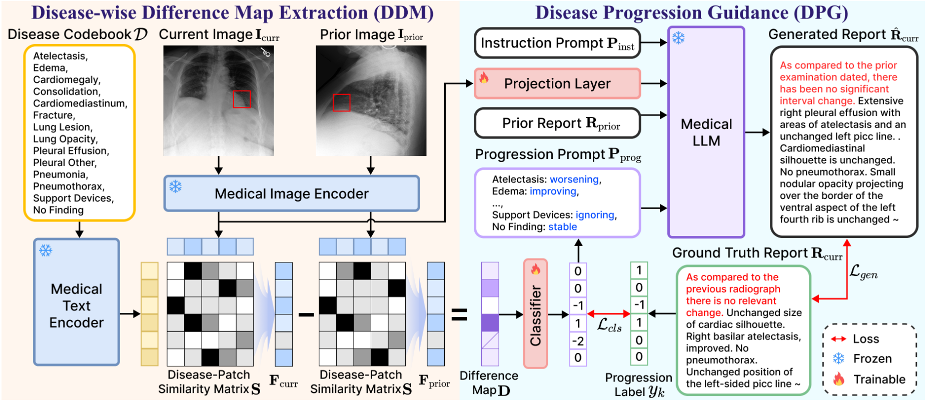
historical information. This limitation not only increases the risk of hallucinations but also restricts the model from incorporating the comprehensive clinical context available to radiologists, thereby precipitating a substantial knowledge gap between automated systems and diagnostic experts [8].

To overcome the limitations of traditional RRG models, recent studies [8,14,18] have focused on leveraging longitudinal data, incorporating current and previous images and reports. These efforts strive to produce context-aware reports, a critical aspect in monitoring patient's medical conditions over time. For instance, Prefilling [18] integrates multi-modal data through a cross-attention module and a hierarchical memory-driven transformer, effectively reducing hallucinations by incorporating historical information. HERGen [14] employs a group causal transformer to process entire patient-level images and utilizes auxiliary contrastive alignment, thereby narrowing the knowledge gap between the model and diagnostic experts. HC-LLM [8] introduces temporal constraints to differentiate between consistent (*e.g.*, time-shared) and evolving (*e.g.*, time-specific) disease patterns, which in turn indirectly guides the large language model (LLM). However, despite these significant advancements, they still face two major shortcomings: 1) dependence on coarse-grained image representations, which struggle to capture disease-specific fine-grained pathological details for precise disease analysis, and 2) lack of direct guidance for disease progression, thus limiting the ability to detect critical clinical changes.

To address these challenges, we propose a novel framework **Diff-RRG** that introduces longitudinal disease-wise patch **Diff**erence as guidance for LLM-based **R**adiology **R**eport **G**eneration. Our method comprises two key components: Disease-wise Difference Map extraction (DDM) and Disease Progression Guidance (DPG). The DDM generates disease-wise difference maps by analyzing differences between current and prior chest X-ray images at the patch level, allowing the model to capture fine-grained spatial details in disease-related regions. These detailed difference maps serve as the foundation for the DPG, which provides explicit temporal guidance for report generation. By inferring disease progression states, the DPG ensures that the generated reports preserve clinical context while accurately capturing the temporal dynamics of disease evolution. Together, these components enable Diff-RRG to bridge the gap between automated systems and radiologists, closely aligning with the radiologist's workflow, which prioritizes the identification of progressing pathological lesions. Experimental results on the Longitudinal-MIMIC dataset demonstrate superior performance in both linguistic quality and clinical accuracy while improving explainability by visualizing disease-relevant regions.

Our contributions are summarized as follows:

1. We propose a novel longitudinal RRG framework, named Diff-RRG, designed to align with the diagnostic process in real clinical scenarios by generating fine-grained and progression-focused reports.
2. We devise the DDM module that extracts disease-wise difference maps, enabling the model to capture localized pathological variations.


Disease-wise Difference Map Extraction (DDM)
Disease Codebook $\mathcal{D}$
Current Image $\mathbf{I}_{\text{curr}}$
Prior Image $\mathbf{I}_{\text{prior}}$
Atelectasis, Edema, Cardiomegaly, Consolidation, Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices, No Finding
Medical Image Encoder
Medical Text Encoder
Disease-Patch Similarity Matrix $\mathbf{S}$
$\mathbf{F}_{\text{curr}}$
Disease-Patch Similarity Matrix $\mathbf{S}$
$\mathbf{F}_{\text{prior}}$
Difference Map $\mathbf{D}$
Disease Progression Guidance (DPG)
Instruction Prompt $\mathbf{P}_{\text{inst}}$
Projection Layer
Prior Report $\mathbf{R}_{\text{prior}}$
Progression Prompt $\mathbf{P}_{\text{prog}}$
Atelectasis: worsening,
Edema: improving,
...,
Support Devices: ignoring,
No Finding: stable
Medical LLM
Generated Report $\hat{\mathbf{R}}_{\text{curr}}$
Classifier
0
0
-1
1
-2
0
$\mathcal{L}_{cls}$
1
0
-1
1
0
0
Progression Label y_k
Ground Truth Report $\mathbf{R}_{\text{curr}}$
$\mathcal{L}_{gen}$
Loss
Frozen
Trainable

As compared to the prior examination dated, there has been no significant interval change. Extensive right pleural effusion with areas of atelectasis and an unchanged left picc line. . Cardiomediastinal silhouette is unchanged. No pneumothorax. Small nodular opacity projecting over the border of the ventral aspect of the left fourth rib is unchanged ~

As compared to the previous radiograph there is no relevant change. Unchanged size of cardiac silhouette. Right basilar atelectasis, improved. No pneumothorax. Unchanged position of the left-sided picc line ~

Fig. 1. An overview of our proposed Diff-RRG framework.

3. We introduce the DPG module, which provides explicit disease progression guidance, allowing the model to generate clinically meaningful reports.
4. Experiments on the Longitudinal-MIMIC dataset demonstrate the effectiveness of our proposed method, and the visualization of disease-related patches enhances explainability by highlighting localized pathological changes.

2 Method

In this section, we introduce the overall framework of our proposed model and the two modules, namely DDM and DPG, and then outline the radiology report generation process.

2.1 Overall Framework

The overall framework of Diff-RRG is illustrated in Fig. 1. The primary objective of our proposed method is to effectively capture the disease progression across serial chest X-ray images and generate a clinically accurate diagnostic report considering the historical information. To achieve this, we utilize a patient's prior image, prior report, and current image to generate a corresponding current report that closely resembles the ground truth report. The report generation process is formulated as follows:

$$\hat{\mathbf{R}}_{\text{curr}} := \text{Diff-RRG}(\mathbf{I}_{\text{curr}}, \mathbf{I}_{\text{prior}}, \mathbf{R}_{\text{prior}}), \tag{1}$$

where $\mathbf{I}_{\text{curr}}, \mathbf{I}_{\text{prior}} \in \mathbb{R}^{C \times H \times W}$ denote the current and prior chest X-ray images, respectively, while $\mathbf{R}_{\text{prior}}$ represents the prior radiology report and $\hat{\mathbf{R}}_{\text{curr}}$ refers to the predicted current report.

2.2 Disease-wise Difference Map Extraction (DDM)

Accurately identifying disease progression in radiology requires capturing fine-grained pathological changes rather than relying solely on global image features. Therefore, we design the DDM to extract a difference map by comparing radiographic images taken at different time points, enabling patch-level analysis of disease evolution. The DDM consists of an image encoder, text encoder, and disease-wise patch selection. Specifically, we adopt a pre-trained CLIP [12]-based model for image and text encoders to extract aligned representations of medical images and disease texts.

Image Encoder As a robust image feature extractor, the image encoder captures fine-grained visual patterns crucial for disease identification and progression analysis. Given an input X-ray image $\mathbf{I}$ (both current and prior), the image encoder $f_{\text{ie}}(\cdot)$ extracts patch-wise feature embeddings, producing $\mathbf{V} = f_{\text{ie}}(\mathbf{I})$, where $\mathbf{V} \in \mathbb{R}^{p \times d}$ denotes the patch-wise visual representations, p denotes the number of patches, and d corresponds to the size of the embedding dimension.

Text Encoder Serving as a medical terminology feature representation extractor, the text encoder $f_{\text{te}}(\cdot)$ encodes the pre-defined disease codebook $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, which comprises the names of n diseases. The encoder outputs the disease embedding $\mathbf{C} = f_{\text{te}}(\mathcal{D})$, where $\mathbf{C} \in \mathbb{R}^{n \times d}$ is the corresponding text embedding matrix.

Disease-wise Patch Selection To quantify the association between disease categories and image patches, we compute the similarity matrix $\mathbf{S} \in \mathbb{R}^{n \times p}$ between the patch-wise visual representation $\mathbf{V}$ and the disease embedding matrix $\mathbf{C}$, as follows [17]. Subsequently, we utilize the Gumbel-Softmax [4] to dynamically select the most relevant patches for each disease. Consequently, the number of selected patches varies across patients and diseases.

We construct a mask matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$, where each element is set to 1 if the corresponding value in the Gumbel-Softmax output exceeds a threshold δ, and 0 otherwise:

$$\mathbf{M}_{i,j} = \mathbb{1}(\text{Gumbel-Softmax}((\mathbf{S}_{i,j}) > \delta)), \tag{2}$$

where $\mathbb{1}(\cdot)$ is the indicator function, i corresponds to the index of the disease codebook, and j denotes the index of image patches. The mask $\mathbf{M}$ is then applied to the patch embedding $\mathbf{V}$ via matrix multiplication, ensuring that only the selected patches contribute to the final disease-wise representation. Given that the number of selected patches varies, we compute the mean of the retained patch embeddings to obtain a fixed-dimensional representation for each disease:

$$\mathbf{F} = (\mathbf{M} \cdot \mathbf{V}) \odot \mathbf{s}^{-1} + \mathbf{E}_{\text{pos}}, \quad s.t. \quad \mathbf{s}_i = \textstyle\sum_j \mathbf{M}_{i,j} \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{n \times d}$ represents disease-wise selected patch embeddings, $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{n \times d}$ is the positional encoding for each disease, $\odot$ denotes element-wise multiplication, and $\mathbf{s}^{-1}$ is the reciprocal of vector $\mathbf{s}$, broadcasted to match the dimensions of $\mathbf{M} \cdot \mathbf{V}$. The computation is performed only for diseases with at least one selected patch ($s_i \neq 0$), while others are ignored. At this stage, the selected patches correspond to potential regions where each disease may manifest. This highlights the model's attention to pathological regions, enhancing explainability.

The process is identically applied to both the current and prior images, ultimately yielding the respective representations $\mathbf{F}_{\text{curr}}, \mathbf{F}_{\text{prior}}$. Finally, the disease-wise difference map $\mathbf{D} \in \mathbb{R}^{n \times d}$ is computed as:

$$\mathbf{D} = \mathbf{F}_{\text{curr}} - \mathbf{F}_{\text{prior}}. \tag{4}$$

2.3 Disease Progression Guidance (DPG)

Building upon the extracted disease-wise difference map, DPG performs classification to guide the large language model (LLM) by providing explicit disease progression information. This module comprises disease progression label construction and disease progression classification.

Disease Progression Classification To classifiy disease progression, we first construct disease progression labels y_k to encode changes in each disease by comparing prior and current images derived from ground truth disease annotations. Each disease is classified into one of three progression states: "worsening", "stable", and "improving". Additionally, if a disease does not have any selected patches in either prior or current images (*i.e.*, $\mathbf{s}_i = 0$), it is excluded from classification and assigned an "N/A" label.

The disease-wise difference map $\mathbf{D}$ is fed into a progression classifier, which predicts a disease progression label for each disease. The classification loss $\mathcal{L}_{\text{cls}}$ is optimized using a cross-entropy loss function as follows:

$$\mathcal{L}_{\text{cls}} = -\sum_{k=1}^{n} y_k \log(\hat{y}_k), \tag{5}$$

where $\hat{y}_k$ denotes the predicted probability distribution for k-th disease.

The predicted disease progression labels are then combined with the corresponding disease names to form a progression prompt $\mathbf{P}_{\text{prog}}$, which serves as explicit guidance for the text decoder.

2.4 Radiology Report Generation

With the direct progression guidance, our model integrates the current image representation and historical information from the prior report. For report generation, we leverage an LLM, inspired by its remarkable ability to generate medical reports and effectively capture the temporal context [8,15].

The report generation process follows an autoregressive formulation, where the language modeling loss $\mathcal{L}_{\text{gen}}$ is defined as the summation of the negative log-likelihood of each token r_t:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log p(r_t | r_1, \ldots, r_{t-1}, \mathbf{P}_{\text{inst}}, \mathbf{V}_{\text{curr}}, \mathbf{R}_{\text{prior}}, \mathbf{P}_{\text{prog}}). \tag{6}$$

Here, $\mathbf{P}_{\text{inst}}$ refers to the instruction prompt, $\mathbf{V}_{\text{curr}}$ represents the current image feature embeddings, and T denotes the total number of generated tokens.

Finally, the total training objective of the model is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda \mathcal{L}_{\text{cls}}, \tag{7}$$

where λ denotes the balancing coefficient. By combining the disease progression classification loss with the report generation loss, our approach ensures that the generated report accurately reflects disease progression in serial radiographs.

3 Experiments and Results

Dataset and Evaluation Metrics We utilize the publicly available Longitudinal-MIMIC dataset [18], derived from the MIMIC-CXR dataset [5], to train and evaluate our model. The dataset includes 26,625 patients and a total of 94,169 samples with at least two visit records. Each sample consists of the current image, the current report, the previous image, and the previous report. We follow the official split for dividing training, validation, and test sets.

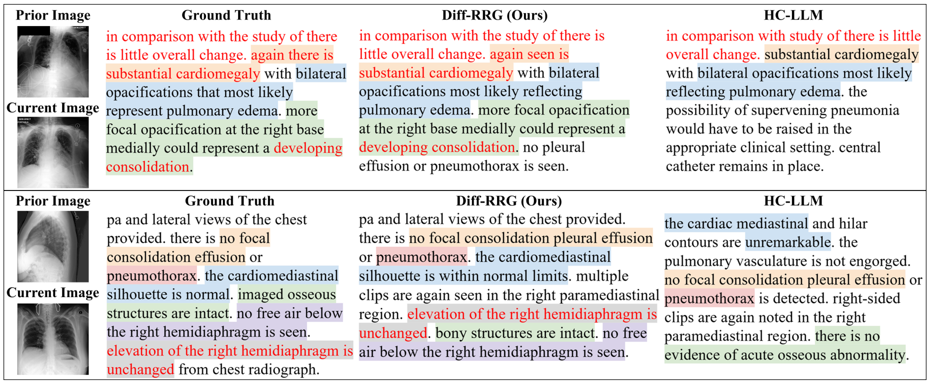
To assess the quality of our generated reports, we evaluate both Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics. For NLG, we use BLEU-n [11], METEOR [1], and ROUGE-L [7]. For CE, we compute micro-averaged Precision, Recall, and F1-score utilizing the CheXbert labeler [13].

Implementation Details We employ pre-trained BiomedCLIP [17] as both the image and text encoders and BioMistral-7B [6] as the text decoder. In the Gumbel-Softmax function, the temperature τ and threshold δ are set to 0.3 and 0.2, respectively. The coefficient λ is set to 0.1, and we leverage the AdamW [9] optimizer with an initial learning rate of 1e-4. The model was trained for 5 epochs with a mini-batch size of 6 on a single NVIDIA RTX A6000 48GB GPU.

Quantitative Results Table 1 presents a quantitative evaluation of our model against existing models using NLG and CE metrics. To begin with, we compare our method against conventional methods using a single image as an input, including R2Gen [3], R2GenCMN [2], CvT2DistilGPT2 [10], R2GenGPT [15], and GMoD [16]. Our model consistently outperforms all single-input approaches in both NLG and CE metrics, and these results substantiate the importance of utilizing temporal information to enhance generated reports' linguistic quality and clinical accuracy. Furthermore, we compare our model with recent longitudinal methods, Prefilling [18], HERGen [14], and HC-LLM [8], which utilize the same longitudinal data. Our approach achieves superior performance in most

Table 1. Evaluation on the Longitudinal MIMIC-CXR dataset. † indicates the results are cited from the original papers.

Model	NLG metrics						CE metrics		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	Precision	Recall	F1-Score
Single Input									
R2Gen [3]	0.308	0.190	0.126	0.089	0.266	0.124	0.457	0.290	0.355
R2GenCMN [2]	0.328	0.203	0.135	0.096	0.273	0.133	0.512	0.359	0.422
CvT2DistilGPT2 [10]	0.363	0.225	0.151	0.109	0.266	0.138	0.485	0.310	0.378
R2GenGPT [15]	0.345	0.202	0.128	0.087	0.240	0.125	0.331	0.253	0.287
GMoD [16]	0.378	0.234	0.155	0.107	0.276	0.162	0.496	0.429	0.460
Longitudinal Input									
Prefilling [18]	0.343	0.210	0.141	0.100	0.274	0.137	0.506	0.364	0.423
HERGen† [14]	0.389	0.242	0.163	0.117	**0.282**	0.155	0.421	0.289	0.295
HC-LLM [8]	0.404	0.247	0.164	0.116	0.271	0.163	0.488	0.415	0.448
Ours									
Baseline	0.390	0.231	0.150	0.104	0.262	0.146	0.434	0.353	0.389
+ Prior Image, Report	0.397	0.242	0.160	0.113	0.272	0.161	0.510	0.407	0.453
w/ DDM	0.402	0.248	0.167	0.119	0.275	0.163	0.518	0.429	0.469
w/ DPG	0.401	0.246	0.165	0.117	0.272	0.161	0.516	0.414	0.459
Diff-RRG	**0.405**	**0.251**	**0.169**	**0.120**	0.276	**0.164**	**0.528**	**0.430**	**0.474**

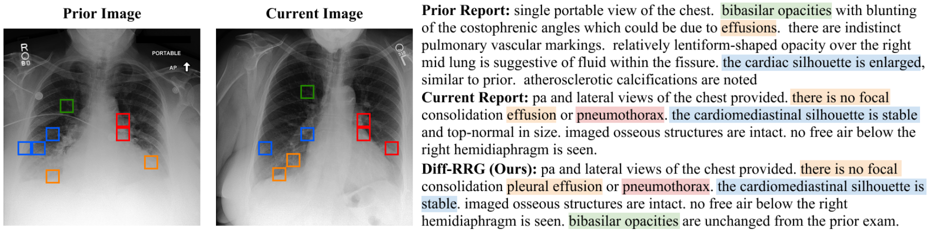


Prior Image Current Image	Ground Truth	Diff-RRG (Ours)	HC-LLM
	in comparison with the study of there is little overall change. again there is substantial cardiomegaly with bilateral opacifications that most likely represent pulmonary edema. more focal opacification at the right base medially could represent a developing consolidation.	in comparison with the study of there is little overall change. again seen is substantial cardiomegaly with bilateral opacifications most likely reflecting pulmonary edema. more focal opacification at the right base medially could represent a developing consolidation. no pleural effusion or pneumothorax is seen.	in comparison with study of there is little overall change. substantial cardiomegaly with bilateral opacifications most likely reflecting pulmonary edema. the possibility of supervening pneumonia would have to be raised in the appropriate clinical setting. central catheter remains in place.
Prior Image Current Image	Ground Truth	Diff-RRG (Ours)	HC-LLM
	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 pleural effusion or pneumothorax. the cardiomediastinal silhouette is within normal limits. multiple clips are again seen in the right paramediastinal region. elevation of the right hemidiaphragm is unchanged. bony structures are intact. no free air below the right hemidiaphragm is seen.	the cardiac mediastinal and hilar contours are unremarkable. the pulmonary vasculature is not engorged. no focal consolidation pleural effusion or pneumothorax is detected. right-sided clips are again noted in the right paramediastinal region. there is no evidence of acute osseous abnormality.

Fig. 2. Qualitative analysis of the generated reports. The red text denotes comparative statements between prior and current images, while the colored highlights indicate descriptions of distinct disease entities.

cases, demonstrating the effectiveness in capturing disease progression and accurately reflecting temporal context. These results further validate the capability of our method to generate clinically meaningful reports that align with real-world radiological assessments.

Qualitative Results To qualitatively assess the effectiveness of our proposed method, we compare the generated reports with the ground truth and outputs from the HC-LLM [8], as shown in Fig. 2. In two cases, our method not only accurately captures the historical aspects of the patient's condition but also provides a detailed and correct description of the presence or absence of diseases. In contrast, HC-LLM fails to mention certain diseases or provide comparative



Prior Report: single portable view of the chest. bibasilar opacities with blunting of the costophrenic angles which could be due to effusions. there are indistinct pulmonary vascular markings. relatively lentiform-shaped opacity over the right mid lung is suggestive of fluid within the fissure. the cardiac silhouette is enlarged, similar to prior. atherosclerotic calcifications are noted

Current Report: pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomediastinal silhouette is stable and top-normal in size. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.

Diff-RRG (Ours): pa and lateral views of the chest provided. there is no focal consolidation pleural effusion or pneumothorax. the cardiomediastinal silhouette is stable. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen. bibasilar opacities are unchanged from the prior exam.

Fig. 3. Visualization of disease-relevant patches extracted by the DDM module. The colored bounding boxes in the X-ray images represent the most informative patches identified by our model, while the corresponding highlighted texts in the reports indicate disease-relevant descriptions.

descriptions between the prior and current images. These results clearly illustrate the superior performance of our model, which effectively integrates historical data to offer comprehensive and accurate disease descriptions, improving the overall diagnostic quality.

Ablation Study Table 1 shows an ablation study of each component. First, we implement a baseline model with a BiomedCLIP [17] image encoder and a BioMistral-7B [6] text decoder, processing a single input image. Next, we compare this baseline against a model incorporating longitudinal multi-modal inputs such as prior images and reports that improve performance, demonstrating the benefits of historical context in report generation. To assess individual contributions, we evaluate the DDM and DPG modules separately. The DDM module indicates a substantial improvement in both NLG and CE metrics, underscoring the importance of extracting fine-grained difference maps. Similarly, integrating the DPG module yields improvements across most evaluation metrics, generating progression-aware reports. Overall, our proposed Diff-RRG model achieves the highest NLG and CE performance, validating their complementary effectiveness between components.

Model Explainability To enhance the explainability of our proposed method, we visualize the disease-wise extracted patches in the DDM module, as illustrated in Fig. 3. The extracted patches effectively identify disease-relevant regions across consecutive images, demonstrating the model's ability to track disease progression at a fine-grained level. Notably, our model consistently identifies pathological regions even when the disease manifestation shifts in location. Furthermore, the highlighted texts in the reports illustrate how these extracted visual features directly contribute to generating disease-relevant descriptions, establishing a clear connection between the visual findings and textual reporting. Such explainability further supports the clinical utility of our framework by providing transparent and interpretable predictions.

4 Conclusion

In this paper, we devise a novel radiology report generation framework, named Diff-RRG, designed to capture disease progression by leveraging longitudinal multi-modal information. We further propose the DDM and DPG modules to extract disease-wise difference maps and provide explicit disease progression guidance to the LLM decoder. Furthermore, the identified disease-related patches enhance the explainability of the model. Experimental results on the Longitudinal-MIMIC dataset demonstrate the superiority of our approach, bridging the gap between automated report generation and real-world clinical processes. This enables the generation of clinically meaningful reports, contributing to improved clinical decision-making.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University)) and No. RS-2022-II220959 ((Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

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. pp. 5904–5914 (2021)
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. pp. 1439–1449 (2020)
4. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (2017)
5. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.1.0) (2024)
6. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.a., Rouvier, M., Dufour, R.: BioMistral: A collection of open-source pretrained large language models for medical domains. In: Findings of the Association for Computational Linguistics. pp. 5848–5864 (2024)
7. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
8. Liu, T., Wang, J., Hu, Y., Li, M., Yi, J., Chang, X., Gao, J., Yin, B.: HC-LLM: Historical-constrained large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)

9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
10. Nicolson, A., Dowling, J., Koopman, B.: Improving chest X-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023)
11. 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)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
13. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: CheXbert: 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. pp. 1500–1519 (2020)
14. Wang, F., Du, S., Yu, L.: HERGen: Elevating radiology report generation with longitudinal data. In: European Conference on Computer Vision. pp. 183–200 (2024)
15. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology report generation with frozen llms. Meta-Radiology **1**, 100033 (2023)
16. Xiang, Z., Cui, S., Shang, C., Jiang, J., Zhang, L.: GMoD: Graph-driven momentum distillation framework with active perception of disease severity for radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 295–305 (2024)
17. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2** (2024)
18. Zhu, Q., Mathai, T.S., Mukherjee, P., Peng, Y., Summers, R.M., Lu, Z.: Utilizing longitudinal chest X-rays and reports to pre-fill radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–198 (2023)