# Improving the CXR Reports Generation with Multi-modal feature Alignment and Self-Refining strategy

Fatima CHEDDI
Smart Systems Laboratory
ENSIAS, Mohammed V University in Rabat, Rabat, Morocco
Fatima.chedi@gmail.com

Ahmed HABBANI
Smart Systems Laboratory
ENSIAS, Mohammed V University in Rabat, Rabat, Morocco
ahmed.habbani@ensias.um5.ac.ma

Hammadi Nait-Charif
National Center for Computer Animation
Bournemouth University
United Kingdom
hncharif@bournemouth.ac.uk

***Abstract*—Medical chest X-ray images are necessary for the diagnosis of different diseases. The need for automated interpretation and report generation of these images is crucial. It not only saves radiologists' time but also minimizes the risk of diagnostic mistakes. However, several challenges impede this task due to the employment of uni-directional image-to-report in the encoder-decoder deep learning model and the absence of contextual details in the process of report generation can lead to incomplete or inaccurate descriptions. To address this, we propose an approach based on Multi-modal feature Alignment and Self-Refining mechanism RG-MASR in order to generate an improved medical report from chest X-ray images automatically. Our method comprises three modules: visual and textual characteristics extraction to extract the semantic characteristics from X-ray images and their paired reports. Second, a multimodal feature alignment module is employed to leverage both textual and visual features. Finally, we integrate a self-refining technique in the report generator module to refine alignment and improve the output to generate a comprehensive report. We evaluate our method on the IU X-ray and NIH public chest X-ray datasets. The results demonstrate that our proposed RG-MASR surpasses existing approaches in terms of ROUGE and BLEU metrics.**

***Keywords—CXR report generation, Multi-modal alignment, medical image, Deep learning, Medical report, Transformers***

## I. INTRODUCTION

The interpretation of medical images using the techniques of machine learning has revolutionized the field of radiology, offering significant improvements in analyzing complex cases in radiographic images (x-rays). Recently, there have been significant advancements in developing automated methods for generating diagnosis reports based on medical images. These reports are essential for patient assessment and clinical evaluation. Thereby assisting radiologists in making more informed decisions.

The generation of detailed and accurate reports from chest X-ray images are a critical mission, offering important insights for identifying and managing various medical situations. Despite the advancements in medical technologies, the manual interpretation and reporting by radiologists remain a process and subject to variation due to differing individual styles and terminologies[1].

Automated report generation based on machine learning can significantly decrease the time needed for doctors to analyze and interpret medical images.

Recent developments in machine learning have facilitated automated report generation. The Initial approaches use an encoder-decoder model [2]. Furthermore, recent improvements have integrated transformers [3][4] to enhance these methods using multi-head attention and self-attention; these aid in focusing on relationships between nodes in the input data and capturing long-range semantic associations.

Multiple studies demonstrate the effectiveness of transformers in automatic report-generation applications [5][6]. However, these models still have limitations in accurately identifying and describing periodic and difficult diseases due to biased interpretation of specialized medical vocabulary and lesion characteristics. Furthermore, the generated reports require professionalism and flexibility, making the reports difficult to interpret and reducing their practical usability in clinical contexts.

To resolve these issues, we introduce a method named RG-MASR based on Multi-modal feature Alignment and Self-Refining mechanism. Our RG-MASR model provides the following contributions :

- We combine the power of multi-modal analysis by using advanced deep learning models to give chest X-ray images along with the ground truth diagnosis report as input to our system. This method allows the model to capture complex patterns between image and text, ensuring that the generated text is aligned with medical knowledge.
- We integrate into our system a multi-modal alignment-based attention mechanism to enable the alignment

**979-8-3315-2205-6/24/$31.00 ©2024 IEEE**

between hierarchical CXR Features Extraction and the report feature.

- We integrate a Self-Refining strategy into RG-MASR providing iterative model feedback and refinement to improve the initial output.
- We also present the results of experiments applied to the IU X-ray and NIH datasets, demonstrating that our RG-MASR surpasses existing approaches across several evaluation metrics. Also based on experiment results, our method shows its ability to produce relevant reports, potentially easing the burden on radiologists and improving diagnostic workflows.

## II. Related Work

In this section, we provide a summary of current automatic CXR report generation methods. A basic method in this field often employs CNN-RNN architectures [7]. To focus on more relevant data, CNN-RNN architecture with an attention layer [8] was introduced to synthesize multi-view visual features. Subsequent research used Transformer architectures [4], which leverage self-attention mechanisms. This is followed by the integration of reinforcement learning, contrastive learning, and memory-driven [9] techniques to enhance report generation. For instance, Zhang et al. [10] introduce a model embedded within a knowledge graph framework, integrating disease-specific attributes and relationships to enhance report accuracy and clinical relevance. They develop a graph convolutional neural network for structured learning from an existing knowledge graph, improving their predictions' interpretability and clinical relevance. Furthermore, Wang et al. propose a model named TMRGM, it uses a library of structured templates for normal cases and a multi-attention mechanism for abnormal reports so that it can produce reports more accurately and specifically by distinguishing the two types. Also, TMRGM integrates visual features with relevant textual information by combining co-attention, which learns a joint representation, and adaptive attention, which finds highly related information. Kaur and Mittal present the CADxReport model [12], a novel system using a co-attention and reinforcement learning framework. Using a VGG19 model trained separately, CADxReport uses the extracted visual features alongside a multi-label classifier-derived semantic feature that is fused with co-attention to generate more clinically useful reports. Further, CADxReport uses reinforcement learning optimization for CIDEr rewards to improve report completeness and correctness. Additionally, Shuxin Yang introduced M2KT [13], which is based on the concept of a learned knowledge base and its alignment with multiple modalities. For improving report quality, the multi-modal alignment module aligns textual embeddings to visual features and even disease annotations, thereby allowing for an integration of clinical semantics. Also, Liu et al. proposed a model named PPKED [14] to address common data biases in radiology report generation by mimicking the diagnostic cycles of radiologists. Their architecture consists of a PoKE module that provides information on rare anomalies, a PrKE module that integrates with previous radiology reports and knowledge learned in the scientific literature, along a Multi-domain Knowledge Distiller (MKD) for optimizing outputs. Li et al. proposed the HRGR [15], which uniquely combines retrieval and generation processes within a reinforcement learning framework. Through reinforcement learning, HRGR-Agent optimizes sentence- and word-level accuracy, producing well-structured reports that balance normal and abnormal findings. Chen et al.'s R2Gen [9] introduces a memory network-equipped transformer model to generate the final report capturing long-range dependencies of both images and texts. With the memory insertions, the model is able to store relevant information and pull from it when needed during the report generation process.

In recent times, novel methods have been developed to integrate and analyze multiple data modalities. For example, Aksoy et al. [16] introduce an integrative multi-modal approach, focusing on combining visual and textual data to enhance the coherence and accuracy of generated reports. This model overcomes limitations in unidirectional models by fostering bidirectional connections between images and reports. Similarly, Pan et al. [17] suggest cross-modal multi-scale feature fusion to blend visual and textual features at multiple scales for high-quality chest radiology report generation. This model cleverly aligns multi-scale features of chest X-ray images and associated reports using advanced fusion techniques. Also, Cheddi et al. [18] propose a multi-modal feature fusion-based method for automated report generation, MM-CXRG. This model uses visual and textual data using deep learning algorithms and multi-modal fusion using channel attention networks, offering a harmonized association of image and text features. The use of channel attention networks allows the model to automatically select the pertinent features from all modalities, improving its capability to generate accurate reports.

## I. Proposed Method

### A. System Overview

Fig. 1 presents our proposed method for generating chest X-ray reports. Our system is partitioned into three interdependent parts: First, a visual and textual features extraction module is employed to extract semantic information from both input images and text descriptions, representing them as multi-dimensional vectors. In this case, we employ a convolutional

neural network model, and Vision Transformers (ViT) to extract fine-grained and more contextual-level visual information from chest X-ray images, while Word2Vec is used to encode corresponding textual data into multi-dimensional vectors. Then, a Multi-modal Feature Alignment (MFA) module is based on a cross-attention mechanism to fuse information from the two modalities with attention weights, focusing on discriminative features. In the end, a Report Generator (RG) module is responsible for converting the concatenated multimodal features into accurate and appropriate contextualized textual reports.

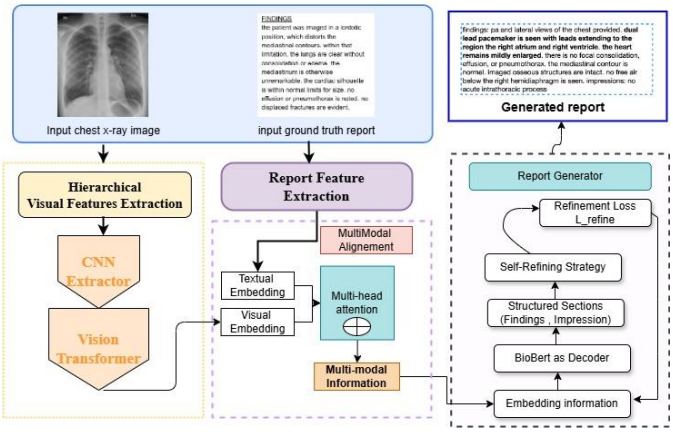


Fig. 1. The proposed Architecture

## B. *Hierarchical CXR Features Extraction*

We use a hierarchical visual extraction method that uses the convolutional neural network (ResNet-50 [19]) and vision transformers (ViTs) to get a visual embedding from chest X-ray (CXR) images. This method produces fine-grained and multi-level spatial and semantic information about the CXR image, which is very useful for analysis tasks that need local pattern recognition of the abnormalities and global structural patterns. Every CXR image undergoes resizing to 224x224 pixels as input to ensure consistency in feature extraction. Given an X-ray image I, ResNet-50 is used as the backbone feature extractor, processing the image through multiple convolutional layers while using residual connections to maintain gradient flow, allowing for the training of very deep networks (eq. 1). The output from ResNet's layers X' forms a set of hierarchical feature maps, each representing varying levels of abstraction within the CXR image. After the division of features obtained from The ResNet (eq. 2), each is treated as a "token" for input to the transformer. If the ResNet output feature map is h×w, each patch size p×p would divide the feature map into (h/p) × (w/p) tokens, preserving spatial information (eq. 3, eq. 4, eq. 5). In the transformer, self-attention mechanisms allow each token to interact with every other token, capturing interdependencies across patches. ViT builds a comprehensive representation of the CXR image, linking local with global contexts (eq. 6).

$$X' = f(Ix) \tag{1}$$

$$Patch_i = Reshape(X') \rightarrow E_i \tag{2}$$

- All patches are linearly transformed into an embedding vector:

$$E_i = Linear(Patch_i) \tag{3}$$

$$X_{patch}\text{=Linearize(X')} \tag{4}$$

- Positional Encoding (for Vision Transformer): Add positional encodings to the patch embeddings:

$$X_i = X_{patch} + PE_i \tag{5}$$

- Transformer Encoder (for Vision Transformer): Pass the combined patch embeddings through transformer encoder layers:

$$Z_{vit} = ViTransformer(X_i) \tag{6}$$

## C. *Ground-truth report Feature Extraction*

In this section, we employ the Word2Vec model to extract semantic features from reports where frequently co-occurring terms result in similar vectors. The skip-gram model predicts the context terms to learn text descriptions and captures the relatively detailed relationships between words that co-occur often in similar types of diagnostic descriptions. In the same tradition, the CBOW model predicts a word from the target context words, which is a useful property for representing the general semantic content of medical text. Each report R is then represented as a single vector by averaging its embeddings, providing a summary of the diagnostic context (eq. 7).

$$Z_t\text{= Word2vec(R)} \tag{7}$$

## D. *Multimodal Alignment Module*

In this module, we propose a multimodal alignment module that is able to concatenate visual features of chest X-ray images and textual features of clinical reports in a unified manner through a cross-attention technique and multi-head attention(Eq. 8). These mechanisms help the model to focalize on important details in the text based on the visual context, which is achieved by computing query, key, and value matrices for each modality. Output is a combined feature vector Zf that retains important image-text relationships. The transformer

decoder takes this boosted representation as input to generate accurate, semantically contextual, clinical decision-making diagnostic reports.

$$Z_f = \text{Fusion}\ (Z_v, Z_t) \quad (8)$$

### *E. Report Generator*

In this module, we use BioBert [20] as the decoder based on transformer architecture. The embeddings obtained by the previous module are processed inside BioBert, ordered, and fed into the attention part of the model. By using the trainable parameters, the attention directs the cross-attention layer of this module to attend only to important clinical features embedded in the embeddings. In particular, each of the Q vectors from the text produced attends to the K vectors in the embeddings to find relevant diagnostic information, while the V vectors provide content for each focused feature. We use the self-attention and encoder-decoder cross-attention mechanism of BioBert to obtain structured sections, where each section is separated by newlines (e.g., "Findings" and "Impressions") so that the generated text satisfies the clinical documentation standards.

$$R = BioBert(Z_f) \quad (9)$$

To enhance the result accuracy of the generated report and ensure better adaptation to data variations, such as the complex features in medical images, we introduce a self-refining mechanism.

### *F. Self Refining mechanism*

To refine alignment, we integrate a self-refinement technique [21], that iteratively improves the coherence output. As depicted in (Eq. 10) this involves leveraging feedback loops, uncertainty quantification, and domain-specific knowledge to evaluate and enhance initial drafts produced from multimodal inputs. His iterative refinement not only addresses inconsistencies but also improves the contextual relevance of the generated reports. This is achieved by minimizing a composite loss function:

$$\mathrm{L} = \lambda_1\ L_{align} + \lambda_2\ L_{report} \quad (10)$$

where $\mathbf{L_{align}}$ measures that errors and inconsistencies between the generated report and the input data are corrected, maintaining alignment with the original data, while $\mathbf{L_{report}}$ enhances the coherence and readability of the refined output. The **λ1** and **λ2** balance these components. This iterative refinement ensures that the final medical report is comprehensive, reliable, and actionable.

## II. EXPERIMENTAL SETUP

We evaluate, compare, and test our proposed approach on two datasets Indiana University [22] and NIH Chest X-ray [23]. In subsections, we offer a comprehensive description of these datasets, and the metrics utilized.

### *A. Datasets*

#### *1) Indiana University (IU) X-Ray dataset[22]*

Provides a valuable resource for medical research. It includes over 7,470 chest X-rays paired with detailed radiologist reports. This allows researchers to develop machine learning for analyzing X-rays and improve natural language processing for medical reports.

#### *2)* NIH Chest X-ray *[23]*:

Is a comprehensive benchmark released by the National Institutes of Health. Is a giant in medical imaging, offering a staggering 112,120 images from more than 30,000 patients. Each is labeled with a maximum of 14 distinct thoracic disease categories.

### *B. Evaluation metrics:*

In the phase of evaluation, we employ evaluation metrics, such as BLEU [24], which calculate the precision of the n-gram mismatch of the generated text and ground-truth; ROUGE-L[25] considers both recall (how much of the reference is captured) and precision (how relevant the generated text is) based on n-gram overlap. And METEOR[26], which considers precision, recall, synonymy, and word order. These metrics collectively ensure a comprehensive evaluation of semantic relevance and the linguistic accuracy of the generated texts.

## *III. Results and discussion*

In this part, we offer a detailed evaluation of the results obtained from our implementation, incorporating both quantitative and qualitative assessments. Finally, we are comparing our proposed method with some existing methods.

### *A. Qualitative Results*

Fig. 2 presents ground truth reports and generated reports for medical X-ray images applied on NIH and IU datasets.

The generated reports demonstrate a high degree of accuracy in capturing the essential findings and impressions of chest X-ray images from both the NIH and IU datasets. The generated text is slightly more concise, occasionally omitting less critical details, but it maintains the integrity and primary

diagnostic information of the original reports. This performance highlights the model's potential for effective automated report generation, aiding in the efficient and accurate interpretation of medical images.

| *Dataset & Medical X-ray image* | *Ground truth* | *Generated report* |
|---|---|---|
| NIH dataset  | **Findings**: The chest X-ray shows patchy infiltrates in the right lower lobe with air bronchograms, indicative of pneumonia. No pleural effusion or pneumothorax is observed. there is mild tortuosity to the descending thoracic aorta, The heart size appears normal. **Impression**: Right lower lobe pneumonia with patchy infiltrates and consolidation. No evidence of pleural effusion or pneumothorax. Normal heart size. | **Findings**: The chest X-ray shows patchy in the right lower lobe, suggesting the presence of pneumonia. There is no evidence of pleural effusion or pneumothorax. The heart size appears normal. **Impression**: Right lower lobe pneumonia with patchy infiltrates. Absence of pleural effusion or pneumothorax. Normal heart size. |
| IU dataset  | **Findings**: The PA chest radiograph shows a right lower lobe consolidation with air bronchograms. There is a mild right pleural effusion. The cardiac silhouette is within normal limits. No evidence of pneumothorax. The bony structures appear intact." **Impression:** "Right lower lobe pneumonia with mild pleural effusion. No pneumothorax. Heart size normal." | **Findings**: The PA chest X-ray demonstrates a right lower lobe consolidation characterized by air bronchograms. A mild right pleural effusion is also observed. The cardiac silhouette appears normal. No evidence of pneumothorax. The bony structures are unremarkable." Impression: "Right lower lobe pneumonia accompanied by a mild pleural effusion. No pneumothorax. Normal heart size |

Fig. 2. Visualization of medical report generation

## B. *Quantitative Results*

*Table I: performance results*

| Dataset | B. 1 | B. 2 | B.3 | R L | M | CIDER |
|---|---|---|---|---|---|---|
| NIH | 0.412 | 0.341 | 0.229 | 0.413 | 0.401 | ------ |
| IU XR | 0.486 | 0.360 | 0.262 | 0.379 | 0.205 | 0,432 |

Our proposed method employs BioBert as a decoder and is performed on NIH and IU datasets. To evaluate the performance of the RG-MASR approach, we utilize a suite of evaluation metrics that measure the semantic and syntactic alignment between real reports and the generated sentences. As illustrated in Table I, the scores are 0.486, 0.360, 0.262, 0.379, and 0.205 in B-1, B-2, B-3, R-L, and METEOR and CIDER, while being trained on the IU dataset. And achieve 0,412, 0,341, 0.229, 0.413, and 0.401 on the NIH dataset. These results highlight the model's ability to produce reports that are both syntactically accurate and semantically relevant.

## C. *Comparison and discussion:*

In this section, we present a detailed comparison and discussion of our approach for generating CXR reports using multi-modal feature fusion with existing methods. The performance of different methods is evaluated on the IU X-ray dataset and the NIH dataset using evaluation metrics, including B-1, B-2, B-3, R-L, and METEOR. To achieve this, the models PPKED, HRGR, R2GEN, VFE+GK+SKE+RG, and CNX-B2 are chosen for comparison. The results of this comparative study are presented in Table II. Our method demonstrates competitive performance across all evaluation metrics on both the IU X-ray and NIH datasets. Specifically, it achieves a BLEU-1 score of 0.486 on the IU X-ray dataset, which is comparable to the method VFE+GKE+SKE+RG (0.496). For BLEU-2 and BLEU-3, our scores are 0.360 and 0.262, respectively, indicating superior performance in generating more complex n-grams compared to all other methods. Our method also achieves the high ROUGE-L value of 0.379 and the highest METEOR score of 0.205, indicating its effectiveness in generating coherent and contextually accurate reports on two datasets.

The comparison reveals that our method surpasses the performance of existing methods across various metrics. Our higher BLEU-3, ROUGE-L, and METEOR scores suggest that our multi-modal feature fusion approach effectively captures both local and global features, leading to more accurate and comprehensive report generation.

On the NIH dataset, our method achieves a BLEU-1 score of 0.412 and a BLEU-2 score of 0.341, which are slightly better than the XRSG method. Additionally, our method achieves the highest ROUGE-L score of 0.413, demonstrating its effectiveness in generating coherent and contextually relevant reports. Our METEOR score of 0.401 matches that of XRSG, further validating the versatility of the RG-MASR approach on different datasets.

*Table II: PERFORMANCE COMPARISON RESULTS.*

| Data set | Method | B-1 | B- 2 | B- 3 | R-L | M. |
|---|---|---|---|---|---|---|
| IU Xray | R2Gen [9] | 0.470 | 0.304 | 0.219 | 0.371 | 0.187 |
| | HRGR [15] | 0.438 | 0.298 | 0.208 | 0.322 | ---- |
| | PPKED [14] | 0.483 | 0.315 | 0.224 | 0.376 | 0.191 |
| | VFE+GK +SKE +RG[29] | **0.496** | 0.327 | 0.238 | 0.381 | ---- |
| | CNX-B2 [27] | 0.479 | **0.363** | 0.261 | 0.354 | 0.188 |
| | Ours | 0.486 | 0.360 | **0.262** | **0.379** | **0.205** |
| HIN | Xr. S.G [28] | 0.396 | 0.329 | ---- | 0.404 | 0.393 |
| | ours | **0.412** | **0.341** | **0.229** | **0.413** | **0.401** |

One key advantage of our method is its ability to leverage the strengths of both CNNs and transformers. Allowing it to identify significant characteristics and correlations in the data effectively. Also, the use of hierarchical CXR feature extraction to extract fine-grained details and global context. Additionally, adopting the self-refining mechanism helps the model to involve a feedback process to evaluate its output and highlight points of improvement.

Moreover, the training manipulation enhances the model's capacity to manage multi-modal data by finely aligning visual and textual features. This refinement enables the model to offer a more comprehensive description found in the image, resulting in improved overall performance compared to some current models.

This indicates that our model can handle such scenarios. When doctors describe the same case using diverse writing vocabularies and styles. However, it is worth noting that while our method performs exceptionally well in most metrics, there is always potential for further enhancement. Extended work could focus on enhancing the training process with more diverse datasets and exploring additional fine-tuning techniques. Additionally, exploring advanced fine-tuning approaches could improve reports' technical detail and specificity. The inclusion of complementary data modalities, such as clinical notes, laboratory test results, imaging metadata, and patient histories, could greatly improve the capabilities of the model beyond existing methods by enhancing its awareness of the medical context. Integrating these multi-modal data sources would enable the model to provide richer, more detailed, and clinically insightful reports. Such improvements can both improve the model performance as well as lay the groundwork for constructing a more generalizable and versatile automated medical report generation tool.

## IV. CONCLUSION AND PERSPECTIVES

Our method is based on multi-modal feature alignment and self-refining strategy (RG-MASR) is a new framework for radiology reports generation that incorporates multi-modality inputs, and a self-refining strategy pays. Using multi-modal feature alignment, the model focuses more on the important information contained in both images and texts while also relying on a self-refining strategy that refines outputs iteratively for higher-quality reports over iterations. The obtained results affirm that RG-MASR outperforms the existing state-of-the-art. Future perspectives include widening the RG-MASR scope to other imaging modalities, including additional clinical data for more detailed reports, and improving the self-refining mechanism to make it more accurate.

## REFERENCES


[1] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When Radiology Report Generation Meets Knowledge Graph", AAAI, vol. 34, no. 07, pp. 12910-12917, 2020.

[2] Baoyu Jing, Zeya Wang, and Eric Xing. ," Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports". In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6570–6580, Florence, Italy. Association for Computational Linguistics, 2019.

[3] Von Luxburg, Ulrike, et al. "Advances in neural information processing systems 30." 31st annual conference on neural information processing systems, Long Beach, California, USA, 2017.

[4] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. ,"Attention is All you Need", Neural Information Processing Systems, 2017.

[5] Wang, S.; Fang, G.; Liu, L.; Wang, J.; Zhu, K.; Melo, S.N. "Transformer-Based Visual Object Tracking with Global Feature Enhancement", Appl. Sci., 13, 12712, 2023.

[6] Tsaniya, Hilya & Fatichah, Chastine & Suciati, Nanik., "Automatic Radiology Report Generator Using Transformer With Contrast-Based Image Enhancement". IEEE Access. PP. 1-1, 2024.

[7] Imane Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz , "Automatic Caption Generation for Medical Images", In Proceedings of the 3rd International Conference on Smart City Applications (SCA '18). Association for Computing Machinery, New York, NY, USA, Article 86, 1–6, 2018.

[8] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 , 2017.

[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 (EMNLP), pages 1439–1449, Online. Association for Computational Linguistics, 2020.

[10] Zhang, Yixiao et al. “When Radiology Report Generation Meets Knowledge Graph.” AAAI Conference on Artificial Intelligence , 2020.

[11] Wang, Xuwen & Zhang, Yu & Guo, Zhen & Li, Jiao. TMRGM: A Template-Based Multi-Attention Model for X-Ray Imaging Report Generation. Journal of Artificial Intelligence for Medical Sciences, 2021.

[12] Navdeep Kaur, Ajay Mittal, “CADxReport: Chest x-ray report generation using co-attention mechanism and reinforcement learning”, Computers in Biology and Medicine, 2022.

[13] Yang S, Wu X, Ge S, Zheng Z, Zhou SK, Xiao L. ,”Radiology report generation with a learned knowledge base and multi-modal alignment”. Med Image Anal. ;86:102798. Epub 2023 , Mar 2023.

[14] F. Liu, X. Wu, S. Ge, W. Fan and Y. Zou, "Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 13748-13757 , 2021.

[15] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. , “Hybrid retrieval-generation reinforced agent for medical image report generation”, In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 1537–1547, 2018.

[16] Aksoy, N., Sharoff, S., Baser, S., Ravikumar, N., & Frangi, A. F. ,” Beyond images: An integrative multi-modal approach to chest x-ray report generation. Frontiers in Radiology”, 4, 1339612 , 2024.

[17] Yu Pan, Li-Jun Liu, Xiao-Bing Yang, Wei Peng, Qing-Song Huang, “Chest radiology report generation based on cross-modal multi-scale feature fusion”, Journal of Radiation Research and Applied Sciences, Volume 17, Issue 1, 2024.

[18] F. Cheddi, A. Habbani and H. Nait-Charif, "A Multi-Modal Feature Fusion-Based Approach for Chest X-Ray Report Generation," 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM), Leeds, United Kingdom, 2024.

[19] He, K., Zhang, X., Ren, S., & Sun, J. “Deep Residual Learning for Image Recognition”, 2015 .

[20] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. , “BioBERT: A pre-trained biomedical language representation model for biomedical text mining”. ArXiv, 2019.

[21] Madaan et al. ,”Self-Refine: Iterative Refinement with Self-Feedback. ArXiv, 2023.

[22] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, “Preparing a collection of radiology examinations for distribution and retrieval”, J. Am. Med. Inform. Assoc. 23 (2) 304–310, 2016.

[23] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2097–2106, 2017.

[24] 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics, 2002.

[25] Chin-Yew Lin.. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics , 2004.

[26] Michael Denkowski and Alon Lavie, “ Meteor Universal: Language Specific Translation Evaluation for Any Target Language”, In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics , 2014.

[27] F. F. Alqahtani, M. M. Mohsan, K. Alshamrani, J. Zeb, S. Alhamami and D. Alqarni, "CNX-B2: A Novel CNN-Transformer Approach For Chest X-Ray Medical Report Generation," in IEEE Access, vol. 12, pp. 26626-26635, 2024.

[28] Veras Magalhães G, L de S Santos R, H S Vogado L, Cardoso de Paiva A, de Alcântara Dos Santos Neto P. XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model. Heliyon. Mar 12;10(7) , 2024.

[29] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, ‘‘Knowledge matters: Chest radiology report generation with general and specific knowledge,’’ Med. Image Anal., vol. 80, Aug. 2022.