

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

A FRAMEWORK FOR EXPLAINABLE AND HIGH-
QUALITY RADIOLOGY REPORT GENERATION
USING AUTOMATIC KEYWORD ADAPTATION,
FREQUENCY-BASED MULTI-LABEL
CLASSIFICATION AND TEXT-TO-TEXT LARGE
LANGUAGE MODELS

HE, ZEBANG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Health Technology and Informatics

**A Framework for Explainable and High-Quality
Radiology Report Generation Using Automatic Keyword
Adaptation, Frequency-Based Multi-Label Classification
and Text-to-text Large Language Models**

HE, Zebang

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

April 2025

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

HE, Zebang (Name of Student)

Abstract

Background: Radiology reports are essential in medical imaging, providing critical insights for diagnosis, treatment, and patient management by bridging the gap between radiologists and clinicians. However, the manual generation of these reports is time-consuming and labor-intensive, leading to inefficiencies and delays in clinical workflows, particularly as case volumes increase. Although deep learning approaches have shown promise in automating radiology report generation, existing methods, particularly those based on the encoder-decoder framework, suffer from significant limitations. These include a lack of explainability due to black-box features generated by encoder and limited adaptability to diverse clinical settings.

Purpose: This study aims to develop a deep learning-based radiology report generation framework that could generate high-quality and explainable radiology report for chest X-Ray images.

Methods and Materials: In this study, we address these challenges by proposing a novel deep learning framework for radiology report generation that enhances explainability, accuracy, and adaptability. Our approach replaces traditional black-box features in computer vision with transparent keyword lists, improving the interpretability of the feature extraction process. To generate these keyword lists, we apply a multi-label classification technique, which is further enhanced by an automatic keyword adaptation mechanism. This adaptation dynamically configures the multi-label classification to better adapt specific clinical environments, reducing the reliance on manually curated reference keyword lists and improving model adaptability across diverse datasets. We also introduce a frequency-based multi-label classification strategy to address the issue of keyword imbalance, ensuring that

rare but clinically significant terms are accurately identified. Finally, we leverage a pre-trained text-to-text large language model (LLM) to generate human-like, clinically relevant radiology reports from the extracted keyword lists, ensuring linguistic quality and clinical coherence.

Results: We evaluate our method using two public datasets, IU-XRay and MIMIC-CXR, demonstrating superior performance over state-of-the-art methods. Our framework not only improves the accuracy and reliability of radiology report generation but also enhances the explainability of the process, fostering greater trust and adoption of AI-driven solutions in clinical practice. Comprehensive ablation studies confirm the robustness and effectiveness of each component, highlighting the significant contributions of our framework to advancing automated radiology reporting.

Conclusion: In this study, we developed a novel Deep-Learning based Radiology Report Generation framework for generating high-quality and explainable radiology report for chest X-Ray images using the multi-label classification and text-to-text large language model. Through replacing the black-box semantic features into visible keyword lists, our framework could solve the unexplainability of the current workflow and provide the clear and flexible automatic pipeline for reducing the workload of radiologists and the further applications related to Human-AI interactive communication.

Publications

Journal Articles

1. **He, Zebang**, Alex Ngai Nick Wong and Jung Sun Yoo. “Co-ERA-Net: Co-Supervision and Enhanced Region Attention for Accurate Segmentation in COVID-19 Chest Infection Images.” *Bioengineering* 10 (2023): n. pag.
2. Wong, Alex Ngai Nick, **Zebang He**, Ka Long Leung, Curtis Chun Kit To, Chun Yin Wong, Sze Chuen Cesar Wong, Jung Sun Yoo, Cheong Kin Ronald Chan, Angela Zaneta Chan, Maribel D Lacambra and Martin Ho Yin Yeung. “Current Developments of Artificial Intelligence in Digital Pathology and Its Future Clinical Applications in Gastrointestinal Cancers.” *Cancers* 14 (2022): n. pag.
3. **He, Zebang**, Alex Ngai Nick Wong and Jung Sun Yoo. Radiology report generation using automatic keyword adaptation, frequency-based multi-label classification and text-to-text large language models, *Computers in Biology and Medicine*, Volume 196, Part A, 2025, 110625, ISSN 0010-4825.

Conference Abstract

1. **Mr. He Zebang**, Dr. Ho-yin Anson Cheung, Dr. Ting Hei, Edward Wong and Dr. Jung Sun Yoo. “Enhancing Chest X-Ray Report Generation in Radiology through Deep Learning: Leveraging Keywords from Existing Reports and Multi-Label Classification.” *Journal of Medical Imaging and Radiation Sciences* (2024): n. pag. 10 (2023): n. pag.

2. Dr. Yoo Jung Sun, Dr Chung Ting Tang and **Mr. Zebang He**. “Fast and low dose myocardial perfusion imaging SPECT using deep learning-based denoising.” Journal of Medical Imaging and Radiation Sciences (2024): n. pag.
3. **He, Zebang**, Alex Ngai Nick Wong and Jung Sun Yoo. “Integrating Dual-Supervision and Region Fusion Attention for Infection Segmentation in Chest Radiology Image” WMIC 2024, Montreal September 9th-13th, Poster Presentation.
4. **He, Zebang**, Alex Ngai Nick Wong and Jung Sun Yoo. “Integrating Dual-Supervision and Region Fusion Attention for Infection Segmentation in Chest Radiology Image” American College of Radiology Annual Meeting 2024, April 13-17, 2024, at Washington Hilton in Washington, DC, ePoster Presentation.

Invited Talk

1. **Mr. He Zebang** and Dr. Jung Sun Yoo. “Deep learning based differentiation of COVID-19 and non-COVID-19 infection cases and segmentation of COVID-19 infected regions in Chest X-ray images” The Hong Kong Radiographers' Association Annual Meeting, August 12, 2023, Hong Kong.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my chief supervisor, Dr. Jung Sun Yoo, and my co-supervisor, Prof. Jing Cai, for their invaluable guidance, expertise, and unwavering support throughout my research journey. Their profound knowledge in the medical field, which was initially outside my comfort zone, has been instrumental in shaping this thesis. Their motivation, efficiency, and keen insight into emerging research trends have been a constant source of inspiration for me.

I am also deeply indebted to Dr. Alex Ngai Nick Wong, whose academic mentorship and unwavering support have been pivotal to the success of this project. His constructive feedback, thoughtful revisions, and encouragement have significantly enriched my work.

My sincere thanks go to all the academic staff in the Department of Health Technology and Informatics for their professional advice and assistance throughout my academic journey. Their expertise and dedication have been invaluable in helping me navigate the challenges of this research.

I am equally grateful to my labmates for their generous support and collaboration. Special thanks to Mr. HongZhao Chen, Mr. HeXiao Ding, Mr. Genald Cheng, and Mr. Sam Ng for their insightful suggestions and camaraderie during this research.

Finally, I would like to extend my heartfelt appreciation to my parents, friends, and all those who have supported me academically, emotionally, and personally. Their encouragement, advice, and unwavering belief in me have been my greatest strength.

This research marks a significant milestone in my life, and I am committed to leveraging this experience to contribute further to both academic and industrial advancements in the medical field through continuous learning and dedication.

Contents

1.	Literature Review.....	1
1.1	Chest X-Ray and Radiology Report	1
1.1.1	Chest X-Ray.....	1
1.1.2	Radiology Report	4
1.1.3	The Conflict Between Low Efficiency of Manual Radiology Report Writing and the Increasing Demand for Timely Radiological Diagnosis	8
1.2	Deep-Learning based Radiology Report Generation.....	11
1.2.1	From Traditional AI to Deep Learning in radiology: The Expanding Scale of AI Models.....	11
1.2.2	Vision-Language Deep Learning Application	13
1.2.3	Current Radiology Report Generation Work: Dataset and Algorithms.....	17
1.3	Automatic Keyword Adaption.....	23
1.3.1	Radiology Dictionary	23
1.3.2	Keyword Extraction from Text.....	25
1.4	Frequency-based Multi-Label Classification.....	27
1.4.1	Development of Classification in Deep Learning.....	27
1.4.2	Multi-Label Classification	30
1.5	Text-to-text Large Language Model.....	33
1.5.1	Brief Introduction of Language Model	33
1.5.2	Large Language Model	36
2.	Research Aims and Objectives	39
2.1	Research Gaps	39
2.2	Research Aims	40
2.3	Research Objectives	41
3.	Materials and Methods.....	42

3.1	Dataset Collection.....	42
3.1.1	IU X-Ray Dataset.....	42
3.1.2	MIMIC-CXR Dataset.....	44
3.2	Radiology Report Generation Framework based on Automatic Keyword Adaption, Frequency-based Multi-Label Classification and Text-to-text Large Language Model..	45
3.2.1	Overview of whole framework	45
3.2.2	Automatic Keyword Adaption.....	47
3.2.3	Frequency-based Multi-Label Classification	50
3.2.4	Radiology Report Generation from keywords based on Large Language Model	53
3.2.5	Loss Function.....	55
4.	Experiments and Results	58
4.1	Implementation Details.....	58
4.2	Evaluation Details and Metrics.....	61
4.3	Samples of Extracting Keywords from Radiology Reports and Failure Cases	62
4.4	Keyword Distribution Analysis	64
4.5	Quantitative Analysis of Radiology Report Generation.....	68
4.6	Qualitative Analysis of Radiology Report Generation.....	73
4.7	Ablation Study.....	76
4.7.1	Performance of Multi-Label Classification Across Network Architectures.....	76
4.7.2	Influence of Keyword Numbers and the Combination of High- and Low- Frequency Keywords on Radiology Report Generation Performance.....	80
4.7.3	Performance of Radiology Report Generation with Different Text-to-text Large Language Models and Pretrained Materials.....	85
5.	Discussion	88
5.1	Relationship between keyword list length and radiology report generation performance	88
5.2	Commercial Large Language Model in Radiology Report Generation Method ...	93

5.3	Performance effect of the keyword prediction quality across datasets	101
5.4	Automatic Generation of Impression from Findings by Text-to-text Large Language Models.....	105
6.	Conclusion	108
7.	Reference	112

List of Tables

Table 1. Comparison of state-of-the-art radiology report generation methods on the IU X-Ray dataset. The table highlights the image encoder, text decoder, and performance metrics for each method. Notably, for methods incorporating a language model during text decoding, the language model is listed under the "Text Decoder" column. The reference list of the state-of-the-art radiology report generation methods is shown in the next table.	20
Table 2. Comparison of state-of-the-art radiology report generation methods on the MIMIC-CXR dataset. The table presents the image encoder, text decoder, and associated performance metrics for each method. Methods that incorporate a language model for text decoding are explicitly listed with the language model under the "Text Decoder" column. The reference list of the state-of-the-art radiology report generation methods is shown in the next table.	22
Table 3. Overview of classic and state-of-the-art classification networks. The networks are systematically collected and reorganized from the source of MMPretrained Contributors (2023) [156]. The table includes key architectures that have significantly influenced the field of image classification, ranging from foundational models to advanced deep learning frameworks. Each network is categorized based on its architectural characteristics, such as convolutional layers, residual connections, or attention mechanisms, and their contributions to improving classification accuracy and computational efficiency are highlighted. The table serves as a comprehensive reference for understanding the evolution and performance of classification networks in both general and domain-specific applications.	30
Table 4. Dataset descriptions for the IU X-Ray and MIMIC-CXR datasets, including details of keyword cluster information derived from the automatic keyword adaptation process. Additionally, the table provides information on the radiology dictionary used in the	

experiments and presents the proportion of keywords in the two experimental datasets relative to the full version of the dictionary.....60

Table 5. Coverage ratio of keywords generated through Automatic Keyword Adaptation in the test and validation sets of the IU X-Ray and MIMIC-CXR datasets, evaluated using two strategies. In the Keyword-Based Strategy, radiology reports are split into unique keywords by removing duplicates, and the coverage ratio is calculated as the percentage of matched keywords from the generated keyword set relative to the total unique keywords in the report. In the Text-Based Strategy, the generated keyword set is searched directly within the report text, with matching words or phrases highlighted. The coverage ratio is then computed as the proportion of the total character length of matched words or phrases to the total character length of the report.....67

Table 6. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the IU X-Ray dataset. For most methods, the results are cited directly from their respective publications, presented under "Paper Report Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean±standard deviation, and results from re-training and the proposed pipeline are presented with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.....70

Table 7. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the MIMIC-CXR dataset. For most methods, the results are cited directly from their respective publications, presented under "Paper Report

Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean \pm standard deviation, and results from re-training and the proposed pipeline are presented with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.....71

Table 8. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the MIMIC-CXR dataset for the Clinical Evaluation (CE) Metrics in CheXpert Label Accuracy. For most methods, the results are cited directly from their respective publications, presented under "Paper Report Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean \pm standard deviation, and results from re-training and the proposed pipeline are presented with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.....72

Table 9. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the MIMIC-CXR dataset for the Clinical Evaluation (CE) Metrics in RadGraph F1. For most methods, the results are cited directly from their respective publications, presented under "Paper Report Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean \pm standard deviation, and results from re-training and the proposed pipeline are presented

with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.....73

Table 10. Performance comparison of different networks for multi-label classification on the IU X-Ray dataset. Results are reported as mean±standard deviation with precision up to four decimal places. Additionally, the average performance across all frequency clusters is calculated and presented as a mean value for reference. The bolded "ConvNeXt" represents the network configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods78

Table 11. Performance comparison of different networks for multi-label classification on the MIMIC-CXR dataset. Results are reported as mean±standard deviation with precision up to four decimal places. Additionally, the average performance across all frequency clusters is calculated and presented as a mean value for reference. The bolded "ConvNeXt" represents the network configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.79

Table 12. Performance comparison of different networks for radiology report generation on the IU X-Ray and MIMIC-CXR datasets. Results are presented as mean±standard deviation with precision up to four decimal places. The bolded "ConvNeXt" denotes the network configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.80

Table 13. Performance comparison of radiology report generation on the IU X-Ray dataset across different keyword clusters. Results are reported as mean ± standard deviation with precision up to four decimal places. The bolded "Cluster All Keywords" represents the

configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.	82
Table 14. Performance comparison of radiology report generation on the MIMIC-CXR dataset across different keyword clusters. Results are presented as mean \pm standard deviation with precision up to four decimal places. The bolded "Cluster All Keywords" represents the configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.	83
Table 15. Performance comparison of radiology report generation on the IU X-Ray dataset across different text-to-text large language model versions and their pretraining materials. Results are reported as mean \pm standard deviation with precision up to four decimal places. Additionally, the trainable model parameters are provided to indicate the complexity of each language model. The suffixes "-base" and "-small" denote the model size of the respective versions.	87
Table 16. Performance comparison of radiology report generation on the MIMIC-CXR dataset across different text-to-text large language model versions and their pretraining materials. Results are reported as mean \pm standard deviation with precision up to four decimal places. Additionally, the trainable model parameters are provided to indicate the complexity of each language model. The suffixes "-base" and "-small" denote the model size of the respective versions.	87
Table 17. Average keyword list length in the IU X-Ray and MIMIC-CXR datasets. The keyword list length refers to the number of keywords associated with each case. The row "Before Filtering" indicates the initial keyword lists generated through the keyword extraction from the automatic keyword adaptation process. The row "After Filtering (Low Frequency)" represents the keyword lists after removing terms that appear fewer than 10 times, aiming to	

address extreme class imbalance and reduce the complexity of multi-label classification. The row "After Filtering (Radiology Dictionary)" further refines the keyword lists by validating them against the RadLex dictionary to ensure clinical appropriateness; these filtered lists are used as the final keyword sets for frequency-based multi-label classification. The row "Input to TT-LLM" shows the average number of keywords provided as input to the TT-LLM, based on predictions from the classification network. Since the validation and test sets are not involved in training, the TT-LLM input lengths are not reported for these splits.90

Table 18. Performance comparison of radiology report generation across the IU X-Ray and MIMIC-CXR datasets using a cross-dataset evaluation strategy (training on one dataset and testing on the other). The table presents the results of our proposed framework, reported as mean \pm standard deviation with precision up to four decimal places. Key evaluation metrics, such as BLEU, ROUGE, and clinical accuracy, are included to assess the quality and robustness of the generated reports..... 103

Table 19. Performance comparison of generating the Impression section from Findings in the IU X-Ray and MIMIC-CXR datasets. The table evaluates the ability of our proposed framework to generate the Impression section based on Findings, with results reported as mean \pm standard deviation and precision up to four decimal places. The “Findings generated by our framework” setting refers to Findings produced under the experimental setup described in Section 4.4 (“Experiments and Results”). This comparison highlights the framework’s effectiveness in synthesizing concise and clinically relevant Impression sections from both ground truth and framework-generated Findings, demonstrating its utility in automating radiology report generation..... 106

List of Figures

Figure 1. Sample chest X-ray imaging device used in Hong Kong. The figure illustrates a typical chest X-ray setup, including the imaging equipment and positioning framework commonly used in clinical practice.....	3
Figure 2. Sample chest X-ray images in posteroanterior (PA) and lateral views. The figure presents two standard radiographic projections used in chest imaging. The PA view provides a frontal perspective, while the lateral view offers a complementary side profile, aiding in comprehensive assessment of thoracic structures.....	4
Figure 3. The Different imaging modalities of chest radiographs and their corresponding radiology report types. From left to right, the figure presents examples of Chest X-ray, MRI, and CT scans, each associated with distinct reporting styles tailored to their imaging characteristics and diagnostic requirements.....	7
Figure 4. The general time of manual radiology report writing in X-Ray. From [117]. It shows that for chest X-Ray, the average time for manual writing of radiology report for each case is 1 min 30s to 2 min 30s.....	9
Figure 5. The disparity between patient demand for radiology diagnosis and the availability of expert radiologists. The figure highlights the significant gap in access to radiological expertise, based on data from the 2023 Hospital Authority report and the Hong Kong College of Radiologists. This shortage contributes to increased workload, longer diagnosis times, and potential delays in patient care.....	10
Figure 6. Overview of representative tasks in vision-language (VL) problems and the CLIP model pipeline. (a) Illustration of VL tasks, including image-text transformation (orange), video-text transformation (green), and the use of text as auxiliary information in vision tasks (pink) [143]. (b) The CLIP model pipeline, representing a general framework for leveraging	

language text as auxiliary information to enhance pretraining in vision-related tasks [92]. Additionally, various CLIP-based models, such as MedCLIP[130] and PubmedCLIP[140], have been introduced in medical vision-language applications, differing primarily in their training language datasets.16

Figure 7. General pipeline for radiology report generation. The process begins with an input chest X-ray image, which is processed by an image encoder to extract high-level semantic features that are not directly interpretable by humans. A text decoder then utilizes these features to generate a corresponding radiology report in a structured paragraph format.19

Figure 8. The RadLex Dictionary from RSNA. Various radiology dictionaries have been developed to standardize terminology and improve consistency in radiology reporting. Among them, the RadLex Dictionary from the Radiological Society of North America (RSNA) is widely used, offering both online and offline versions for flexible access. Other dictionaries have been introduced for educational and reference purposes, though they often lack direct extraction from radiology reports, making RadLex a preferred resource for structured reporting and automated processing.25

Figure 9. General pipeline for keyword extraction using a language model, illustrated with KeyBERT [154]. The keyword extraction process begins with an input texts, from which candidate keywords are identified using contextual embeddings generated by a pre-trained language model. KeyBERT, as an example, leverages BERT-based embeddings to rank and extract the most relevant keywords based on their semantic similarity to the input text.27

Figure 10. Illustration of the difference between multi-label classification and multi-class classification (in single-label classification). Adapted from HuggingFace’s beginner tutorial [170]. This figure highlights the key distinctions between the two classification paradigms. In multi-class classification, each input is assigned a single label from a set of mutually exclusive

categories, whereas in multi-label classification, an input can be associated with multiple labels simultaneously. The figure visually demonstrates these concepts using example scenarios, emphasizing how multi-label classification accommodates overlapping or concurrent labels, while multi-class classification enforces exclusivity. This distinction is critical for understanding the appropriate application of each approach.....32

Figure 11. Example of multi-label classification applied to medical imaging. This sample, captured from [167] demonstrates the process of performing multi-label classification on Chest X-Ray images. The figure illustrates how multiple diagnostic labels, such as pneumothorax, effusion, or cardiomegaly, can be assigned to a single image, reflecting the presence of concurrent or overlapping conditions. This example highlights the utility of multi-label classification in medical imaging, where complex and co-occurring pathologies often require simultaneous identification for accurate diagnosis and treatment planning.33

Figure 12. A brief timeline illustrating the development of language models. The figure highlights key milestones in the evolution of language models, starting from traditional statistical approaches to the emergence of deep learning-based models. It is noted that after the development of RNNs (Recurrent Neural Networks) and LSTMs (Long Short-Term Memory networks), language models transitioned to deep learning-based architectures. Following the introduction of Transformers, language models began to scale significantly in size and complexity, leading to the era of Large Language Models (LLMs). This timeline, adapted from [172], showcases the progression of language models and their increasing capabilities in natural language understanding and generation tasks.....35

Figure 13. A brief timeline depicting the development of large language models (LLMs). The figure highlights the transformative impact of the Transformer architecture, introduced in 2017, which marked the emergence of the LLM. Following the development of foundational models

such as GPT-1 and BERT, the technology underlying LLMs stabilized, enabling widespread adoption and further advancements. Subsequent progress has focused on refining training materials, optimizing training methods, and enhancing user interfaces (GUIs) to improve accessibility and usability. This timeline, adapted from [173], illustrates the evolution of LLMs and their growing influence on natural language processing and AI-driven applications.....38

Figure 14. Example of a radiology report from the IU X-Ray dataset. The figure showcases a typical radiology report, which includes a detailed textual description of findings derived from medical imaging. The report is structured to provide clinical observations, interpretations, and recommendations, highlighting key aspects such as anatomical structures, abnormalities, and potential diagnoses. Please note that, although the original version of IU X-Ray dataset provide the annotations in Manual/MTI, these annotations are for the MeSH[174] codes supplemented by Radiology Lexicon (RadLex) codes, which is not actually matched the keywords appeared in Finding of radiology report. Therefore, we do not utilize it for radiology report generation.43

Figure 15. Example of a radiology report from the MIMIC-CXR dataset. The figure presents a representative radiology report, which includes a comprehensive textual description of imaging findings, clinical observations, and diagnostic interpretations.44

Figure 16. Overview of the proposed radiology report generation pipeline integrating automatic keyword adaptation and frequency-based multi-label classification. The process begins with automatic keyword adaptation, which processes radiology reports from the training set to extract keyword clusters. These clusters are then used to configure the frequency-based multi-label classification. Subsequently, the frequency-based multi-label classification predicts keyword lists for each cluster, which are combined through keyword list fusion to generate the

final keyword list. Finally, a large text to text language model generates the corresponding radiology report using the fused keyword list.....47

Figure 17. Block diagram of the proposed automatic keyword adaptation process. Beginning with the provided radiology reports, candidate keywords are extracted using the KeyBERT tool. Subsequently, the extracted keywords are filtered using the RadLex Dictionary, retaining only those present in the dictionary. The filtered keywords are ranked by frequency, and keyword clusters are constructed based on this ranking for use in frequency-based multi-label classification. If users provide a keyword candidate list, these user-specified keywords are given the highest priority and processed through a separate branch. After validation to ensure the keywords exist within the reports, the user-provided keywords are treated as an individual cluster and integrated into the keyword clusters for multi-label classification.49

Figure 18. Comparison of our proposed multi-label classification approach with general single-label classification using classic classification networks. The figure illustrates the key differences between the two methodologies, highlighting how our proposed method accommodates multiple labels per input, enabling the identification of concurrent or overlapping conditions. In contrast, single-label classification assigns only one label per input, limiting its applicability in scenarios where multiple attributes or diagnoses need to be captured simultaneously.52

Figure 19. Diagram illustrating keyword list fusion in the frequency-based multi-label classification process. To account for the information density in high-frequency clusters, the threshold for these clusters is increased, reducing the likelihood of incorrect predictions for high-frequency keywords. Conversely, for low-frequency clusters that may contain rare but clinically important keywords, the threshold is slightly decreased to allow for the inclusion of more keywords. For individual clusters, a standard classification threshold of 0.5 is applied.

The keywords from all clusters, after threshold-based filtering, are combined to form the final keyword list, which is then used for radiology report generation.....52

Figure 20. Full workflow of Automatic Keyword Adaption. It includes keyword filtering by radiology dictionary and frequency, and then clustering for multi-label classification53

Figure 20. Diagram illustrating the workflow of radiology report generation based on multi-label classification within our proposed framework. The figure outlines the step-by-step process, beginning with the multi-label classification which contains relevant diagnostic labels. These labels are then used to guide the generation of structured and coherent radiology reports after adding prefixes to construct the command to the large language model. The diagram highlights the integration of multi-label classification with natural language generation techniques, demonstrating how our framework ensures accurate and contextually relevant report generation. This approach leverages the strengths of multi-label classification to capture complex and concurrent findings, enhancing the overall quality and utility of automated radiology reporting.....55

Figure 22. Sample cases with corresponding radiology reports and estimated keyword counts from the IU X-Ray and MIMIC CXR datasets. Representative cases were randomly selected from each dataset, and their associated radiology reports were obtained from the official sources. For each cases, Keywords were manually annotated and verified in the RadLex radiology dictionary in Line 1 and we also show the extraction result by our framework in Line 2. The total keyword list length reflects the number of identified keywords in each case.63

Figure 22. Sample failure cases from the IU X-Ray and MIMIC-CXR datasets. Representative examples were randomly selected from both datasets, with radiology reports sourced from official repositories. Keywords were initially extracted using the Automatic Keyword

Adaptation process, and subsequently cross-checked with the RadLex radiology dictionary to identify potentially missing or unrecognized clinical terms.	64
Figure 23. Methodology for calculating the keyword-based coverage ratio and text-based coverage ratio of the generated keywords in comparison to the ground truth radiology reports.	66
Figure 24. Pareto chart illustrating the keyword distribution in the IU X-Ray dataset. Blue bars represent the frequency of each keyword in the training set, while the orange line indicates the cumulative frequency ratio from the most frequent keyword to the current keyword compared with the total frequency of keyword.	67
Figure 25. Pareto chart illustrating the keyword distribution in the MIMIC-CXR dataset. Blue bars represent the frequency of each keyword in the training set, while the orange line indicates the cumulative frequency ratio from the most frequent keyword to the current keyword compared with the total frequency of keyword.	68
Figure 26. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the IU X-Ray dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue. The keywords in red are not shown in the generated radiology report by our framework because it is not forced to contain in the generated reports in our commands to the large language model.	75
Figure 27. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the MIMIC-CXR dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue.	75

Figure 28. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the IU X-Ray dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue.	83
Figure 29. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the MIMIC-CXR dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue.	84
Figure 30. Keyword list length distribution in the IU X-Ray and MIMIC-CXR datasets. This figure presents the relationship between keyword list length and the frequency of corresponding lengths observed in the two datasets. The results indicate that most keyword lists fall within medium-length ranges, while extremely short and long keyword lists occur less frequently. This distribution generally aligns with a normal-like pattern, reflecting typical variability in radiology report complexity	90
Figure 31. Performance and keyword list length distribution in the IU X-Ray dataset. This figure illustrates the relationship between keyword list length and the performance of the proposed method on the IU X-Ray dataset. The results show a general decline in performance across all evaluation metrics as the keyword list length increases. This trend suggests that longer keyword lists correspond to more complex radiology reports, which pose greater challenges for accurate generation. The analysis highlights the difficulty of handling lengthy and detailed inputs, emphasizing an area for future improvement in radiology report generation models	91
Figure 32. Performance and keyword list length distribution in the MIMIC-CXR dataset. This figure illustrates the relationship between keyword list length and the performance of the	

proposed method on the MIMIC-CXR dataset. The results show a general decline in performance across all evaluation metrics as the keyword list length increases. This trend suggests that longer keyword lists correspond to more complex radiology reports, which pose greater challenges for accurate generation. The analysis highlights the difficulty of handling lengthy and detailed inputs, emphasizing an area for future improvement in radiology report generation models.92

Figure 33. Graphical User Interface (GUI) of the PolyU GenAI platform. The figure showcases the interface designed to interact with and evaluate multiple commercial large language models (LLMs), including GPT-4, Qwen, and Llama. The platform provides a user-friendly environment for testing and comparing the capabilities of these LLMs in various tasks, such as text generation, summarization, and question answering. We utilize the GPT-4o, Qwen and Llama as the sample commercial LLMs.96

Figure 34. Process of obtaining generated radiology reports using GPT-4 and DeepSeek on their respective online platforms. The figure illustrates the workflow for generating radiology reports, starting from inputting medical data to receiving the final output. While both platforms offer free access, it is noted that both DeepSeek and GPT-4 imposes traffic control restrictions, preventing continuous generation of reports on their online platform.97

Figure 35. Deployment of the DeepSeek model with distillation in a lightweight structure on our workstation using the vLLM framework. The figure demonstrates the process of optimizing and deploying the DeepSeek model for efficient inference, leveraging model distillation techniques to reduce computational requirements while maintaining performance.98

Figure 36. Visualization of results generated by the proposed framework compared to commercial large language models (LLMs) on the IU X-Ray dataset. The figure provides a side-by-side comparison of radiology reports generated by our framework and those produced

by commercial LLMs, such as GPT-4 and DeepSeek. The visualization highlights the accuracy, coherence, and clinical relevance of the generated reports, emphasizing the strengths of our framework in capturing detailed medical findings and producing structured outputs. This comparison underscores the potential of our approach to enhance radiology report generation, offering a viable alternative to existing commercial solutions.99

Figure 37. Visualization of results generated by the proposed framework compared to commercial large language models (LLMs) on the MIMIC-CXR dataset. The figure presents a comparative analysis of radiology reports produced by our framework and those generated by commercial LLMs, such as GPT-4 and DeepSeek. The side-by-side visualization highlights the accuracy, detail, and clinical relevance of the reports, showcasing the ability of our framework to effectively interpret complex medical imaging data and produce structured, actionable outputs. This comparison demonstrates the potential of our approach to improve radiology report generation, offering a robust alternative to existing commercial solutions in handling diverse and challenging datasets like MIMIC-CXR..... 101

Figure 38. Process of evaluating the robustness of radiology report generation using our proposed framework through cross-dataset experimentation. The figure illustrates the methodology of training the framework on one dataset (e.g., IU X-Ray) and testing it on another dataset (e.g., MIMIC-CXR) to assess its generalizability and robustness. This experiment demonstrates the framework's potential for real-world deployment, where variability in data sources and clinical contexts is common. 103

Figure 39. Visualization of results generated by the proposed framework across the two datasets, using IU X-Ray as the test set. The figure compares two scenarios: (1) training and testing on the IU X-Ray dataset (column 3) and (2) training on the MIMIC-CXR dataset and testing on the IU X-Ray dataset (column 4). The visualization highlights the differences in report quality,

coherence, and clinical relevance between the two scenarios, demonstrating the framework's adaptability and robustness when trained on different datasets. This comparison underscores the importance of dataset diversity in training and the framework's ability to generalize across varying clinical contexts. 104

Figure 40. Visualization of results generated by the proposed framework across the two datasets, using MIMIC-CXR as the test set. The figure compares two scenarios: (1) training and testing on the MIMIC-CXR dataset (column 3) and (2) training on the IU X-Ray dataset and testing on the MIMIC-CXR dataset (column 4). The visualization highlights the differences in report quality, coherence, and clinical relevance between the two scenarios, demonstrating the framework's adaptability and robustness when trained on different datasets. This comparison underscores the importance of dataset diversity in training and the framework's ability to generalize across varying clinical contexts..... 104

Figure 41. Visualization of generated Impression sections from Findings by our proposed framework on the IU X-Ray dataset. The figure showcases examples of Impression sections generated by the framework based on both ground truth and framework-generated Findings. The visualization highlights the coherence, clinical relevance, and accuracy of the generated Impression sections, demonstrating the framework's ability to synthesize concise summaries from detailed Findings. This example underscores the effectiveness of the framework in automating the generation of clinically useful radiology reports. 106

Figure 42. Visualization of generated Impression sections from Findings by our proposed framework on the MIMIC-CXR dataset. The figure presents examples of Impression sections generated by the framework based on both ground truth and framework-generated Findings. The visualization emphasizes the coherence, clinical relevance, and accuracy of the generated Impression sections, illustrating the framework's capability to produce concise and actionable

summaries from detailed Findings. This example further validates the framework's robustness and applicability across diverse datasets in radiology report generation..... 107

Figure 43. Proposed future collaboration workflows between the pipeline and radiologists for refining keywords and finalizing radiology reports. The diagram illustrates three approaches: the proposed pipeline (a) and two collaborative workflows (b that contain b-1 and b-2). In the proposed pipeline (a), medical imaging data is processed for keyword extraction, undergoing automatic verification and refinement before being input into a pretrained large language model (LLM) to generate clinically relevant reports. In the (b-1) workflow, the process is enhanced by radiologists double-checking the refined keywords before they are input into the LLM for report generation. In the (b-2) workflow, the refined keywords undergo the same process as (b-1), but the generated reports are further reviewed and revised by radiologists to produce the final radiology report, ensuring the highest quality and clinical accuracy. 111

List of Abbreviations

1. CXR -> Chest X-ray
2. CT -> computed tomography
3. MRI -> Magnetic Resonance Imaging
4. PA -> posteroanterior view
5. COPD -> chronic obstructive pulmonary disease
6. COVID-19 -> coronavirus disease 2019
7. WHO -> World Health Organization
8. UNSCEAR -> United Nations Scientific Committee on the Effects of Atomic Radiation
9. AI -> artificial intelligence
10. VGG -> Visual Geometry Group
11. Seq2Seq -> sequence-to-sequence
12. GPU -> graphics processing units
13. BERT -> Bidirectional Encoder Representations from Transformers
14. NLP -> natural language processing
15. CLIP -> Contrastive Language-Image Pre-Training
16. RNN -> Recurrent Neural Networks
17. LSTM -> Long Short-Term Memory
18. LLM -> Large Language Models
19. GPT -> Generative Pre-trained Transformer
20. VL -> vision-language
21. EEG -> electroencephalogram
22. SE -> Squeeze-and-Excitation
23. CBAM -> Convolutional Block Attention Module

- 24. Word2Vec -> Word Representations in Vector Space
- 25. BART -> Bidirectional and Auto-Regressive Transformers
- 26. GUI -> Graphical User Interface
- 27. LLaMA -> Large Language Model Meta AI
- 28. T5 -> Text to Text Transfer Transformer
- 29. CNN -> Convolutional Neural Network
- 30. GCN -> Graph Convolution Networks
- 31. RSNA -> Radiological Society of North America
- 32. SVM -> Support Vector Machines
- 33. ViT -> Vision Transformer
- 34. SOTA -> state-of-the-art
- 35. BLEU -> Bilingual Evaluation Understudy
- 36. ROUGE-L -> Recall-Oriented Understudy for Gisting Evaluation
- 37. CIDEr -> Consensus-based Image Description Evaluation
- 38. METEOR -> Metric for Evaluation of Translation with Explicit Ordering

List of State-of-the-art Radiology Report Generation Model in Literature Review 1.2.4 and Experiments and Results 4.4

1. On the Automatic Generation of Medical Imaging Reports [1]
2. Multimodal Recurrent Model with Attention for Automated Radiology Report Generation [2]
3. Addressing Data Bias Problems for Chest X-ray Image Report Generation[3]
4. Attention-Based Abnormal-Aware Fusion Network for Radiology Report Generation [4]
5. Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment [5]
6. Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation [6]
7. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports [7]
8. Generating Radiology Reports via Memory-driven Transformer [8]
9. When Radiology Report Generation Meets Knowledge Graph [9]
10. A Self-boosting Framework for Automated Radiographic Report Generation [10]
11. Automated radiology report generation using conditioned transformers [11]
12. Contrastive Attention for Automatic Chest X-ray Report Generation [12]
13. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation [13]
14. Knowledge matters: Chest radiology report generation with general and specific knowledge[14]
15. Progressive Transformer-Based Generation of Radiology Reports [15]
16. Radiology Report Generation with a Learned Knowledge Base and Multi-modal Alignment [16]

17. RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting [17]
18. Visual-Textual Attentive Semantic Consistency for Medical Report Generation [18]
19. Weakly supervised contrastive learning for chest x-ray report generation [19]
20. A Medical Semantic-Assisted Transformer for Radiographic Report Generation [20]
21. A Self-Guided Framework for Radiology Report Generation [21]
22. AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation [22]
23. An Inclusive Task-Aware Framework for Radiology Report Generation [23]
24. Attention based automated radiology report generation using CNN and LSTM [24]
25. Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation [25]
26. Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach [26]
27. Clinically Coherent Radiology Report Generation with Imbalanced Chest X-rays [27]
28. Cross-modal Memory Networks for Radiology Report Generation [28]
29. Cross-modal Prototype Driven Network for Radiology Report Generation [29]
30. Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation [30]
31. Improving Chest X-Ray Report Generation by Leveraging Warm-Starting [31]
32. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards [32]
33. JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation [33]
34. Multimodal contrastive learning for radiology report generation [34]

35. Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations [35]
36. Prior Guided Transformer for Accurate Radiology Reports Generation [36]
37. Prior Knowledge Enhances Radiology Report Generation [37]
38. Reinforced Cross-modal Alignment for Radiology Report Generation [38]
39. Self adaptive global-local feature enhancement for radiology report generation [39]
40. Transq: Transformer-based semantic query for medical report generation [40]
41. Dynamic graph enhanced contrastive learning for chest x-ray report generation [41]
42. Interactive and explainable region-guided radiology report generation [42]
43. Joint embedding of deep visual and semantic features for medical image report generation [43]
44. KGVl-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation [44]
45. Kiut: Knowledge-injected u-transformer for radiology report generation [45]
46. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens [46]
47. ORGAN: observation-guided radiology report generation via tree reasoning [47]
48. R2gengpt: Radiology report generation with frozen llms [48]
49. Replace and report: NLP assisted radiology report generation [49]
50. Style-aware radiology report generation with radgraph and few-shot prompting [50]
51. Unify, align and refine: Multi-level semantic alignment for radiology report generation [51]
52. Vision transformer and language model based radiology report generation [52]
53. Visual-linguistic causal intervention for radiology report generation [53]
54. Semi-supervised medical report generation via graph-guided hybrid feature consistency [54]
55. Bootstrapping Large Language Models for Radiology Report Generation [55]

56. Large Model driven Radiology Report Generation with Clinical Quality Reinforcement Learning [56]
57. TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation [57]
58. Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation [58]
59. Attribute Prototype-guided Iterative Scene Graph for Explainable Radiology Report Generation [59]
60. LHR-RFL: Linear Hybrid-Reward-Based Reinforced Focal Learning for Automatic Radiology Report Generation [60]

List of State-of-the-art Classification Network in Literature Review 1.4.1

1. ImageNet Classification with Deep Convolutional Neural Networks [61]
2. Very Deep Convolutional Networks for Large-Scale Image Recognition [62]
3. Going deeper with convolutions [63]
4. Deep Residual Learning for Image Recognition [64]
5. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[65]
6. Rethinking the Inception Architecture for Computer Vision [66]
7. Aggregated Residual Transformations for Deep Neural Networks [67]
8. Wide Residual Networks [68]
9. Densely Connected Convolutional Networks[69]
10. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning [70]
11. Squeeze-and-Excitation Networks [71]
12. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[72]
13. MobileNetV2: Inverted Residuals and Linear Bottlenecks[73]
14. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design[74]
15. Searching for MobileNetV3 [75]
16. Res2Net: A New Multi-scale Backbone Architecture[76]
17. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [77]
18. Deep High-Resolution Representation Learning for Visual Recognition [78]
19. CSPNet: A New Backbone that can Enhance Learning Capability of CNN [79]
20. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [80]

21. Designing Network Design Spaces [81]
22. Training data-efficient image transformers and distillation through attention [82]
23. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [83]
24. RepVGG: Making VGG-style ConvNets Great Again [84]
25. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet[85]
26. Conformer: Local Features Coupling Global Representations for Visual Recognition[86]
27. Twins: Revisiting the Design of Spatial Attention in Vision Transformers[87]
28. MetaFormer is Actually What You Need for Vision[88]
29. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection[89]
30. Swin Transformer V2: Scaling Up Capacity and Resolution[90]
31. BEiT: BERT Pre-Training of Image Transformers[91]
32. Learning Transferable Visual Models From Natural Language Supervision[92]
33. EfficientNetV2: Smaller Models and Faster Training[93]
34. XCiT: Cross-Covariance Image Transformers[94]
35. An Empirical Study of Training Self-Supervised Vision Transformers[95]
36. Masked Autoencoders Are Scalable Vision Learners[96]
37. TinyViT: Fast Pretraining Distillation for Small Vision Transformers[97]
38. A ConvNet for the 2020s[98]
39. Visual Attention Network[99]
40. Patches Are All You Need? [100]
41. EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications [101]
42. An Improved One millisecond Mobile Backbone[102]
43. EfficientFormer: Vision Transformers at MobileNet Speed[103]
44. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers[104]

- 45. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation[105]
- 46. Co-designing and Scaling ConvNets with Masked Autoencoders[106]
- 47. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Model [107]

1. Literature Review

1.1 Chest X-Ray and Radiology Report

1.1.1 Chest X-Ray

Chest X-ray (CXR), also known as a chest radiograph (see Figure 1), is a widely used non-invasive imaging technique for evaluating conditions of the thoracic cavity, including the lungs, heart, ribs, and surrounding tissues. Like other imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI), CXR employs ionizing radiation to generate images. However, CXR is particularly valued for its speed, simplicity, and cost-effectiveness, making it the most frequently performed radiological examination globally.

The standard imaging process typically includes two primary views: the posteroanterior (PA) view, capturing the chest from back to front, and the lateral view, taken from the side (illustrated in Figure 2). These complementary perspectives enable a more comprehensive evaluation. For example, while the PA view is effective for assessing the lung fields, cardiac silhouette, and bony structures, the lateral view adds depth perception and helps localize lesions, effusions, or other abnormalities that may not be fully visible from a single angle [108].

CXR is widely used for diagnosing and monitoring both acute and chronic thoracic conditions. Pulmonary applications include pneumonia, pneumothorax, tuberculosis, chronic obstructive pulmonary disease (COPD), and lung cancer. It is also essential for identifying cardiac issues such as cardiomegaly and congestive heart failure, as well as skeletal problems like rib fractures and spinal deformities. Moreover, CXRs assist in evaluating pleural conditions (e.g., pleural effusion, hemothorax) and diaphragmatic abnormalities. The

modality also plays an important role in occupational health screening, especially for lung diseases caused by exposure to asbestos or silica [109]. During the COVID-19 pandemic, CXRs proved vital for detecting lung infections and monitoring disease progression [110].

The global scale of CXR utilization is substantial. According to the World Health Organization (WHO) and the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) [111], more than 4.2 billion diagnostic radiology examinations are conducted worldwide each year, including 40 million nuclear medicine procedures and 8.5 million radiotherapy treatments. In Hong Kong, the Hospital Authority Statistical Report (2023) recorded 4,203,136 radiology attendances between 2022 and 2023, highlighting the substantial demand for radiological services within a single healthcare system. This immense volume of diagnostic imaging underscores the critical role of radiology in modern medicine, enabling timely diagnosis, treatment planning, and monitoring of various health conditions.

However, this demand has outpaced the supply of qualified radiologists. As of 2024, only 472 radiologists were registered as fellows in the Hong Kong College of Radiologists—a figure insufficient to meet growing diagnostic needs. This shortage contributes to delayed report turnaround times and may impact patient care. Additionally, the repetitive nature of interpreting large volumes of CXRs—many of which are routine or normal—can lead to radiologist fatigue and increase the risk of oversight or diagnostic error.

To mitigate these issues, artificial intelligence (AI)-powered solutions have emerged as promising tools in medical imaging workflows. By leveraging machine learning, AI systems can analyze CXRs rapidly and accurately, helping to detect conditions such as pneumonia, lung nodules, or pleural effusions. [112] These models can prioritize urgent cases for human review, reduce diagnostic workloads, and improve consistency by minimizing inter-reader variability. As these technologies mature, their integration into clinical practice holds the

potential to enhance productivity, lower turnaround times, and support better patient outcomes.

In conclusion, chest X-ray remains a fundamental tool in diagnostic radiology due to its speed, accessibility, and clinical utility. While increasing demand has placed strain on radiological services, the rise of AI-based solutions offers a scalable pathway to improve efficiency and maintain diagnostic quality, reinforcing the central role of CXR in modern healthcare.



X-Ray in Department of Diagnostic & Interventional Radiology, Hong Kong Sanatorium & Hospital Limited



X-Ray in Imaging & Interventional Radiology Centre of CUHK Hospital

Figure 1. Sample chest X-ray imaging device used in Hong Kong. The figure illustrates a typical chest X-ray setup, including the imaging equipment and positioning framework commonly used in clinical practice.



posteroanterior (PA) View



Lateral View

Figure 2. Sample chest X-ray images in posteroanterior (PA) and lateral views. The figure presents two standard radiographic projections used in chest imaging. The PA view provides a frontal perspective, while the lateral view offers a complementary side profile, aiding in comprehensive assessment of thoracic structures.

1.1.2 Radiology Report

A radiology report is a structured clinical document that conveys and interprets findings from medical imaging examinations, serving as a critical communication tool in diagnostic radiology. These reports are widely utilized across imaging modalities such as chest X-ray (CXR), computed tomography (CT), and magnetic resonance imaging (MRI), as illustrated in Figure 3. Radiology reports are generally classified into two types: free-text and structured reports. Free-text reports are composed in a narrative form and allow greater flexibility in expression; however, they often follow an implicit structure that can be parsed and converted into a standardized format. Structured reports, in contrast, follow predefined templates aligned with radiological guidelines and practices, offering enhanced clarity, consistency, and suitability for automated processing. Due to these advantages, most publicly available radiology report datasets adopt structured formats, which form the focus of this study, except in cases where free-text reports are explicitly addressed.[113] [114]

In the context of chest radiography, structured reports typically consist of five main sections: Indication, Technique, Comparison, Findings, and Impression. The Indication section outlines the clinical motivation for the examination, incorporating patient demographics, symptoms, and relevant history to contextualize image interpretation. Technique describes the imaging protocols used, such as the posteroanterior (PA) and lateral views for CXRs, ensuring reproducibility and documentation of procedural details. Comparison refers to prior imaging, if available, enabling longitudinal assessment of disease progression or treatment response. Findings provide an objective description of observed anatomical structures and abnormalities, including signs like lung opacities, pleural effusions, or cardiac enlargement. Finally, the Impression synthesizes these findings into a concise diagnostic summary, often including differential diagnoses, follow-up recommendations, and clinical interpretations relevant to decision-making. [115, 116] This structured format ensures effective communication among healthcare providers, supporting coordinated and informed clinical care.

Despite the benefits of structured reporting, the manual creation of radiology reports remains time-consuming and labor-intensive, particularly under increasing imaging workloads. In many healthcare settings, the demand for radiological interpretations outpaces the availability of trained radiologists, leading to reporting delays and potential bottlenecks in patient management. To address these inefficiencies, artificial intelligence (AI) and deep learning techniques have driven the development of automatic radiology report generation systems. These AI-driven approaches aim to alleviate radiologists' workloads by automatically analyzing medical images and generating preliminary drafts of reports. Given that the Indication, Technique, and Comparison sections require specific clinical knowledge and contextual input from physicians, current AI models primarily target the Findings and Impression sections. By automating these diagnostic components, AI-based systems can

enhance reporting efficiency, ensure greater consistency, and reduce human error. As these technologies continue to evolve, they offer promising support for radiologists in managing rising case volumes and improving the speed and quality of diagnostic reporting, ultimately contributing to better patient outcomes.

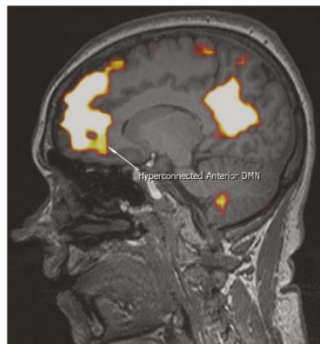


Report

Motivo de consulta: Preoperatorio.

Rx PA tórax:

Impresión diagnóstica: Ateromatosis aórtica calcificada, Engrosamiento pleural biapical, Atelectasia laminar basal izquierda, Elongación aórtica. Sin otros hallazgos radiológicos significativos.



A

MRI OF THE BRAIN

HISTORY

Name: DOB: 12/28/1954 Female

Exam Date: 9/8/14

Referring Phys: GunnarHuser.M.D.

This is a 59-year-old female with exposure to mold and mercury. The patient has symptoms of seizures, memory loss, and numbness in hands and left arm.

PROCEDURE

Using a 3 Tesla Siemens Verio MRI Open system, the following sequences were obtained:

1) Localizer.

4) DWI axial.

2) T1 3D sagittal MPRAGE.

5) SWI axial.

3) T2 FLAIR sagittal.

6) T2 FLAIR axial. 7) T2 TSE axial.

FINDINGS

No cavum septum pellucidum is seen. No blood products are visualized. There is no diffusion restriction.

There is no focal or global volume loss.

The posterior fossa is normal.

The sella and parasellar regions are normal. The flow voids of the vessels of the skull base are identified and normal.

The mastoid air cells, paranasal sinuses, and orbits are normal.

There are scattered foci of T2/FLAIR hyperintensity within the supratentorial subcortical white matter located in bifrontal lobes.

IMPRESSION

Scattered foci of T2/FLAIR hyperintensity in the supratentorial subcortical white matter representing gliosis. This can be seen in the setting of prior head trauma, chronic migraine headaches, less likely chronic small vessel ischemic disease given distribution.

MRI

	<div>0001 REPORT : REASON FOR EXAM (Entered by ordering clinician into CRIS): hx of head and neck cancer. needs scan CT of the nasopharynx.</div> <div>HISTORY: Head and neck cancer.</div> <div>TECHNIQUE: Contiguous 2.5 mm axial images of the nasopharynx were performed without IV contrast. COMPARISON: xx/xx/xxxx.</div> <div>FINDINGS: No soft tissue masses are seen within the soft tissues of the neck. The parotid and submandibular glands are predominantly fatty-replaced. Soft tissues of the Naso, oropharynx are unremarkable. There may be mild fat stranding of the right parapharyngeal soft tissues (series 1001, image 32). No abnormal masses are seen at that site. No bulky lymphadenopathy is seen. There is a fusiform aneurysm of the basilar artery as previously described. It appears to be mildly increased in size and currently measures 2.0 cm in transverse dimensions and previously measured 1.8 cm. It measured 1.5 cm in transverse dimensions on xx/xx/xxxx. Atherosclerotic calcifications are also seen within the carotid arteries bilaterally. There is near-complete opacification of the maxillary sinuses bilaterally. This has increased predominantly within the left maxillary sinus and mildly within the right maxillary sinus. The ethmoidal air cells are clear. Sphenoidal and frontal sinuses are clear. Degenerative changes of the cervical spine are noted.</div> <div>IMPRESSIONS: 1. No soft tissue masses however, mild right parapharyngeal fat stranding is seen it may be postoperative or post radiation in nature. 2. Basilar artery aneurysm that has gradually increased in size when compared to prior examinations. 3. Atherosclerotic disease of the coronary arteries bilaterally.</div>
	<div>0001 Report: CHEST, ABDOMEN, PELVIS CT: Multidetector helical (5 mm, quad) images following, and abdomen images prior to vascular contrast infusion (45 s delay, 2 cc/s, 130 cc Isovue) obtained without apparent complication. History: renal cell pt on Medarex protocol here for end of course evaluation.</div> <div>CHEST: Multiple right, and at least one left lung masses minimally-moderately increasing since xx/xx/xxxx, compatible with metastases despite moderate decrease in at least one right mid-lung mass (e.g. series 4 image 30). Minimal pretracheal and subcarinal adenopathy increasing. Spine osteophytes. Enlargement thyroid on right side, and thyroid heterogeneity unchanged, possibly due to goiter. No evidence of pleural or pericardial effusion, axilla or left hilum adenopathy.</div> <div>ABDOMEN, PELVIS: Few right and left liver foci, left periaortic and left adrenal fossa, and right sacrum mass and lytic lesion (series 3 image 88-95) increasing minimally, compatible with metastases. Scattered lumbar vertebra and bilateral ilium foci (e.g. series 3 image 55, 60, 80, 84-7, 96) foci possibly due to bone metastases. Uterine fundus focus (series 3 image 95) increasing in density since xx/xx/xxxx, possibly due to fibroid, metastasis. No evidence of splenomegaly, hydronephrosis, gallbladder calcification, or bulky mesenteric adenopathy.</div>

CT

Figure 3. The Different imaging modalities of chest radiographs and their corresponding radiology report types. From left to right, the figure presents examples of Chest X-ray, MRI, and CT scans, each associated with distinct reporting styles tailored to their imaging characteristics and diagnostic requirements.

1.1.3 The Conflict Between Low Efficiency of Manual Radiology Report

Writing and the Increasing Demand for Timely Radiological Diagnosis

Radiology reports are essential for communicating diagnostic findings between radiologists and referring clinicians. However, the increasing demand for timely imaging interpretations has created a growing conflict between the limited efficiency of manual report writing and the operational needs of modern healthcare systems. Manual report drafting remains time-consuming and labor-intensive, contributing to workflow bottlenecks and delays in patient care.

Studies [117] estimate that composing a single radiology report manually takes between 1 minute 30 seconds and 2 minutes 30 seconds (Which is shown in Figure 4). While this may seem manageable per case, the cumulative time becomes substantial in high-volume environments. For example, according to data from the Hong Kong Hospital Authority Annual Report (2023), Canadian Medical Imaging Inventory [118], and the UK Royal College of Radiologists [119], radiologists face significant caseloads, especially in the well-welfare countries. In Hong Kong, approximately 1,000 radiologists interpret over 4 million cases annually (Figure 5). In Canada, around 2,000 radiologists are responsible for more than 20 million X-ray studies per year. The UK reports a similar burden, with about 4,000 radiologists managing 22 million X-ray examinations annually. These figures underscore a global shortage of radiologists relative to the rapidly growing diagnostic workload.

This imbalance exacerbates the inefficiencies of manual reporting, resulting in diagnostic delays, increased occupational stress, and prolonged patient waiting times. In response, recent advances in deep learning have enabled the development of automated radiology report generation systems. These systems can generate preliminary report drafts in approximately 0.5 seconds per case—dramatically faster than manual methods. By automating the most time-

consuming aspects of reporting, these systems allow radiologists to focus on validating and refining drafts rather than composing reports from scratch.

Unlike the traditional manual workflow, automated pipelines streamline the reporting process and significantly reduce turnaround times. As highlighted in [120], faster report delivery is directly associated with improved clinical decision-making and better patient management.

In summary, the adoption of deep learning-based report generation technologies offers a scalable and efficient solution to bridge the gap between diagnostic demand and radiologist capacity, enhancing healthcare delivery and relieving the pressure on radiology services.

	Examination	Time mm : ss	RVU	First quartile	Third quartile	Sample size
XR	Abdomen	01:51	1.13	01:03	03:20	4 464
XR	Abdomen + chest	02:28	1.51	01:21	04:22	4 420
XR	Ankle	01:33	0.95	00:53	02:58	1 392
XR	Ankle + foot	02:27	1.50	01:35	04:05	209
XR	Bone Age	05:32	3.39	03:30	08:22	91
XR	Chest	01:38	1.00	00:52	03:09	48 888
XR	Chest + facial bones	02:31	1.54	01:41	06:05	21
XR	Chest, bed	02:14	1.37	01:17	03:56	13 166
XR	Clavicle	01:38	1.00	00:44	02:44	96
XR	Elbow	01:38	1.00	00:53	02:57	716
XR	ERCP (reporting images)	01:30	0.92	00:32	02:37	265
XR	Facial bones	01:57	1.19	01:08	03:36	724
XR	Facial bones + mandible	02:58	1.82	01:33	04:41	139
XR	Femur	01:46	1.08	00:56	03:07	1 122
XR	Finger/thumb	01:24	0.86	00:47	02:28	2 193

Figure 4. The general time of manual radiology report writing in X-Ray. From [117]. It shows that for chest X-Ray, the average time for manual writing of radiology report for each case is 1 min 30s to 2 min 30s.

Service Capacity as at 31 March 2023 截至2023年3月31日的服務容量		
No. of hospital beds 醫院病床數目		30 568
No. of psychiatric day places 精神科日間醫院名額		909
No. of geriatric day places 老人科日間醫院名額		1 017
Service Throughput in 2022-23 2022-23 年度服務量		
Inpatient and Day Inpatient Services 住院及日間住院服務	No. of discharges and deaths 出院人次及死亡人數	1 726 026
	No. of live births 活產嬰兒數目	19 838
	No. of allied health attendances 專職醫療就診人次	6 784 872
Ambulatory Services 日間服務	No. of accident and emergency attendances 急症室就診人次	1 741 091
	No. of specialist outpatient (clinical) attendances 專科門診（臨床）就診人次	8 043 744
	No. of primary care attendances 基層醫療就診人次	5 317 610
	No. of allied health attendances 專職醫療就診人次	3 009 437
	No. of geriatric day attendances 老人科日間醫院就診人次	119 108
	No. of psychiatric day attendances 精神科日間醫院就診人次	79 620
Outreach Services 外展服務	No. of community nurse attendances 接受社區護士服務人次	869 158
	No. of allied health attendances 專職醫療就診人次	27 058
	No. of geriatric outreach attendances 接受老人科外展服務人次	770 143
Other Services 其他服務	No. of psychiatric outreach attendances 接受精神科外展服務人次	272 346
	No. of radiology attendances 放射科檢查病人人次	4 203 136
	No. of operations done inside operating theatres 手術室內進行的手術數目	173 279

Number of Fellows	691
<ul style="list-style-type: none"> Radiology (472) <ul style="list-style-type: none"> Fellows accredited in the Subspecialty of Interventional Radiology (Specialty of Radiology) (74) Clinical Oncology (190) <ul style="list-style-type: none"> Fellows accredited in the Subspecialty of Palliative Medicine (Specialty of Clinical Oncology) (57) Nuclear Medicine (29) 	
Number of Members	235
Number of Trainee Members	74
Number of Associate Members	8
Number of Honorary Fellows	21
Number of Honorary Members	9
Total no. of members as of November 2024	1038

Figure 5. The disparity between patient demand for radiology diagnosis and the availability of expert radiologists. The figure highlights the significant gap in access to radiological expertise, based on data from the 2023 Hospital Authority report and the Hong Kong College of Radiologists. This shortage contributes to increased workload, longer diagnosis times, and potential delays in patient care.

1.2 Deep-Learning based Radiology Report Generation

1.2.1 From Traditional AI to Deep Learning in radiology: The Expanding Scale of AI Models.

The development of artificial intelligence (AI) in radiology report generation has closely followed the evolution of AI models, particularly in terms of scale and complexity. From early machine learning approaches to modern deep learning frameworks, progress has been driven by increasing computational power and data availability, leading to progressively larger and more capable AI models.

Early AI applications in radiology reporting relied on traditional machine learning techniques, using small-scale models such as Support Vector Machines (SVMs) and Logistic Regression [121]. These models processed manually extracted handcrafted features, with relatively few parameters, as their performance depended more on feature engineering than on model size. Given the computational limitations at the time, their primary function was to classify findings as normal or abnormal rather than generating detailed radiology reports.

A significant breakthrough came with the introduction of deep learning, particularly Convolutional Neural Networks (CNNs), which enabled automatic feature extraction from medical images. Unlike traditional models that relied on predefined features, CNNs learned hierarchical representations directly from data. Early architectures such as AlexNet [61] and VGG [62] contained millions of parameters—substantially larger than machine learning models—but were still limited by available GPU power and dataset sizes. Simultaneously, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were introduced for text generation, making them well-suited for generating short paragraphs to describe images. However, due to their sequential nature, RNNs and LSTMs were

computationally expensive and difficult to scale. The largest models of this era contained tens of millions of parameters but struggled to generate long-form, coherent radiology reports.

With increasing computational resources, AI models began integrating both image and text processing within a unified framework. Encoder-decoder architectures emerged as the dominant paradigm for radiology report generation, where CNNs extracted image features and RNNs or LSTMs converted them into descriptive text. During this phase, model scale increased significantly, reaching hundreds of millions of parameters. The introduction of attention mechanisms[122] and sequence-to-sequence (Seq2Seq) [123] models further improved performance by selectively focusing on relevant image features when generating text. Large-scale datasets such as IU X-Ray [124] and MIMIC-CXR [125] became available, allowing models to learn from diverse clinical cases. Despite these advancements, training such models from scratch remained computationally intensive, requiring dedicated high-performance computing resources.

In recent years, AI model scale has grown exponentially with the rise of large-scale pretrained models. Unlike earlier deep learning approaches that required training from scratch on specific medical datasets, these models are first trained on massive, diverse datasets and then fine-tuned for specialized tasks. Multimodal models like CLIP[92], which align images with textual descriptions, and large language models such as GPT [126] have transformed radiology report generation. These models contain billions of parameters—orders of magnitude larger than previous architectures—allowing them to generate highly coherent, context-aware reports. However, their computational demands are substantial, often requiring multiple GPUs or large-scale distributed computing. While these models offer significant performance improvements, their reliance on vast computational resources and large-scale datasets presents challenges for real-world medical applications.

The evolution of AI in radiology has been characterized by a continuous increase in model scale, from small-scale machine learning classifiers to billion-parameter deep learning models. While larger models generally offer superior performance and robustness, they also require significant resources, raising concerns about accessibility and deployment in clinical settings. Moving forward, balancing model scale with efficiency and practicality will be essential for ensuring the widespread adoption of AI-driven radiology applications.

1.2.2 Vision-Language Deep Learning Application

As the demand for automated language processing increases in various fields constrained by limited manpower and time, it has become essential to develop algorithms capable of efficiently processing both simple, large-scale language data and complex contextual information. With advances in deep learning and growing computational power, language models such as BERT [127] and Transformer [122] have been introduced for natural language processing (NLP) tasks, revolutionizing automated language analysis. These models enable both text comprehension and generation and have been successfully applied to tasks such as text classification, generation, summarization, and various downstream applications. By leveraging large-scale datasets and complex architectures, they reduce workload and improve performance across a wide range of NLP tasks.

More recently, the application of transformer structures in computer vision has highlighted structural similarities between models used for visual and language processing. This synergy has led to the emergence of vision-language integration as a promising research area, where text can support image understanding and images can provide high-level contextual information for language tasks. Vision-language applications in natural image domains include image captioning, visual recognition, and text-to-image generation, as illustrated in Figure 6(a).

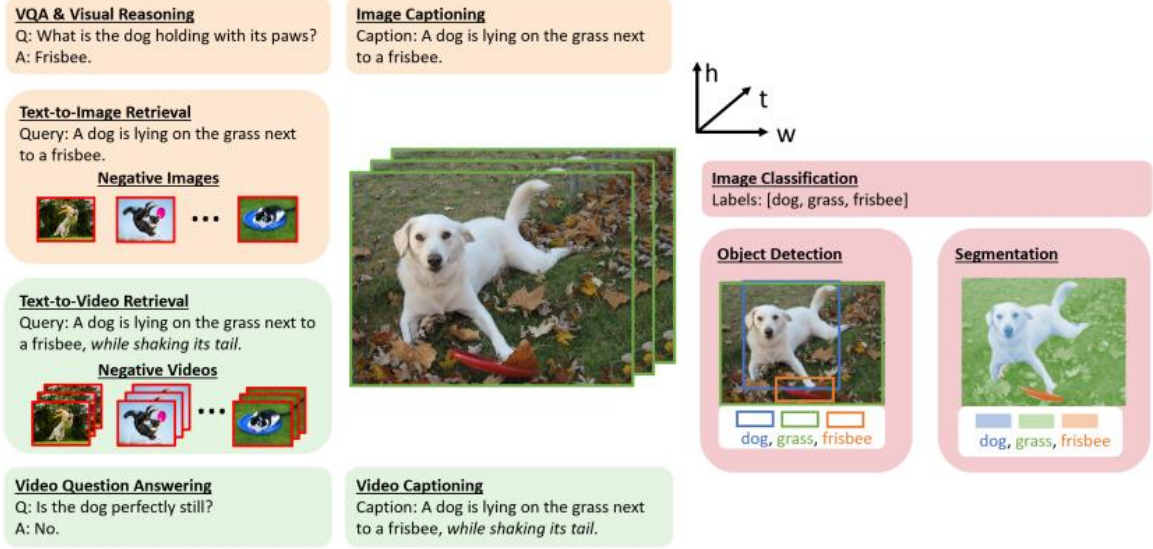
In medical imaging, vision-language applications offer even greater potential due to the critical role of image interpretation in diagnostics. The integration of medical images with corresponding textual data enables automated generation of descriptive content and supports the identification of key diagnostic features. This is particularly relevant in the context of the global radiologist shortage, where such technologies can enhance diagnostic efficiency and accessibility.

Building upon developments in the natural image domain, similar vision-language methods have been increasingly adopted in medical imaging for tasks such as image-text classification ([128], [129], [130]), visual question answering (VQA) ([131], [132]), and medical object detection guided by text ([133], [132]). In these applications, images often serve as the primary input, while associated text provides valuable context. The use of paired image-text datasets is essential for capturing the complex and interdependent relationships between visual and textual components. Prior studies have demonstrated the feasibility of using vision-language learning frameworks for multimodal classification, retrieval, and question answering in medical scenarios.

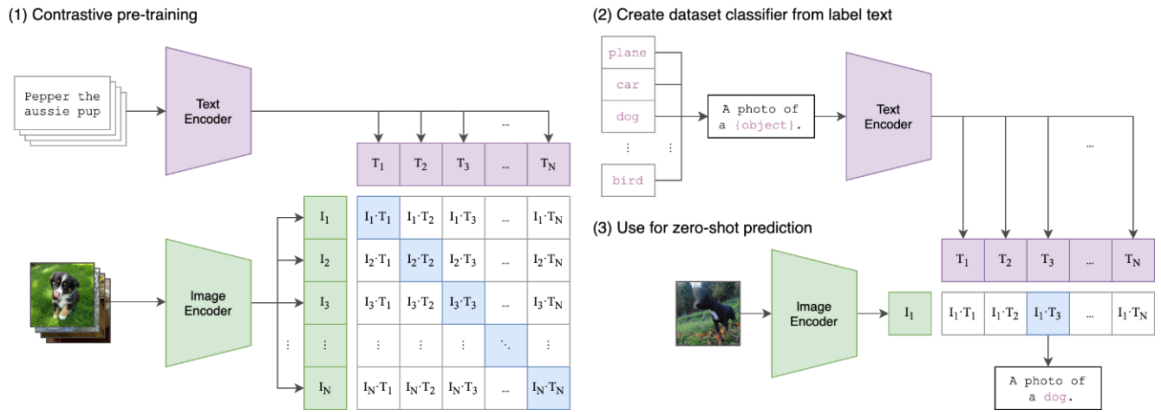
A major advancement in this area is the Contrastive Language-Image Pre-Training (CLIP) framework proposed by OpenAI ([92]), as shown in Figure 6(b). CLIP enables training on large-scale image-text pairs, eliminating the need to train models from scratch for each new task and providing a robust zero-shot prediction capability. This framework has been widely adopted in the medical domain for both fine-tuning and zero-shot inference across a variety of tasks ([134], [135], [136]). These include disease classification ([137], [138], [139]), medical visual question answering ([140], [141], [142]), and representation learning. Models like MedCLIP [130] and PubMedCLIP [140] follow a similar architecture to CLIP but are pretrained specifically on medical text and image pairs, addressing domain-specific

challenges in healthcare applications. Leveraging CLIP as a pre-trained model has led to numerous applications in the medical domain.

Despite their potential, general vision-language models such as CLIP encounter limitations when applied to dynamic medical contexts. Models trained on fixed datasets may have difficulty generalizing across varied clinical environments, particularly when faced with unfamiliar diseases or rare diagnostic patterns not represented in the training data. Although CLIP's large-scale training provides strong generalization in natural image tasks, it may underperform in clinical applications where variability in medical language, image modality, and reporting styles introduces additional complexity. In tasks such as radiology report generation, inconsistencies in clinical language and diverse writing conventions further challenge the robustness of pretrained models. These limitations underscore the need for continued development to improve adaptability and ensure reliable performance of vision-language models across high-variability medical scenarios.



(a) The overview illustration of VL tasks, including the transformation between image and text (in orange), between video and text (in green) and utilizing text as auxiliary information (in pink)



(b) CLIP model pipeline, which is also the general format of utilizing language texts as the auxiliary information to help pretraining in vision-related tasks.

Figure 6. Overview of representative tasks in vision-language (VL) problems and the CLIP model pipeline. (a) Illustration of VL tasks, including image-text transformation (orange), video-text transformation (green), and the use of text as auxiliary information in vision tasks (pink) [143]. (b) The CLIP model pipeline, representing a general framework for leveraging language text as auxiliary information to enhance pretraining in vision-related tasks [92]. Additionally, various CLIP-based models, such as MedCLIP[130] and PubmedCLIP[140], have been introduced in medical vision-language applications, differing primarily in their training language datasets.

1.2.3 Current Radiology Report Generation Work: Dataset and Algorithms

Radiology reports are essential for interpreting medical images such as X-rays, CT scans, and MRIs ([144]), providing critical information for diagnosis, treatment planning, and follow-up care. Traditionally, radiologists manually write these reports, analyzing medical images and documenting their findings. However, with the increasing demand for radiology services, manual writing of reports has become inefficient and often leads to delays in report turnaround, subsequently impacting timely patient diagnosis and treatment([145]).

To address these challenges, deep learning-based radiology report generation has emerged as a promising solution, automating the writing process by extracting visual findings from medical images and translating them into textual descriptions. ([146])

In the early stages of radiology report generation research, foundational datasets and benchmarks such as IU X-Ray ([124]) and MIMIC-CXR ([125]) were introduced, offering chest X-ray images paired with free-text radiology reports. The IU X-Ray dataset consists of 7,470 image-report pairs, while the larger MIMIC-CXR dataset includes 377,110 images linked to 227,835 radiographic studies, alongside additional information such as EEG records. These datasets have been instrumental in validating novel approaches to radiology report generation, driving significant advancements in the field.

In this study, we review current radiology report generation methods applied to these two public datasets. Tables 1 and 2 summarize the dataset characteristics and corresponding performance metrics reported in the original studies for IU X-Ray and MIMIC-CXR, respectively.

The seminal study by [1] introduced a deep learning framework for radiology report generation based on an encoder-decoder architecture. In this model, the feature extractor captures detailed

visual information from medical images, while the text decoder generates radiology reports that mimic human-written descriptions. This encoder-decoder framework, which is shown in Figure 7 as the general pipeline, has since established itself as the standard approach in the field, paving the way for numerous follow-up studies and innovations.

Recent efforts in research have aimed to improve both the image encoder and text decoder to make radiology report generation more accurate and meaningful. For the encoders, more advanced network designs have been developed to capture finer visual details ([8], [20], [22]). Attention mechanisms have also been introduced to better align the encoded features with the content of the reports ([4], [12]). For the development of decoder, more sophisticated designs ([15], [44]) are being used to generate text that captures the detailed and specific language used in radiology reports.

Efforts such as utilizing knowledge graph (RadGraph([147]), [35]; [9]) have further enriched this field by extracting clinical entities and their relationships to build knowledge graphs, enabling more structured and interpretable report generation. These approaches mimic human reasoning and improve coherence and clinical relevance.

Despite these advancements, the traditional encoder-decoder pipeline has significant limitations. The quality of generated reports heavily depends on the features extracted by the encoder, which lack control. This dependency limits the model's adaptability to new clinical settings, reducing its utility across diverse environments. Additionally, the text decoder relies solely on the information provided by extracted features to generate text. This method lacks a broader linguistic understanding, hindering the generation of coherent, high-quality reports.

To overcome these issues, recent work has incorporated large language models (LLMs) such as Llama2 ([48]) and ChatGPT ([148]). These models, pretrained on vast language corpora, enhance fluency and readability when integrated with vision-language mapping. However,

many of these approaches still depend on unexplainable features or rigid templates, posing challenges in terms of interpretability and generalization.

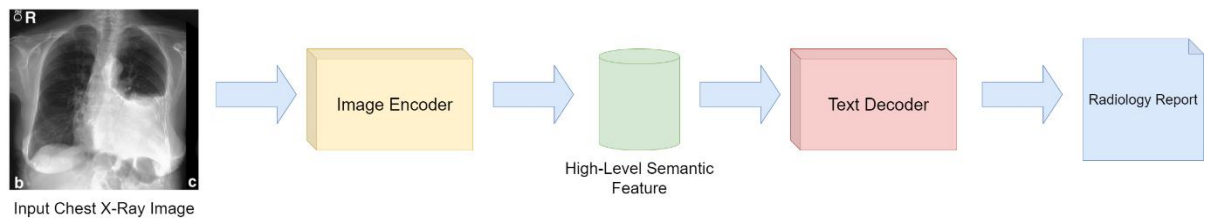


Figure 7. General pipeline for radiology report generation. The process begins with an input chest X-ray image, which is processed by an image encoder to extract high-level semantic features that are not directly interpretable by humans. A text decoder then utilizes these features to generate a corresponding radiology report in a structured paragraph format.

Table 1. Comparison of state-of-the-art radiology report generation methods on the IU X-Ray dataset. The table highlights the image encoder, text decoder, and performance metrics for each method. Notably, for methods incorporating a language model during text decoding, the language model is listed under the "Text Decoder" column. The reference list of the state-of-the-art radiology report generation methods is shown in the next table.

Work	Year	Model (Encoder)	Model (Decoder)	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Jing et al. (2017)	2017	CNN	LSTM	0.517	0.386	0.306	0.247	0.447	0.217	0.327
Xue, Xu, Long, Xue, Antani, Thoma and Huang (2018)	2018	CNN	LSTM	0.464	0.358	0.270	0.195	0.366	0.274	/
Harzig, Chen, Chen and Lienhart (2019)	2019	CNN(ResNet-152)	LSTM	0.373	0.246	0.175	0.126	0.315	0.163	0.359
Xie et al. (2019)	2019	CNN	LSTM	0.443	0.337	0.236	0.181	0.347	/	0.374
Yuan, Liao, Luo and Luo (2019)	2019	CNN(ResNet-152)	LSTM	0.529	0.372	0.315	0.255	0.453	0.343	/
Li, Liang, Hu and Xing (2019)	2019	CNN	Transformer	0.482	0.325	0.226	0.162	0.339	/	0.280
Jing, Wang and Xing (2019)	2019	CNN	LSTM	0.464	0.301	0.210	0.154	0.362	/	0.275
Chen et al. (2020b)	2020	Transformer	Transformer	0.470	0.304	0.219	0.165	0.371	0.187	/
Zhang et al. (2020b)	2020	CNN (DenseNet-121)	LSTM	0.441	0.291	0.203	0.147	0.367	/	0.304
Wang, Zhou, Wang and Li (2021)	2021	CNN	LSTM	0.487	0.346	0.270	0.208	0.359	/	0.452
Alfaraghaly, Khaled, Elkorany, Helal and Fahmy (2021)	2021	CheXnet (DenseNet121-CNN)	GPT2	0.387	0.245	0.166	0.111	0.289	0.164	0.257
Liu et al. (2021b)	2021	CNN(ResNet-50)	LSTM	0.492	0.314	0.222	0.169	0.381	0.193	/
Liu, Wu, Ge, Fan and Zou (2021a)	2021	Transformer	Transformer	0.483	0.315	0.224	0.168	0.376	0.190	0.351
Yang, Wu, Ge, Zhou and Xiao (2021b)	2021	CNN	Transformer	0.496	0.327	0.238	0.178	0.381	/	0.382
Nooralahzadeh et al. (2021)	2021	CNN	Transformer	0.486	0.317	0.232	0.173	0.390	0.192	/
Yang, Wu, Ge, Wu, Zhou and Xiao (2021a)	2021	CNN	Transformer	0.497	0.319	0.230	0.174	0.399	/	0.407
Zhou, Huang, Zhou, Fu and Shao (2021)	2021	CNN	Transformer	0.536	0.391	0.314	0.252	0.448	0.228	0.339
Li, Li, Hu and Tao (2022)	2022	CNN+Transformer	Transformer	0.467	0.334	0.261	0.215	0.415	0.201	/
You et al. (2022a)	2022	Transformer	Transformer	0.484	0.313	0.225	0.173	0.379	0.204	/
Wang, Ning, Lu, Wei, Zheng and Iian Chen (2022b)	2022	CNN	Transformer	0.505	0.340	0.247	0.188	0.382	0.208	/
Sirshar, Paracha, Akram, Alghamdi, Zaidi and Fatima (2022)	2022	CNN	LSTM	0.580	0.342	0.263	0.155	/	/	/
Yan, Cheung, Chiu, Tong, Cheung and See (2022b)	2022	CNN	Transformer	/	/	0.256	/	0.341	/	0.380
Wang, Han, Wang, Li and Zhou (2022d)	2022	Transformer	Transformer	0.496	0.319	0.241	0.175	0.377	/	0.449
Yu and Zhang (2022)	2022	CNN(ResNet-152)	Transformer	0.457	0.305	0.216	0.171	0.391	/	0.426
Chen, Shen, Song and Wan (2022)	2022	CNN	Transformer	0.475	0.309	0.222	0.170	0.375	0.191	/
Wang, Bhale Rao and He (2022a)	2022	CNN(ResNet-101)	Transformer	0.525	0.357	0.262	0.199	0.411	0.220	0.359
Nicolson, Dowling and Koopman (2022)	2022	Transformer	Transformer	0.473	0.303	0.224	0.175	0.375	0.199	0.693
Delbrouck, Chambon, Blüthgen, Tsai, Almusa and Langlotz (2022)	2022	CNN	BERT	/	/	/	0.121	0.306	/	/
You, Li, Okumura and Suzuki (2022b)	2022	CNN (ResNet)	Transformer	0.479	0.319	0.222	0.174	0.377	0.193	/
Wu, Li, Wang and Qian (2022)	2022	CNN	LSTM	0.458	0.324	0.238	0.180	0.369	0.206	0.287
Yan, Pei, Zhao, Shan and Tian (2022a)	2022	CNN	Transformer	0.482	0.313	0.232	0.181	0.381	0.203	0.735
Wang, Tang, Lin, Shih, Ding and Peng (2022c)	2022	Graph Convolution Network, CNN	Transformer	0.450	0.301	0.213	0.158	0.384	/	0.340
Qin and Song (2022)	2022	CNN	Transformer	0.494	0.321	0.235	0.181	0.384	0.201	/
Tanwani, Barral and Freedman (2022)	2022	CNN (ResNeXt-101)	BERT, Transformer	0.580	0.440	0.320	0.270	/	/	/
Wang, Wang, Liu, Gao, Zhang and Wang (2023a)	2022	CNN (ResNet-101)	Transformer	0.505	0.345	0.243	0.176	0.396	0.205	/
Kong, Huang, Kuang, Zhu and Wu (2022)	2022	Transformer	Transformer	0.484	0.333	0.238	0.175	0.415	0.207	/
Li, Lin, Chen, Lin, Liang and Chang (2023a)	2023	Transformer	Transformer	/	/	/	0.163	0.383	0.193	0.586
Yang, Yu, Zhang, Han, Jiang and Huang (2021c)	2023	CNN	LSTM	0.478	0.344	0.248	0.180	0.398	/	0.439
Kale et al. (2023a)	2023	CNN(ResNet-152)	BART	0.423	0.256	0.194	0.165	0.444	0.150	/
Huang, Zhang and Zhang (2023)	2023	CNN(ResNet-101)	Transformer	0.525	0.360	0.251	0.185	0.409	0.242	/
Wang, Liu, Wang and Zhou (2023b)	2023	Transformer	Transformer	0.483	0.322	0.228	0.172	0.380	0.192	0.435
Hou, Xu, Cheng, Li and Liu (2023)	2023	CNN	Transformer	0.510	0.346	0.255	0.195	0.399	0.200	/
Wang et al. (2023c)	2023	Swin-Transformer	LLAMA2	0.488	0.316	0.228	0.173	0.377	0.211	0.438
Kale, Jadhav et al. (2023b)	2023	CNN	Transformer	0.402	0.322	0.285	0.170	0.567	0.455	0.473
Li, Yang, Cheng, Zhu, Li and Zou (2023b)	2023	VAE	Transformer	0.530	0.365	0.263	0.200	0.405	0.218	0.501
Mohsan, Akram, Rasool, Alghamdi, Baqai and Abbas (2022)	2023	Transformer	Transformer	0.532	0.344	0.233	0.158	0.387	0.218	0.500
Chen, Liu, Wang, Li, Zhu and Lin (2023)	2023	Transformer	Transformer	0.505	0.334	0.245	0.190	0.394	0.210	0.592
Zhang, Jiang, Zhang, Huang, Fan, Yu and Han (2023a)	2024	CNN	Transformer	0.482	0.310	0.221	0.165	0.377	0.195	/
Liu, Tian, Chen, Song and Zhang (2024)	2024	MiniGPT-4	MiniGPT-4	0.499	0.323	0.238	0.184	0.390	0.208	/
Zhou, Shi, Wei, Alabi, Yue and Vercouteren (2024)	2024	GPT4, CLIP	BLIP-2	/	/	/	0.208	0.387	0.216	/
Yi, Fu, Liu, Zhang and Hua (2024a)	2024	CNN (ResNet 101)	Transformer	0.500	0.349	0.256	0.194	0.402	0.218	/
Parres, Albiol and Paredes (2024)	2024	Swin Transformer	BERT	/	/	/	0.149	0.341	/	/
Yi, Fu, Yu, Liu, Zhang and Hua (2024b)	2024	CNN (ResNet-101)	Transformer	0.539	0.380	0.278	0.210	0.416	0.223	/

Work	Year	Full Paper
Jing et al. (2017)	2017	On the Automatic Generation of Medical Imaging Reports
Xue et al. (2018)	2018	Multimodal Recurrent Model with Attention for Automated Radiology Report Generation
Harzig et al. (2019)	2019	Addressing Data Bias Problems for Chest X-ray Image Report Generation
Xie et al. (2019)	2019	Attention-Based Abnormal-Aware Fusion Network for Radiology Report Generation
Yuan et al. (2019)	2019	Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment
Li et al. (2019)	2019	Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation
Jing et al. (2019)	2019	Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports
Chen et al. (2020)	2020	Generating Radiology Reports via Memory-driven Transformer
Zhang et al. (2020)	2020	When Radiology Report Generation Meets Knowledge Graph
Wang et al. (2021)	2021	A Self-boosting Framework for Automated Radiographic Report Generation
Alfarghaly et al. (2021)	2021	Automated radiology report generation using conditioned transformers
Liu et al. (2021b)	2021	Contrastive Attention for Automatic Chest X-ray Report Generation
Liu et al. (2021a)	2021	Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation
Yang et al. (2021b)	2021	Knowledge matters: Chest radiology report generation with general and specific knowledge
Nooralahzadeh et al. (2021)	2021	Progressive Transformer-Based Generation of Radiology Reports
Yang et al. (2021a)	2021	Radiology Report Generation with a Learned Knowledge Base and Multi-modal Alignment
Zhou et al. (2021)	2021	Visual-textual Attentive Semantic Consistency for Medical Report Generation
Li et al. (2022)	2022	A Self-Guided Framework for Radiology Report Generation
You et al. (2022a)	2022	AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation
Wang et al. (2022b)	2022	An Inclusive Task-Aware Framework for Radiology Report Generation
Sirshar et al. (2022)	2022	Attention based automated radiology report generation using CNN and LSTM
Yan et al. (2022b)	2022	Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation
Wang et al. (2022d)	2022	Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach
Yu and Zhang (2022)	2022	Clinically Coherent Radiology Report Generation with Imbalanced Chest X-rays
Chen et al. (2022)	2022	Cross-modal Memory Networks for Radiology Report Generation
Wang et al. (2022a)	2022	Cross-modal prototype driven network for radiology report generation
Nicolson et al. (2022)	2022	Improving Chest Xray Report Generation by Leveraging Warm Starting
Delbrouck et al. (2022)	2022	Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards
You et al. (2022b)	2022	JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation
Wu et al. (2022)	2022	Multimodal contrastive learning for radiology report generation
Yan et al. (2022a)	2022	Prior Guided Transformer for Accurate Radiology Reports Generation
Wang et al. (2022c)	2022	Prior Knowledge Enhances Radiology Report Generation
Qin and Song (2022)	2022	Reinforced Cross-modal Alignment for Radiology Report Generation
Tanwani et al. (2022)	2022	Repsnet: Combining vision with language for automated medical reports
Wang et al. (2023a)	2022	Self Adaptive Global-Local Feature Enhancement for Radiology Report Generation
Kong et al. (2022)	2022	TransQ: Transformer-Based Semantic Query for Medical Report Generation
Li et al. (2023a)	2023	Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation
Yang et al. (2021c)	2023	Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation
Kale et al. (2023a)	2023	KGVL-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation
Huang et al. (2023)	2023	KiUT: Knowledge-injected U-Transformer for Radiology Report Generation
Wang et al. (2023b)	2023	METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens
Hou et al. (2023)	2023	ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning
Wang et al. (2023c)	2023	R2GenGPT: Radiology Report Generation with Frozen LLMs
Kale et al. (2023b)	2023	Replace and Report: NLP Assisted Radiology Report Generation
Li et al. (2023b)	2023	Unify, Align and Refine: Multi-Level Semantic Alignment for Radiology Report Generation
Mohsan et al. (2022)	2023	Vision Transformer and Language Model Based Radiology Report Generation
Chen et al. (2023)	2023	Visual-Linguistic Causal Intervention for Radiology Report Generation
Zhang et al. (2023)	2024	Semi-supervised Medical Report Generation via Graph-guided Hybrid Feature Consistency
Liu et al. (2024)	2024	Bootstrapping Large Language Models for Radiology Report Generation
Zhou et al. (2024)	2024	Large Model driven Radiology Report Generation with Clinical Quality Reinforcement Learning
Yi et al. (2024a)	2024	TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation
Parres et al. (2024)	2024	Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation
Yi et al. (2024b)	2024	LHR-RFL: Linear Hybrid-Reward-Based Reinforced Focal Learning for Automatic Radiology Report Generation

Table 2. Comparison of state-of-the-art radiology report generation methods on the MIMIC-CXR dataset. The table presents the image encoder, text decoder, and associated performance metrics for each method. Methods that incorporate a language model for text decoding are explicitly listed with the language model under the "Text Decoder" column. The reference list of the state-of-the-art radiology report generation methods is shown in the next table.

Work	Year	Model (Encoder)	Model (Decoder)	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Chen et al. (2020b)	2020	Transformer	Transformer	0.353	0.218	0.145	0.103	0.277	0.142	/
Liu et al. (2021b)	2021	CNN(ResNet-50)	LSTM	0.350	0.219	0.152	0.109	0.283	0.151	/
Liu et al. (2021a)	2021	Transformer	Transformer	0.360	0.224	0.149	0.106	0.284	0.149	0.237
Yang et al. (2021b)	2021	CNN	Transformer	0.363	0.228	0.156	0.115	0.284	/	0.203
Nooralahzadeh et al. (2021)	2021	CNN	Transformer	0.378	0.232	0.154	0.107	0.272	0.145	/
Yang et al. (2021a)	2021	CNN	Transformer	0.386	0.237	0.157	0.111	0.274	/	0.111
Hou, Kaissis, Summers and Kainz (2021)	2021	CNN	Transformer	0.232	/	/	/	0.240	0.101	0.493
Zhou et al. (2021)	2021	CNN	Transformer	0.372	0.241	0.168	0.123	0.335	0.190	1.121
Yan, He, Lu, Du, Chang, Gentili, McAuley and Hsu (2021)	2021	Transformer	BERT	0.373	/	/	0.107	0.274	0.144	/
Wang et al. (2022e)	2022	Transformer	Transformer	0.413	0.266	0.186	0.136	0.298	0.170	0.429
You et al. (2022a)	2022	Transformer	Transformer	0.378	0.235	0.156	0.112	0.283	0.158	/
Wang et al. (2022b)	2022	CNN	Transformer	0.395	0.253	0.170	0.121	0.284	0.147	/
Yan et al. (2022b)	2022	CNN	Transformer	/	/	0.145	/	0.225	/	0.160
Wang et al. (2022d)	2022	Transformer	Transformer	0.351	0.223	0.157	0.118	0.287	/	0.281
Yu and Zhang (2022)	2022	CNN(ResNet-152)	Transformer	0.347	0.235	0.149	0.106	0.280	/	0.552
Chen et al. (2022)	2022	CNN	Transformer	0.353	0.218	0.148	0.106	0.278	0.142	/
Wang et al. (2022a)	2022	CNN(ResNet-101)	Transformer	0.344	0.215	0.146	0.105	0.279	0.138	/
Nishino, Miura, Taniguchi, Ohkuma, Suzuki, Kido and Tomiyama (2022)	2022	CNN	LSTM	/	/	/	0.168	0.122	/	/
Nicolson et al. (2022)	2022	Transformer	Transformer	0.392	0.247	0.171	0.126	0.286	0.154	0.389
Delbrouck et al. (2022)	2022	CNN	BERT	/	/	/	0.116	0.259	/	/
Wu et al. (2022)	2022	CNN	LSTM	0.34	0.212	0.145	0.103	0.270	0.139	0.109
Serra et al. (2022)	2022	CNN(ResNet-101)	Transformer	0.363	0.245	0.178	0.136	0.313	0.161	/
Yan et al. (2022a)	2022	CNN	Transformer	0.356	0.222	0.151	0.111	0.280	0.140	0.154
Qin and Song (2022)	2022	CNN	Transformer	0.381	0.232	0.155	0.109	0.287	0.151	/
Wang et al. (2023a)	2022	CNN (ResNet-101)	Transformer	0.363	0.235	0.164	0.118	0.301	0.136	/
Kong et al. (2022)	2022	Transformer	Transformer	0.423	0.261	0.171	0.116	0.286	0.168	/
Li et al. (2023a)	2023	Transformer	Transformer	/	/	/	0.109	0.284	0.150	0.281
Tanida, Müller, Kaissis and Rueckert (2023)	2023	CNN	Transformer	0.373	0.249	0.175	0.126	0.264	0.168	0.495
Yang et al. (2021c)	2023	CNN	LSTM	0.362	0.251	0.188	0.143	0.326	/	0.273
Huang et al. (2023)	2023	CNN(ResNet-101)	Transformer	0.393	0.243	0.159	0.113	0.285	0.160	/
Wang et al. (2023b)	2023	Transformer	Transformer	0.386	0.250	0.169	0.124	0.291	0.152	0.362
Hou et al. (2023)	2023	CNN	Transformer	0.407	0.256	0.172	0.123	0.293	0.162	/
Wang et al. (2023c)	2023	Swin-Transformer	LLAMA2	0.411	0.267	0.186	0.134	0.297	0.160	0.269
Kale et al. (2023b)	2023	CNN	Transformer	0.253	0.188	0.169	0.163	0.348	0.268	0.331
Li et al. (2023b)	2023	VAE	Transformer	0.363	0.229	0.158	0.107	0.289	0.157	0.246
Chen et al. (2023)	2023	Transformer	Transformer	0.400	0.245	0.165	0.119	0.28	0.150	0.190
Zhang et al. (2023a)	2024	CNN	Transformer	0.362	0.229	0.157	0.113	0.284	0.153	/
Liu et al. (2024)	2024	MiniGPT-4	MiniGPT-4	0.402	0.262	0.180	0.128	0.291	0.175	/
Zhou et al. (2024)	2024	GPT4, CLIP	BLIP-2	/	/	/	0.122	0.296	0.165	/
Yi et al. (2024a)	2024	CNN (ResNet 101)	Transformer	0.398	0.248	0.169	0.121	0.281	0.149	/
Parres et al. (2024)	2024	Swin Transformer	BERT	/	/	/	0.116	0.265	/	/
Zhang, Yang, Yu, Fan, Jiang, Huang and Han (2024)	2024	Transformer	Transformer	0.391	0.258	0.182	0.129	0.282	0.175	0.526
Yi et al. (2024b)	2024	CNN (ResNet-101)	Transformer	0.400	0.253	0.171	0.120	0.296	0.154	/

Work	Year	Full Paper
Chen et al. (2020)	2020	Generating Radiology Reports via Memory-driven Transformer
Liu et al. (2021b)	2021	Contrastive Attention for Automatic Chest X-ray Report Generation
Liu et al. (2021a)	2021	Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation
Yang et al. (2021b)	2021	Knowledge matters: Chest radiology report generation with general and specific knowledge
Nooralahzadeh et al. (2021)	2021	Progressive Transformer-Based Generation of Radiology Reports
Yang et al. (2021a)	2021	Radiology Report Generation with a Learned Knowledge Base and Multi-modal Alignment
Hou et al. (2021)	2021	RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting
Zhou et al. (2021)	2021	Visual-textual Attentive Semantic Consistency for Medical Report Generation
Yan et al. (2021)	2021	Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation
Wang et al. (2022d)	2021	A Medical Semantic-Assisted Transformer for Radiographic Report Generation
You et al. (2022)	2022	AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation
Wang et al. (2022b)	2022	An Inclusive Task-Aware Framework for Radiology Report Generation
Yan et al. (2022b)	2022	Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation
Wang et al. (2022c)	2022	Automated Radiographic Report Generation Purely on Transformer: A Multicriteria Supervised Approach
Yu and Zhang (2022)	2022	Clinically Coherent Radiology Report Generation with Imbalanced Chest X-rays
Chen et al. (2022)	2022	Cross-modal Memory Networks for Radiology Report Generation
Wang et al. (2022a)	2022	Cross-modal prototype driven network for radiology report generation
Nishino et al. (2022)	2022	Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation
Nicolson et al. (2022)	2022	Improving Chest Xray Report Generation by Leveraging Warm Starting
Delbrouck et al. (2022)	2022	Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards
Wu et al. (2022)	2022	Multimodal contrastive learning for radiology report generation
Serra et al. (2022)	2022	Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations
Yan et al. (2022a)	2022	Prior Guided Transformer for Accurate Radiology Reports Generation
Qin and Song (2022)	2022	Reinforced Cross-modal Alignment for Radiology Report Generation
Wang et al. (2023a)	2022	Self Adaptive Global-Local Feature Enhancement for Radiology Report Generation
Kong et al. (2022)	2022	TransSQ: Transformer-Based Semantic Query for Medical Report Generation
Li et al. (2023a)	2023	Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation
Tanida et al. (2023)	2023	Interactive and Explainable Region-guided Radiology Report Generation
Yang et al. (2021c)	2023	Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation
Huang et al. (2023)	2023	KiUT: Knowledge-injected U-Transformer for Radiology Report Generation
Wang et al. (2023b)	2023	METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens
Hou et al. (2023)	2023	ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning
Wang et al. (2023c)	2023	R2GenGPT: Radiology Report Generation with Frozen LLMs
Kale et al. (2023)	2023	Replace and Report: NLP Assisted Radiology Report Generation
Li et al. (2023b)	2023	Unify, Align and Refine: Multi-Level Semantic Alignment for Radiology Report Generation
Chen et al. (2023)	2023	Visual-Linguistic Causal Intervention for Radiology Report Generation
Zhang et al. (2023)	2024	Semi-supervised Medical Report Generation via Graph-guided Hybrid Feature Consistency
Liu et al. (2024)	2024	Bootstrapping Large Language Models for Radiology Report Generation
Zhou et al. (2024)	2024	Large Model driven Radiology Report Generation with Clinical Quality Reinforcement Learning
Yi et al. (2024a)	2024	TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation
Parres et al. (2024)	2024	Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation
Zhang et al. (2024)	2024	Attribute Prototype-guided Iterative Scene Graph for Explainable Radiology Report Generation
Yi et al. (2024b)	2024	LHR-RFL: Linear Hybrid-Reward-Based Reinforced Focal Learning for Automatic Radiology Report Generation

1.3 Automatic Keyword Adaption

1.3.1 Radiology Dictionary

A radiology dictionary serves as a reference tool that standardizes the terminology used in radiological practice, promoting accuracy and consistency in both clinical communication and keyword-based research applications. In radiology report drafting, radiologists often consult such dictionaries to ensure the terms they use are appropriate, unambiguous, and aligned with accepted medical language.

However, creating a universal radiology dictionary that comprehensively covers all imaging modalities, anatomical regions, and clinical scenarios remains a significant challenge. Instead, more practical approaches involve developing dictionaries tailored to specific imaging modalities or disease areas. Over the years, several radiology dictionaries have been compiled

for educational and reference use [149, 150]. While these resources offer value in teaching and general understanding, they tend to be overly simplistic for use in automated radiology report generation. One of their main limitations is that they are not derived directly from radiology reports and may differ in terminology from the language commonly used in clinical practice.

To overcome these limitations, the Radiological Society of North America (RSNA) introduced RadLex [151]—a comprehensive and structured radiology lexicon designed specifically for enhancing radiology reporting and supporting automated systems (See Figure 8). RadLex provides a standardized vocabulary that encompasses anatomical structures, imaging modalities, pathological conditions, and diagnostic findings, offering broad coverage for both manual and AI-assisted reporting workflows. Its structured format helps unify radiological language across systems and institutions, improving communication and interoperability. In its most recent release (version 4.2), RadLex includes a total of 46,838 curated terms.

The adoption of RadLex facilitates more accurate and efficient automated report generation by reducing the need for radiologists to manually define or select keywords. It supports AI-driven systems by supplying a standardized language foundation that aligns closely with clinical documentation practices, thus enhancing the quality and consistency of automatically generated radiology reports.



Figure 8. The RadLex Dictionary from RSNA. Various radiology dictionaries have been developed to standardize terminology and improve consistency in radiology reporting. Among them, the RadLex Dictionary from the Radiological Society of North America (RSNA) is widely used, offering both online and offline versions for flexible access. Other dictionaries have been introduced for educational and reference purposes, though they often lack direct extraction from radiology reports, making RadLex a preferred resource for structured reporting and automated processing.

1.3.2 Keyword Extraction from Text

Before exploring the role of keywords in radiology report generation, it is essential to review the development of keyword extraction in the medical domain. Keyword extraction serves as a crucial link between radiology reports and the complex, often unexplainable features generated by AI models. By isolating clinically relevant terms, it enables these representations to be transformed into more interpretable and visible outputs. (See Figure 9)

Keyword extraction refers to the process of identifying the most informative and representative terms within a text. This helps language models concentrate on the most critical content while simplifying the complexity of downstream processing tasks. As one of the earliest techniques in natural language processing (NLP), keyword extraction has played a foundational role in structuring and analyzing medical text. However, its application in healthcare settings is particularly challenging due to the presence of domain-specific

terminology, dense and variable sentence structures, and subtle clinical nuances that generic models often fail to capture accurately.

To address these challenges, researchers have developed both rule-based and model-driven approaches tailored to healthcare. For example, one method [152] constructs a medical dictionary and matches terms within reports to extract key clinical information. Another study [153] evaluated language models in extracting keywords from pathology reports within electronic health records, introducing refined techniques to improve extraction accuracy. These studies have demonstrated the feasibility of applying keyword extraction techniques to medical text while emphasizing the importance of incorporating domain knowledge into model design.

Extracting keywords from radiology reports, however, presents additional complexities. Unlike pathology reports, radiology reports often lack well-defined ground truth for keyword identification. Additionally, variations in writing styles across datasets introduce domain shift, making it difficult for models trained on one dataset to generalize effectively to another. As a result, designing robust keyword extraction systems for radiology remains an ongoing challenge, requiring methods that can not only identify clinically meaningful terms but also adapt effectively across diverse reporting environments.

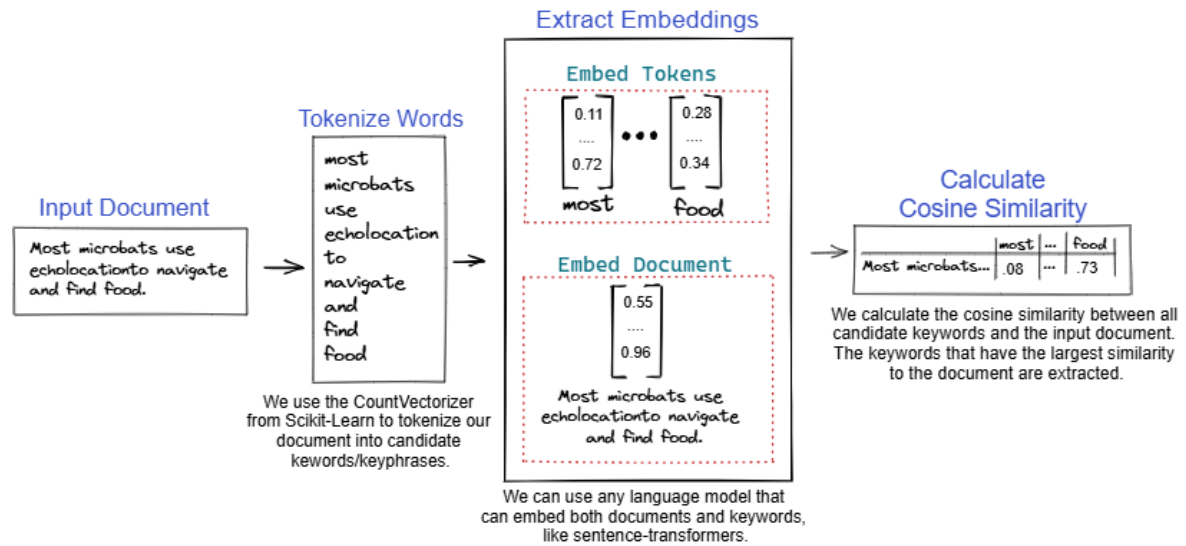


Figure 9. General pipeline for keyword extraction using a language model, illustrated with KeyBERT [154]. The keyword extraction process begins with an input texts, from which candidate keywords are identified using contextual embeddings generated by a pre-trained language model. KeyBERT, as an example, leverages BERT-based embeddings to rank and extract the most relevant keywords based on their semantic similarity to the input text.

1.4 Frequency-based Multi-Label Classification

1.4.1 Development of Classification in Deep Learning

Classification is a fundamental task in artificial intelligence (AI) and has significantly shaped the development of deep learning. Early approaches relied on traditional machine learning methods such as Logistic Regression and Support Vector Machines (SVMs). While effective in certain contexts, these models were limited by their reliance on manual feature engineering and their inability to directly handle raw image data, making them suboptimal for image classification tasks.

The advent of deep learning transformed classification by enabling neural networks to learn feature representations directly from data. A summary of classic classification networks is provided in Table 3. A pivotal moment was the introduction of AlexNet [61], which marked

a breakthrough in image classification by winning the ImageNet challenge. With 60 million parameters, AlexNet utilized five convolutional layers with max-pooling and three fully connected layers, concluding with a 1000-way softmax output. Dropout was employed in the fully connected layers to mitigate overfitting. This architecture laid the groundwork for subsequent convolutional neural networks (CNNs).

Following AlexNet, researchers sought to enhance network depth and connectivity. Key innovations included the Inception architecture [63, 65, 66, 70], which introduced multi-scale feature extraction within the same layer, and ResNet [64], which leveraged residual connections to address the vanishing gradient problem in deeper networks. As architectures advanced, attention mechanisms were integrated to improve feature discrimination. For instance, the Squeeze-and-Excitation (SE) block [71] applied channel-wise attention, while the Convolutional Block Attention Module (CBAM) [155] combined both spatial and channel-wise attention to strengthen the model's ability to focus on salient regions.

While convolution-based networks dominated image classification for years, more recently, a paradigm shift occurred with the adoption of the Transformer architecture [122] in computer vision. The Vision Transformer (ViT) [80] adapted Transformers by partitioning images into fixed-size patches, enabling the self-attention mechanism to model long-range dependencies and global image context. Unlike CNNs, which process information locally, ViTs treat the entire image holistically, providing a more comprehensive representation. Today, CNNs and Transformer-based models represent two major frameworks in image classification, each with distinct advantages.

Despite their strong performance, traditional classification networks often rely on task-specific training, limiting their adaptability to diverse tasks. To overcome this limitation, researchers introduced models capable of general-purpose classification without task-specific

training. A key milestone was CLIP [92], which combines vision and language understanding using a transformer-based architecture. Trained on 400 million image-text pairs from the internet, CLIP enables zero-shot transfer, allowing it to classify images without prior fine-tuning on specific tasks. This approach mirrors the pretraining paradigm in NLP, transforming classification from training models from scratch to fine-tuning pretrained models with domain-relevant data. The zero-shot capability of CLIP has significantly expanded the versatility and efficiency of classification networks, enabling applications across diverse domains, including medical imaging.

Table 3. Overview of classic and state-of-the-art classification networks. The networks are systematically collected and reorganized from the source of MMPretrained Contributors (2023) [156]. The table includes key architectures that have significantly influenced the field of image classification, ranging from foundational models to advanced deep learning frameworks. Each network is categorized based on its architectural characteristics, such as convolutional layers, residual connections, or attention mechanisms, and their contributions to improving classification accuracy and computational efficiency are highlighted. The table serves as a comprehensive reference for understanding the evolution and performance of classification networks in both general and domain-specific applications.

Author et al	Work	Year	Network Structure	CNN? Transformer?	Spotlight Mechanism?
Krizhevsky et al	ImageNet Classification with Deep Convolutional Neural Networks	2012	AlexNet	CNN	Pioneer of Deep Learning based convolutional network, Also proposed the ImageNet Classification
Simonyan et al	Very Deep Convolutional Networks for Large Scale Image Recognition	2014	VGG	CNN	Pioneer of CNN, 11 19 weight layer
Segeedy et al	Going deeper with convolutions	2014	Inception	CNN	while keeping the computational budget constant
He et al	Deep Residual Learning for Image Recognition	2015	ResNet	CNN	Residual Block
Ioffe et al	Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift	2015	Inception V2	CNN	Proposal of Batch Normalization
Segeedy et al	Rethinking the Inception Architecture for Computer Vision	2015	Inception V3	CNN	factorized convolutions and aggressive regularization
Xie et al	Aggregated Residual Transformations for Deep Neural Networks	2016	ResNeXt	CNN	ResNeXt Block (Wider)
Zagoruyko et al	Wide Residual Networks	2016	Wide ResNet	CNN	Wide Dropout
Huang et al	Densely Connected Convolutional Networks	2016	DenseNet	CNN	Dense Connection
Segeedy et al	Inception v4, Inception ResNet and the Impact of Residual Connections on Learning	2016	Inception ResNet	CNN	combining the Inception architecture with residual connections
Hu et al	Squeeze and Excitation Networks	2017	SE ResNet	CNN	Squeeze and Excitation Block
Zhang et al	ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices	2017	ShuffleNet V1	CNN	Channel Shuffle
Sandler et al	MobileNetV2: Inverted Residuals and Linear Bottlenecks	2018	MobileNetV2	CNN	Linear Bottleneck and Inverted Residuals
Ma et al	ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design	2018	ShuffleNet V2	CNN	Channel Split for replace Channel Shuffle
Howard et al	Searching for MobileNetV3	2019	MobileNetV3	CNN	SE channel
Gao et al	Res2Net: A New Multi scale Backbone Architecture	2019	Res2Net	CNN	Res2Net block (Multi Scale)
Tan et al	EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks	2019	EfficientNet	CNN	neural architecture search, Scale of Network
Wang et al	Deep High Resolution Representation Learning for Visual Recognition	2019	HRNet	CNN	high to low in Parallel, exchange the information across resolutions
Wang et al	CSPNet: A New Backbone that can Enhance Learning Capability of CNN	2019	CSPNet	CNN	Cross Stage Partial, Partial Transition
Dosovitskiy et al	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale	2020	Vision Transformer	Transformer	Pioneer of Transformer in computer vision, split image into patches
Radosavovic et al	Designing Network Design Spaces	2020	RegNet	CNN	neural architecture search in Space
Touvron et al	Training data efficient image transformers and distillation through attention	2020	DeiT	Transformer	Student Teacher strategy, Distillation Token
Liu et al	Swin Transformer: Hierarchical Vision Transformer using Shifted Windows	2021	Swin Transformer	Transformer	Shift Windows, Swin Transformer Block
Ding et al	RepVGG: Making VGG-style ConvNets Great Again	2021	RepVGG	CNN	VGG Style, revise structure
Yuan et al	Tokens to Token ViT: Training Vision Transformers from Scratch on ImageNet	2021	Tokens to Token ViT	Transformer	Token to token process
Peng et al	Conformer: Local Features Coupling Global Representations for Visual Recognition	2021	Conformer	CNN, Transformer	Feature Coupling Unit
Chu et al	Twins: Revisiting the Design of Spatial Attention in Vision Transformers	2021	Twins	Transformer	spatial attention mechanism
Yu et al	MetaFormer: Is Actually What You Need for Vision	2021	PoofFormer	Transformer	Pooling replace MLP in Attention
Li et al	MViTv2: Improved Multiscale Vision Transformers for Classification and Detection	2021	MViTv2	Transformer	Residual Pooling Connection
Liu et al	Swin Transformer V2: Scaling Up Capacity and Resolution	2021	Swin Transformer V2	Transformer	post normalization, scaled cosine attention, log spaced continuous position bias
Bao et al	BEiT: BERT Pre-Training of Image Transformers	2021	BEiT	Transformer	BERT Pre-Training
Radford et al	Learning Transferable Visual Models From Natural Language Supervision	2021	CLIP	Transformer	Large Scale Pretrained, Unsupervised and wide task range
Tan et al	EfficientNetV2: Smaller Models and Faster Training	2021	EfficientNet V2	CNN	training aware neural architecture search, Fused MBConv
ElNouby et al	XGCT: Cross Covariance Image Transformers	2021	XGCT	Transformer	cross covariance attention
Chen et al	An Empirical Study of Training Self-Supervised Vision Transformers	2021	MAE V3	Transformer	instability discussion
He et al	Masked Autoencoders Are Scalable Vision Learners	2021	MAE (Mask autoencoders)	Transformer	asymmetric encoder decoder, masking a high proportion of the input image
Wu et al	TinyViT: Fast Pretraining Distillation for Small Vision Transformers	2022	TinyViT	Transformer	Student Teacher Model in Patch Swap
Liu et al	A ConvNet for the 2020s	2022	ConvNeXt	CNN	pyramid structure
Guo et al	Visual Attention Network	2022	Visual Attention Network	CNN	large kernel attention
Trockman et al	Patches Are All You Need?	2022	ConvMixer	CNN	ConvMixer Layer
Maz et al	EdgeNeXt: Efficiently Amalgamated CNN Transformer Architecture for Mobile Vision Applications	2022	EdgeNeXt	CNN, Transformer	split depth wise transpose attention
AnasaluVasu et al	An Improved One millisecond Mobile Backbone	2022	MobileOne	CNN	Depthwise convolution and Pointwise convolution, Remove Residual structure
Li et al	EfficientFormer: Vision Transformers at MobileNet Speed	2022	EfficientFormer	Transformer	dimension consistent pure transformer, latency driven slimming
Peng et al	BEiT v2: Masked Image Modeling with Vector Quantized Visual Tokenizers	2022	BEiT V2	Transformer	Masked image modeling, vector quantized knowledge distillation
Li et al	BLIP: Bootstrapping Language Image Pre-training for Unified Vision Language Understanding and Generation	2022	BLIP	Transformer	noisy web data by bootstrapping
Woo et al	Co-designing and Scaling ConvNets with Masked Autoencoders	2023	ConvNeXt V2	CNN	Global Response Normalization, Remove Layerscale
li et al	BLIP 2: Bootstrapping Language Image Pre-training with Frozen Image Encoders and Large Language Models	2023	BLIP V2	Transformer	lightweight Querying Transformer

1.4.2 Multi-Label Classification

Multi-label classification is a task where each instance may belong to multiple categories simultaneously, making it particularly relevant in medical imaging, where a single radiological image often presents multiple findings or pathologies. Different with the wider range of application for multi-class application, multi-label classification would provide more information and give more possible answers, which Figure 10 shows the main difference. In natural image processing, multi-label classification has been extensively studied, with several

approaches emerging even before the rise of deep learning. Early methods framed multi-label classification as a mathematical problem using classifier chains, where each label was treated as an independent binary problem. This pioneering work demonstrated that by incorporating label correlations, classifier chains could achieve more effective multi-label classification compared to binary relevance approaches ([157]).

The introduction of deep learning—particularly Convolutional Neural Networks (CNNs)—significantly advanced multi-label classification by enabling models to learn complex feature representations directly from raw image data. A variety of strategies have since emerged to improve model performance, including the development of specialized network architectures ([158]; [159]; [159]; [160]),, attention mechanisms ([161]; [162]; [163]), custom loss functions ([164]; [165]), and tailored training methodologies [166]. These innovations have enhanced model accuracy, especially when dealing with complex and imbalanced datasets, which are common in multi-label classification tasks.

In contrast to natural image datasets like MS-COCO, PASCAL VOC, and ImageNet, medical imaging datasets are typically smaller and suffer from highly imbalanced label distributions. This imbalance often causes models to favor frequently occurring labels—typically normal findings—while underperforming on rarer, yet clinically significant, abnormalities. To address this issue, researchers have adapted techniques from natural image classification, including label co-occurrence modeling [167] (as illustrated in Figure 11), attention-based enhancements [168], and custom loss functions [169], all designed to improve model sensitivity to underrepresented classes.

Despite these advancements, applying multi-label classification as a tool for radiology keyword prediction poses unique challenges. In natural image settings, the label set is relatively fixed and consistent, but in radiology, the keyword vocabulary is highly dynamic and context-

dependent. Label distributions can vary significantly across clinical environments, imaging modalities, and datasets, which violates the core assumption of stable label sets in traditional multi-label frameworks. Therefore, while multi-label classification shows strong potential for medical image analysis, its integration into radiology report generation requires further refinement to accommodate the complex, evolving nature of clinical language and reporting styles.

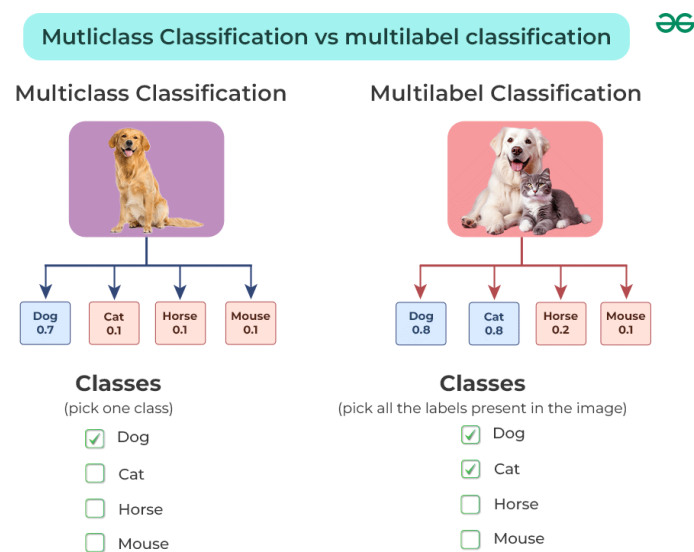


Figure 10. Illustration of the difference between multi-label classification and multi-class classification (in single-label classification). Adapted from HuggingFace’s beginner tutorial [170]. This figure highlights the key distinctions between the two classification paradigms. In multi-class classification, each input is assigned a single label from a set of mutually exclusive categories, whereas in multi-label classification, an input can be associated with multiple labels simultaneously. The figure visually demonstrates these concepts using example scenarios, emphasizing how multi-label classification accommodates overlapping or concurrent labels, while multi-class classification enforces exclusivity. This distinction is critical for understanding the appropriate application of each approach

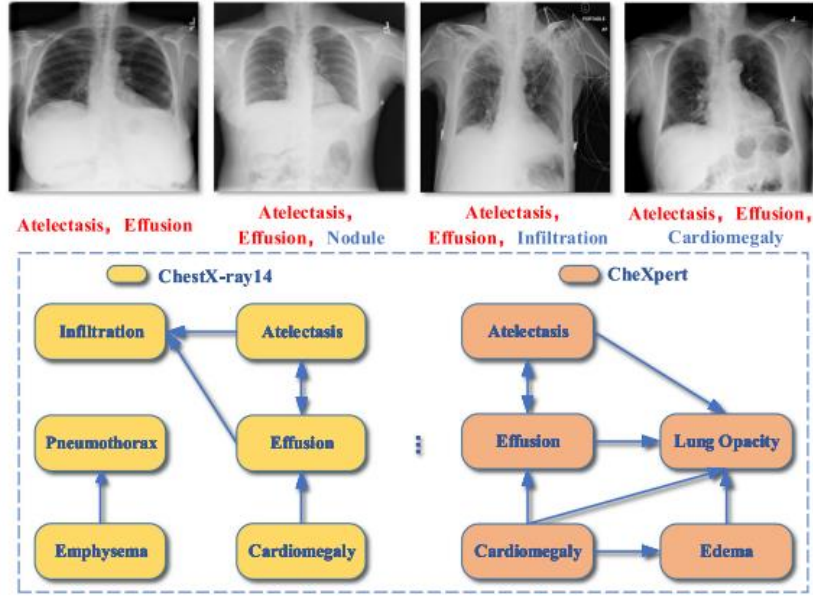


Figure 11. Example of multi-label classification applied to medical imaging. This sample, captured from [167] demonstrates the process of performing multi-label classification on Chest X-Ray images. The figure illustrates how multiple diagnostic labels, such as pneumothorax, effusion, or cardiomegaly, can be assigned to a single image, reflecting the presence of concurrent or overlapping conditions. This example highlights the utility of multi-label classification in medical imaging, where complex and co-occurring pathologies often require simultaneous identification for accurate diagnosis and treatment planning.

1.5 Text-to-text Large Language Model

1.5.1 Brief Introduction of Language Model

In machine learning, a language model enables machines to read, write, and understand human language, thereby simulating human-like communication abilities. Unlike humans, who possess an innate capacity for language comprehension and generation [1], machines rely on artificial intelligence (AI) and natural language processing (NLP) techniques to interpret and produce linguistic data. Language models are specifically designed to estimate the generative likelihood of word sequences and predict the probabilities of language tokens within a given context. This capability plays a pivotal role in the advancement of AI and has drawn significant

attention across a broad range of domains. The historical progression of language model development is illustrated in Figure 12.

The foundational use of neural networks in language modeling began with Recurrent Neural Networks (RNNs) [2], which modeled the probability of word sequences while capturing temporal dependencies. A major advancement followed with the introduction of distributed word representations [3], allowing words to be embedded as vectors that encode semantic relationships. Building on this, Word2Vec [4] enabled efficient learning of these word embeddings from large corpora, further enhancing the model’s ability to capture contextual and semantic patterns in language.

As deep learning evolved, Long Short-Term Memory (LSTM) networks emerged as a key advancement, overcoming the limitations of traditional RNNs by addressing issues of long-term dependencies. The introduction of Bidirectional LSTMs (biLSTMs) [171] further enhanced language modeling by capturing context-aware word representations from both preceding and succeeding words within a sequence. However, the true revolution in NLP came with the development of the Transformer model [122], which replaced recurrent structures with a self-attention mechanism. This architecture enabled the modeling of global word dependencies in parallel, greatly improving both the efficiency and expressiveness of language models.

Building upon the Transformer architecture, BERT (Bidirectional Encoder Representations from Transformers) [127] introduced pre-training techniques that reduced training time and allowed the model to predict general-purpose semantic features. By pre-training on large-scale text corpora and fine-tuning for specific tasks, BERT set a new standard for NLP, achieving state-of-the-art performance across various benchmarks. This paradigm of “pre-

training” and “fine-tuning” became a cornerstone of modern NLP, enabling researchers to adapt language models to diverse applications with minimal task-specific data.

As computational resources and available data continued to expand, language models scaled into what are now referred to as Large Language Models (LLMs). These models, trained on massive datasets with billions of parameters, exhibit advanced capabilities in language understanding and generation. Due to the increasing cost and complexity of training such models from scratch, research has shifted toward leveraging and fine-tuning pre-trained LLMs. This shift has accelerated progress in applications such as automated text generation, machine translation, summarization, and conversational agents, solidifying LLMs as a core component of contemporary AI systems.

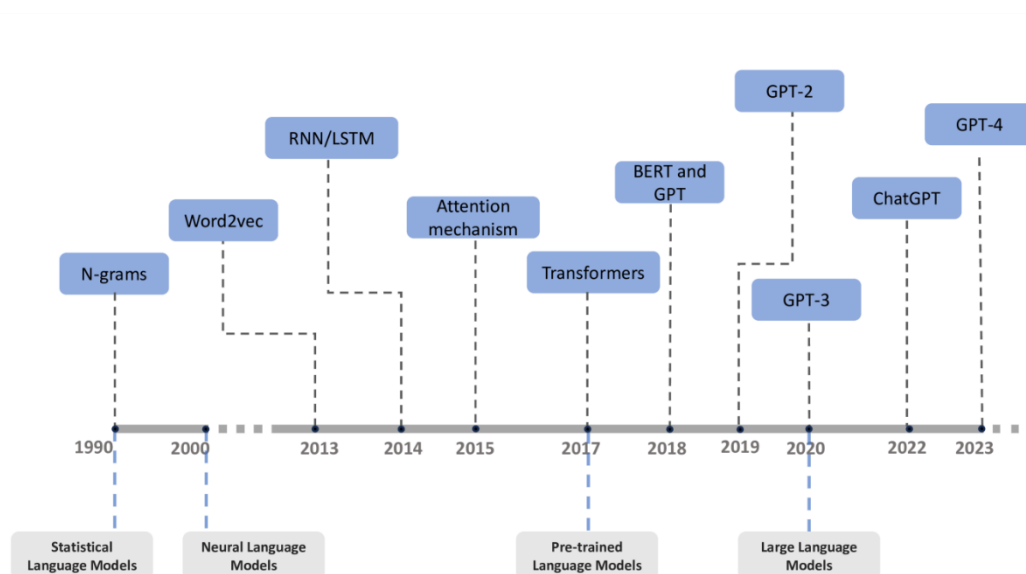


Figure 12. A brief timeline illustrating the development of language models. The figure highlights key milestones in the evolution of language models, starting from traditional statistical approaches to the emergence of deep learning-based models. It is noted that after the development of RNNs (Recurrent Neural Networks) and LSTMs (Long Short-Term Memory networks), language models transitioned to deep learning-based architectures. Following the introduction of Transformers, language models began to scale significantly in size and complexity, leading to the era of Large Language Models (LLMs). This timeline, adapted from [172], showcases the progression of language models and their increasing capabilities in natural language understanding and generation tasks.

1.5.2 Large Language Model

As language models continue to evolve, a clear trend emerges: increasing model complexity and larger datasets lead to stronger language understanding and generation capabilities, enabling models to tackle increasingly complex real-world tasks. To support a broader range of applications, both the structural sophistication of language models and the scale of their training materials have grown significantly. This evolution has given rise to a new category of models known as Large Language Models (LLMs).

Unlike earlier architectures such as pure Transformers or LSTMs, which were typically trained on annotated datasets for task-specific purposes, LLMs are designed to generalize across a wide range of applications without requiring extensive task-specific annotations. A key characteristic of LLMs is their reliance on large-scale pretraining using vast collections of unannotated text. Through self-supervised learning, LLMs uncover linguistic patterns and contextual structures from heterogeneous datasets, allowing them to generate coherent, context-aware outputs with minimal downstream fine-tuning. This approach addresses the impracticality of anticipating and annotating every possible task, especially in open-ended or dynamic domains.

The introduction of the Transformer architecture was instrumental in shaping this evolution, shifting the paradigm toward pretraining and general-purpose language modeling. A pivotal milestone in this direction was the development of OpenAI’s Generative Pre-trained Transformer (GPT) [126], which introduced a two-stage training process: first, pretraining a generative Transformer on diverse unlabeled text; second, fine-tuning the model for specific tasks. This framework significantly reduced reliance on annotated data and facilitated broader task transferability.

Building on GPT’s success, subsequent iterations—GPT-2, GPT-3, and GPT-4—scaled model size, training data diversity, and contextual reasoning capabilities, further extending the utility of LLMs. Other models such as BART, T5, and LLaMA have introduced architectural innovations and training strategies to improve general language tasks and enable zero-shot and few-shot learning. Collectively, these models demonstrate the versatility of LLMs in tasks including text summarization, translation, content generation, and question answering—often with minimal or no additional task-specific supervision. A timeline of large language model development is shown in Figure 13.

Despite their remarkable performance, LLMs also introduce notable challenges. Their broad applicability raises ethical and social concerns related to misinformation, content bias, and data privacy, necessitating transparent governance and responsible deployment. Moreover, in specialized fields such as medicine, general-purpose LLMs often struggle to meet domain-specific demands. Since they are predominantly trained on open-domain corpora, these models may lack the precision and terminology required for clinical tasks. To enhance their reliability in such contexts, domain adaptation strategies—such as fine-tuning on curated, field-specific datasets—are essential. These adaptations ensure that LLMs meet the accuracy, safety, and interpretability requirements of high-stakes applications like medical diagnostics and decision support.

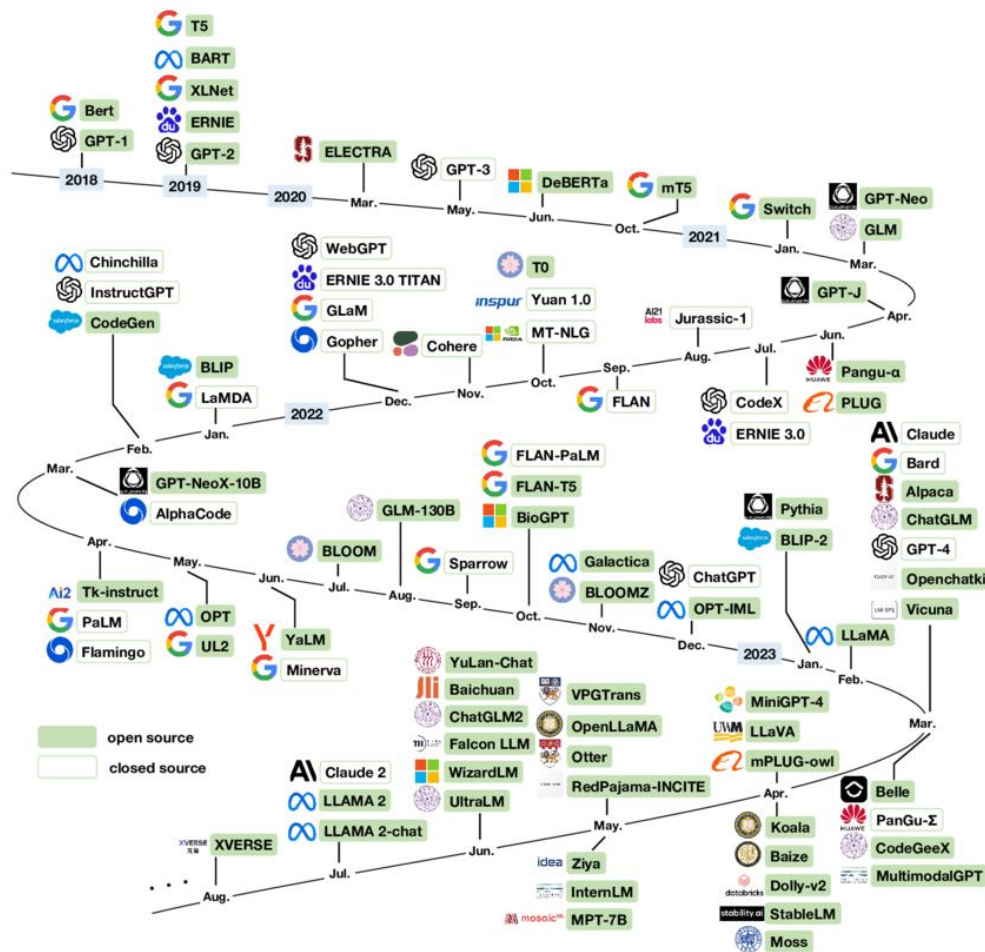


Figure 13. A brief timeline depicting the development of large language models (LLMs). The figure highlights the transformative impact of the Transformer architecture, introduced in 2017, which marked the emergence of the LLM. Following the development of foundational models such as GPT-1 and BERT, the technology underlying LLMs stabilized, enabling widespread adoption and further advancements. Subsequent progress has focused on refining training materials, optimizing training methods, and enhancing user interfaces (GUIs) to improve accessibility and usability. This timeline, adapted from [173], illustrates the evolution of LLMs and their growing influence on natural language processing and AI-driven applications.

2. Research Aims and Objectives

2.1 *Research Gaps*

Radiology report drafting, whether performed manually or through automation, has distinct strengths and limitations. Manual drafting by radiologists remains the most accurate approach, especially for abnormal cases, as it relies on expert interpretation and nuanced clinical judgment. However, this process is time-consuming, particularly for routine or normal cases, which constitute the majority of radiological imaging. This inefficiency impacts diagnostic workflows and prolongs turnaround times, creating bottlenecks in clinical operations.

To address these challenges, various deep learning-based methods for automatic radiology report generation have been developed. These methods typically use an image encoder to extract high-level features from medical images and a text decoder to generate reports. While they reduce radiologists' workload, several limitations hinder their clinical adoption.

A primary limitation is the low quality of generated reports, which often lack the precision and coherence required for reliable clinical use—especially in complex cases or when subtle findings are present. Additionally, these methods lack explainability: the high-level features extracted by image encoders are abstract and uninterpretable, preventing radiologists from verifying whether the features capture essential clinical information. As a result, radiologists can only assess the final report, and any errors require manual revisions. While this may be manageable for normal cases with repetitive text patterns, the increasing case volume reduces the efficiency gains of automation. Moreover, the inability to trace the evidence behind generated reports often

leads radiologists to prefer manual drafting for abnormal cases, limiting the practical application of automated methods.

To address these challenges, this research proposes a novel deep learning-based framework that replaces unexplainable high-level features with visible and interpretable keywords. The framework uses multi-label classification to generate a list of relevant keywords from medical images, which are then used by a text-to-text large language model to generate radiology reports. To overcome the lack of ground-truth keyword annotations, the framework introduces an automatic keyword adaptation mechanism that derives relevant keywords from sample radiology reports. Additionally, a frequency-based multi-label classification method is tailored to these adapted keyword settings. By combining these innovations, the proposed framework aims to improve the quality, accuracy, and interpretability of generated reports, reducing radiologists' workload while enhancing trust and efficiency in automated systems for clinical practice.

2.2 *Research Aims*

This research aims to overcome the limitations of existing radiology report generation methods that rely on general encoder-decoder structures. To achieve this, a novel deep learning framework is developed, emphasizing both explainability and report quality. The framework integrates three key components: (1) an automatic keyword adaptation mechanism, (2) frequency-based multi-label classification for generating keyword lists, and (3) a text-to-text large language model for producing high-quality, interpretable radiology reports. By combining these components, the framework ensures that generated reports are accurate, transparent, and easily verifiable. Additionally, it facilitates seamless collaboration with radiologists, enhancing the efficiency and practicality of automated workflows in real-world clinical settings.

2.3 *Research Objectives*

1. Design and develop a deep learning-based framework for the automatic generation of chest X-ray radiology reports.
2. Implement automatic keyword adaptation and frequency-based multi-label classification to address the lack of ground-truth keywords in radiology reports, and integrate a text-to-text large language model to generate high-quality, explainable reports from these keyword lists.
3. Evaluate the framework's performance by validating its keyword adaptation, multi-label classification, and text-to-text large language model components, and compare its report generation capabilities with state-of-the-art methods.

Compared to existing approaches, the proposed framework aims to deliver more accurate and interpretable radiology reports. By replacing unexplainable high-level features with visible, understandable keyword lists, radiologists can validate and adjust keywords before report generation, improving flexibility and reducing the time required for revisions and verification. Ultimately, this approach seeks to enhance diagnostic and treatment workflows for lung-related diseases by reducing turnaround times and supporting further applications such as AI-assisted patient communication and seamless integration across hospital departments.

3. Materials and Methods

3.1 Dataset Collection

3.1.1 IU X-Ray Dataset

The IU X-Ray dataset [124], developed by Indiana University, is a widely used benchmark for radiology report generation. It contains 7,566 chest X-ray images paired with 3,852 radiology reports in its original version. Following the official dataset split from [8], we divided the data into training, validation, and testing sets. We excluded images without associated radiology reports and assumed that all images (in different views) within a single case share the same report. After filtering and preprocessing, the final dataset comprises 6,659 image-report pairs in the training set, 295 pairs in the validation set, and 590 pairs in the testing set. The sample radiology report of IU X-Ray dataset is shown in Figure 14.

<p>RADIOLOGY REPORT</p> <p>DATE: XXXX, XXXX XXXX XXXX hours</p> <p>Indication: Abdominal pain and distention.</p> <p>Findings: Frontal and lateral views of the chest show an unchanged cardio mediastinal silhouette. There is bibasilar interstitial opacity and left basal plate like opacity XXXX due to discoid atelectasis and/or XXXX scarring. There are emphysematous changes, particularly within the right upper lobe. No XXXX focal airspace consolidation or pleural effusion.</p> <p>Impression: 1. COPD. Basilar probable pulmonary fibrosis and scarring. 2. No acute cardiac or pulmonary disease process identified.</p> <p>DICTATED BY : Dr. XXXX XXXX XXXX XXXX XXXX ELECTRONICALLY SIGNED XXXX. XXXX XXXX XXXX XXXX XXXX TRANSCRIBED XXXX 8 XXXX XXXX RADRES XXXX</p> <p>SIGNATURE XXXX</p>
--

Manual annotation

- Opacity/lung/base/bilateral/interstitial
- Pulmonary Atelectasis/base/left
- Cicatrix/lung/base/left
- Pulmonary Emphysema
- Pulmonary Disease, Chronic Obstructive
- Pulmonary Fibrosis/base

MTI annotation

- Cicatrix
- Pulmonary Fibrosis
- Pulmonary Atelectasis
- Lung
- Pleural Effusion
- Pulmonary Disease, Chronic Obstructive

Figure 14. Example of a radiology report from the IU X-Ray dataset. The figure showcases a typical radiology report, which includes a detailed textual description of findings derived from medical imaging. The report is structured to provide clinical observations, interpretations, and recommendations, highlighting key aspects such as anatomical structures, abnormalities, and potential diagnoses. Please note that, although the original version of IU X-Ray dataset provide the annotations in Manual/MTI, these annotations are for the MeSH[174] codes supplemented by Radiology Lexicon (RadLex) codes, which is not actually matched the keywords appeared in Finding of radiology report. Therefore, we do not utilize it for radiology report generation.

3.1.2 MIMIC-CXR Dataset

The MIMIC-CXR dataset [125] is the largest publicly available collection of chest radiographs with free-text radiology reports. Its latest version, released on July 23, 2024, includes 377,110 images corresponding to 227,835 radiographic studies. We followed the official dataset split provided by PhysioNet and excluded images without corresponding radiology reports. The final dataset consists of 270,790 image-report pairs in the training set, 2,130 pairs in the validation set, and 3,858 pairs in the testing set. The sample radiology report is shown in Figure 15.

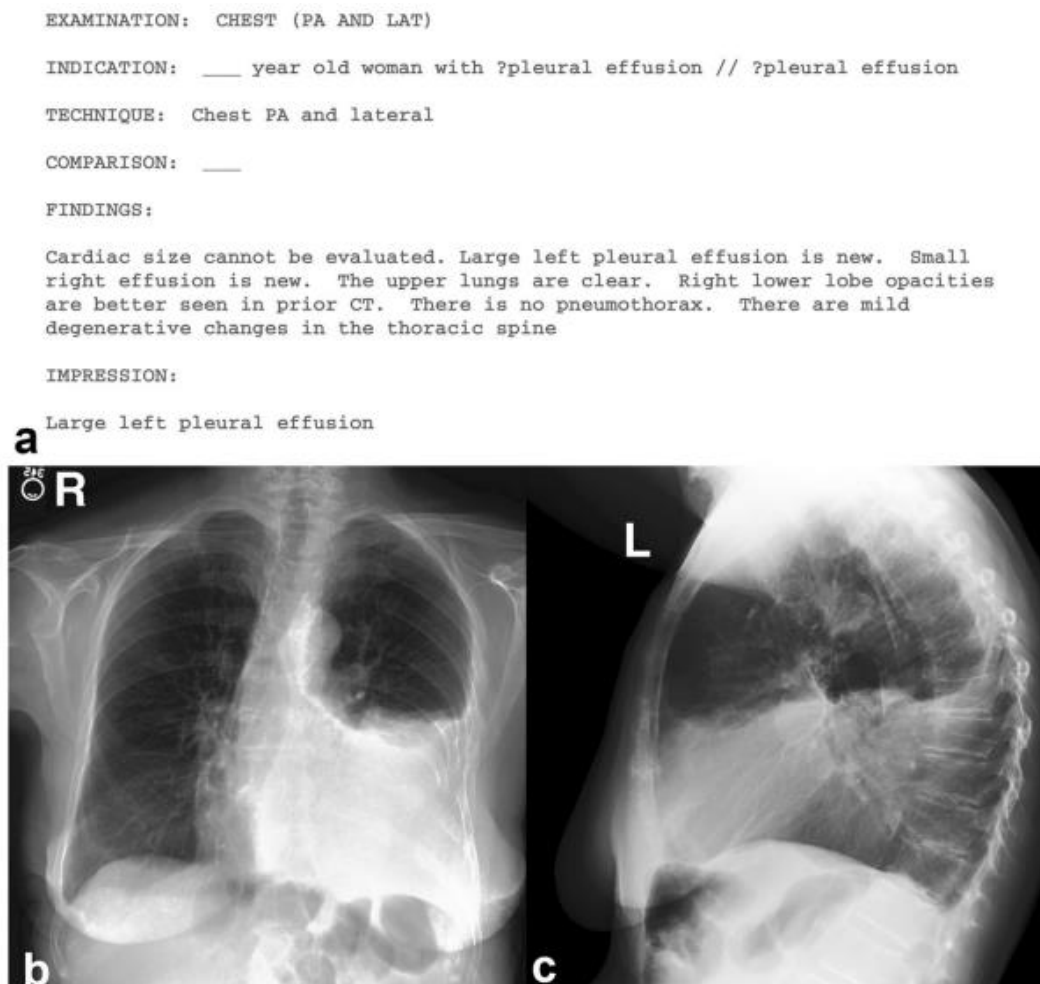


Figure 15. Example of a radiology report from the MIMIC-CXR dataset. The figure presents a representative radiology report, which includes a comprehensive textual description of imaging findings, clinical observations, and diagnostic interpretations.

3.2 *Radiology Report Generation Framework based on Automatic Keyword Adaption, Frequency-based Multi-Label Classification and Text-to-text Large Language Model*

3.2.1 Overview of whole framework

Given a set of radiology X-ray images, denoted as Img_input , the objective is to generate a detailed sequence $W=w_1, w_2, \dots, w_n$ that describes both normal and abnormal findings present in the images. Traditional deep learning models typically generate radiology reports by directly predicting subsequences of text from the images. However, such models often lack transparency and explainability. Recognizing that radiology reports are structured around keywords, we propose a novel approach where keywords are first extracted from the images and subsequently used to generate the final report using a large language model (LLM).

To extract the relevant keywords from the images, our method employs a multi-label classification approach. However, several challenges arise with this method, including the need to classify new medical scenarios and the absence of a pre-existing reference keyword list for specific medical conditions. To address these challenges, we introduce automatic keyword adaptation, which dynamically adjusts the multi-label classification process based on provided radiology reports. This adaptation mechanism tailors the keyword generation to particular medical contexts by leveraging the available reports, ensuring flexibility and relevance for diverse clinical environments.

To further improve the effectiveness of our method, we address the issue of class imbalance—a common problem in medical multi-label classification. We propose a frequency-based classification strategy, in which keywords are grouped into categories

based on their frequency of occurrence. These categories represent varying levels of classification difficulty, from rare to common keywords. For each frequency group, we train a separate neural network to generate the corresponding keyword list. These separate lists are then combined through a process called keyword list fusion, resulting in a comprehensive, balanced keyword list that captures both frequent and rare terms.

Finally, the generated keyword list is used as input to a pre-trained large language model to generate the radiology report. The language model, fine-tuned on medical texts, ensures the report is coherent, contextually relevant, and written in a human-like style. By generating the report from a well-structured keyword list, our approach not only improves the interpretability of the generated report but also ensures its accuracy and clinical relevance.

This workflow, summarized in Figure 16, enables our method to adapt to various medical scenarios with minimal reliance on predefined keyword lists, enhancing its generalizability and robustness across different datasets and clinical conditions.

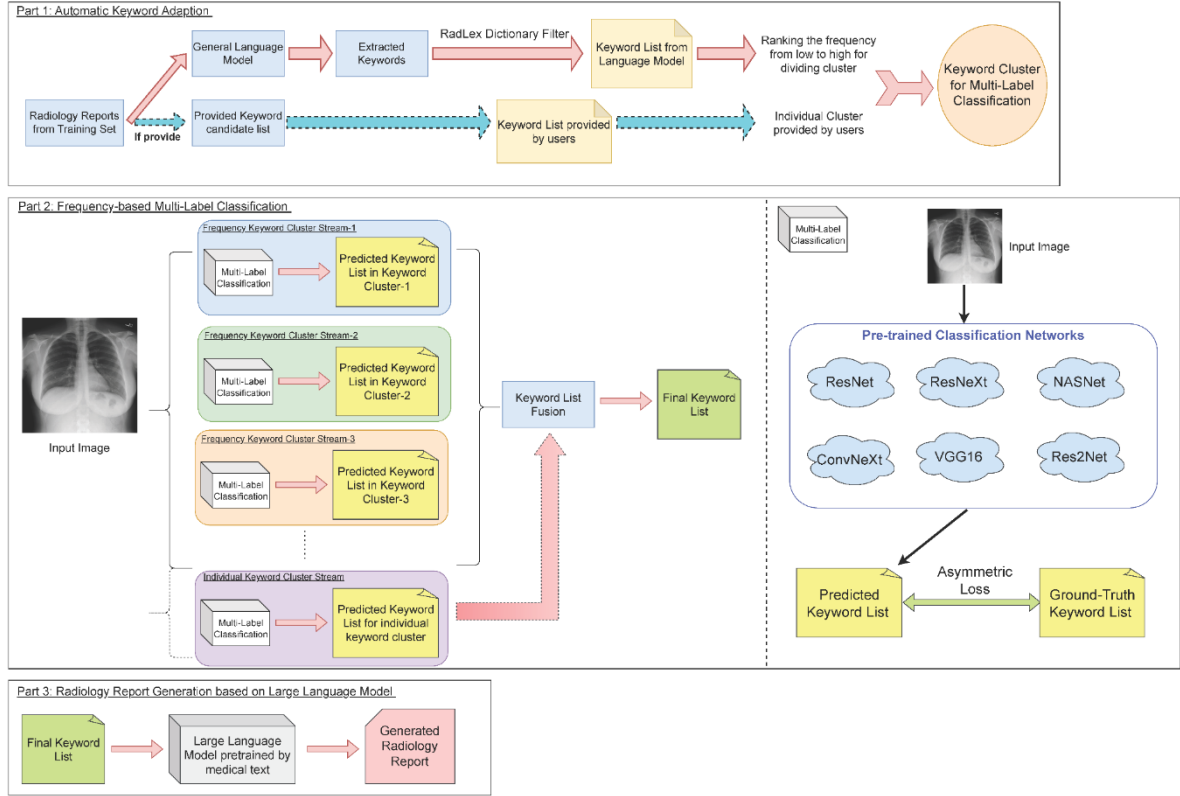


Figure 16. Overview of the proposed radiology report generation pipeline integrating automatic keyword adaptation and frequency-based multi-label classification. The process begins with automatic keyword adaptation, which processes radiology reports from the training set to extract keyword clusters. These clusters are then used to configure the frequency-based multi-label classification. Subsequently, the frequency-based multi-label classification predicts keyword lists for each cluster, which are combined through keyword list fusion to generate the final keyword list. Finally, a large text to text language model generates the corresponding radiology report using the fused keyword list.

3.2.2 Automatic Keyword Adaption

A key challenge in generating radiology reports from keywords is the absence of ground-truth or reference keyword lists for each case. To address this, we simplify the problem by assuming that radiology reports from specific medical contexts are available as references. This assumption enables us to focus on extracting relevant keywords from these reports without the need for prior knowledge or predefined keyword lists.

In situations where no reference keyword list is available, we rely on keywords extracted by the language model, which are further filtered using a radiology-specific dictionary. By leveraging advancements in language models and keyword extraction techniques, we can extract keywords directly from radiology reports using pre-trained models, such as those trained on general and medical text corpora. This allows the identification of relevant terms that reflect the information contained in the reports.

For keyword extraction, we utilize the KeyBERT model ([53], [154]) to generate an initial set of keywords from the provided radiology reports. This process does not include any filtering, as the model is designed to extract keywords without specific domain constraints. However, it is important to note that these keywords may not always meet the structural requirements of a radiology report, as the language model is not explicitly trained for medical report generation.

To refine the extracted keywords and ensure they align with the needs of multi-label classification, we perform a post-processing step using a radiology-specific dictionary, which is RadLex ([54]) in our pipeline. This filtering step ensures that only relevant and widely used radiology terms are retained, while preventing the omission of critical keywords. After the keywords are extracted and filtered, they are ranked by frequency and grouped into clusters for further multi-label classification.

Our proposed automatic keyword adaptation mechanism allows for flexibility across various medical contexts, eliminating the need for predefined reference keyword lists. This adaptability is key to ensuring that the model can generate meaningful and contextually relevant reports, even when reference lists are absent or incomplete.

In cases where a reference keyword list is provided, we prioritize these user-provided keywords to align with specific requirements. The provided list is cross-referenced with

the radiology dictionary RadLex to verify the validity of the terms. If any keyword is missing from the dictionary, the system alerts the user and offers the option to retain or discard the term. This process helps maintain the robustness and accuracy of the model, even in cases where the reference keyword list may contain errors or inconsistencies.

In practice, the most common scenario involves partial or incomplete reference keyword lists, which may not cover all the necessary terms for generating high-quality radiology reports. To address this, we adopt a parallel approach, merging the keywords generated by the language model with the user-provided list. This combined keyword set ensures that the generated reports are accurate, comprehensive, and relevant to the specific clinical context.

A detailed block diagram of the proposed automatic keyword adaptation process is shown in Figure 17. Through these strategies, our method ensures that the model can effectively generate radiology reports across a wide range of medical scenarios, while maintaining high quality and relevance. The visualization of the workflow for Automatic Keyword Adaption process connected with Multi-Label Classification is shown in Figure 20.

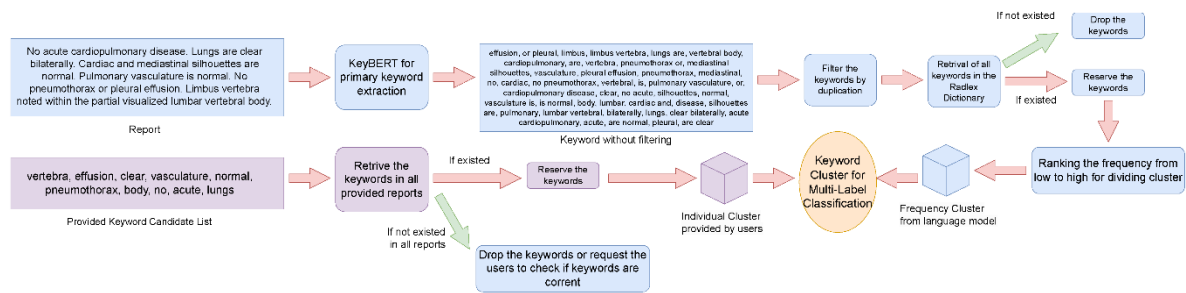


Figure 17. Block diagram of the proposed automatic keyword adaptation process. Beginning with the provided radiology reports, candidate keywords are extracted using the KeyBERT tool. Subsequently, the extracted keywords are filtered using the RadLex Dictionary, retaining only those present in the dictionary. The filtered keywords are ranked by frequency, and keyword clusters are constructed based on this ranking for use in frequency-based multi-label classification. If users provide a keyword candidate list, these user-specified keywords are given the highest priority and processed through a separate branch. After validation to ensure

the keywords exist within the reports, the user-provided keywords are treated as an individual cluster and integrated into the keyword clusters for multi-label classification.

3.2.3 Frequency-based Multi-Label Classification

Medical imaging presents unique challenges for multi-label classification, particularly due to the imbalance between common and rare conditions in radiology images.

Common findings are much more frequent than rare ones, which often leads to models focusing predominantly on predicting common findings while neglecting less frequent, yet clinically significant, conditions. This issue becomes even more pronounced in our dynamic keyword adaptation process, where keywords are generated based on provided radiology reports from specific medical contexts. To address this, we propose a frequency-based multi-label classification approach that divides keywords into different frequency groups to enhance classification accuracy and balance.

Frequency Categorization. In typical image classification tasks, such as natural image recognition, label distribution is often balanced. However, in medical imaging, there is a significant imbalance, with certain conditions appearing much more frequently than others. This imbalance can lead to biased predictions, where rare conditions are overlooked simply because they are less frequently observed.

To mitigate this, we categorize keywords into frequency groups based on how often they appear in the dataset, which is shown in Figure 18. The frequency categorization is dynamic, allowing it to adjust according to the current dataset or be manually set by the user. By grouping keywords into different frequency categories, we can better tailor the classification process to each cluster, improving the model’s ability to detect both common and rare

conditions. This approach ensures that keywords related to rare conditions are given sufficient attention, preventing their underrepresentation in the final reports.

Multi-Label Classification and Keyword List Fusion. Once the keywords are divided into frequency groups, the multi-label classification process is applied within each group, which is shown in Figure 20. While the frequency of the keywords determines the categorization, to maintain consistency and performance, we initially use the same network structure across all frequency groups. This uniform approach allows us to optimize the classification performance without presupposing the ideal network settings for each specific medical scenario. However, if users have specific classification performance requirements, they can adjust the network settings for each group in our framework after the initial setup.

After the classification step, the outputs from each frequency group are combined into a comprehensive keyword list through a process we call Keyword List Fusion, shown in Figure 19. This process integrates the classified keywords based on a common threshold (e.g., 0.5) for each frequency group, while also incorporating keywords from the reference keyword list and those generated by individual frequency clusters. The final fused list represents the most relevant and contextually appropriate keywords for generating the radiology report.

By focusing on frequency-based classification, our approach effectively mitigates the impact of class imbalance, ensuring that both common and rare keywords are accurately predicted. This results in radiology reports that are both comprehensive and accurate.

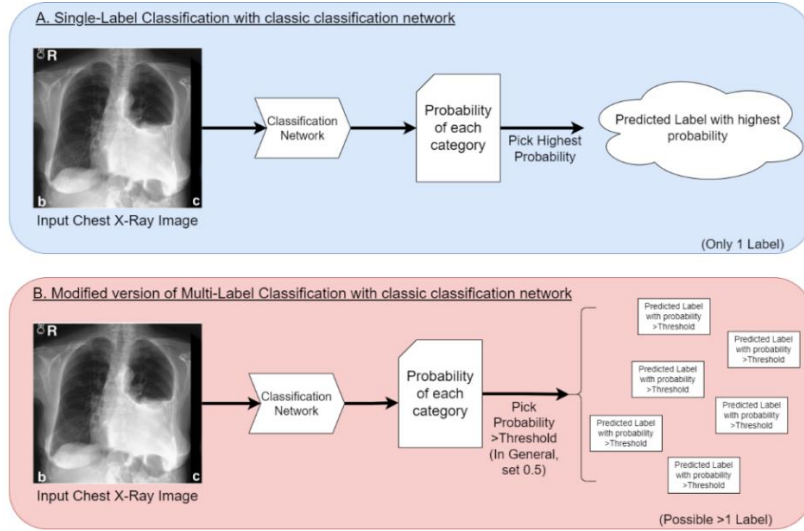


Figure 18. Comparison of our proposed multi-label classification approach with general single-label classification using classic classification networks. The figure illustrates the key differences between the two methodologies, highlighting how our proposed method accommodates multiple labels per input, enabling the identification of concurrent or overlapping conditions. In contrast, single-label classification assigns only one label per input, limiting its applicability in scenarios where multiple attributes or diagnoses need to be captured simultaneously.

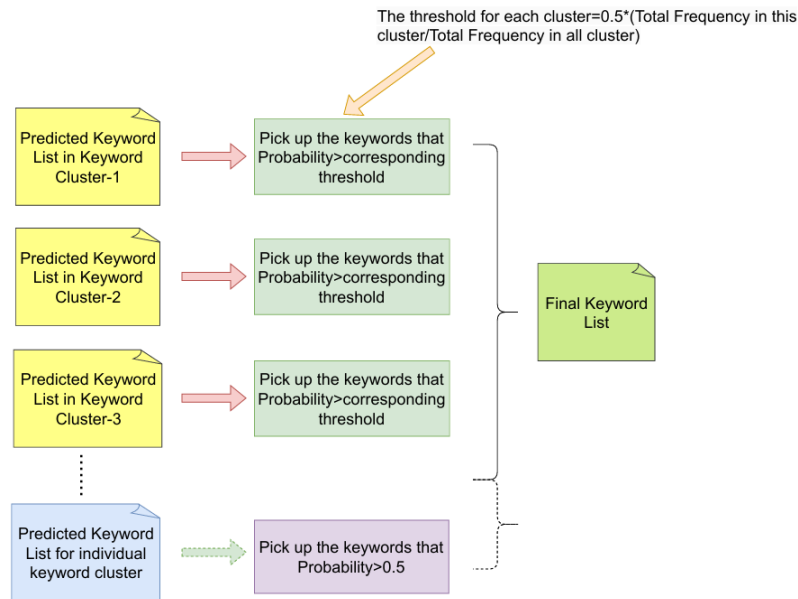


Figure 19. Diagram illustrating keyword list fusion in the frequency-based multi-label classification process. To account for the information density in high-frequency clusters, the threshold for these clusters is increased, reducing the likelihood of incorrect predictions for high-frequency keywords. Conversely, for low-frequency clusters that may contain rare but clinically important keywords, the threshold is slightly decreased to allow for the inclusion of more keywords. For individual clusters, a standard classification threshold of 0.5 is applied. The keywords from all clusters, after threshold-based filtering, are combined to form the final keyword list, which is then used for radiology report generation.

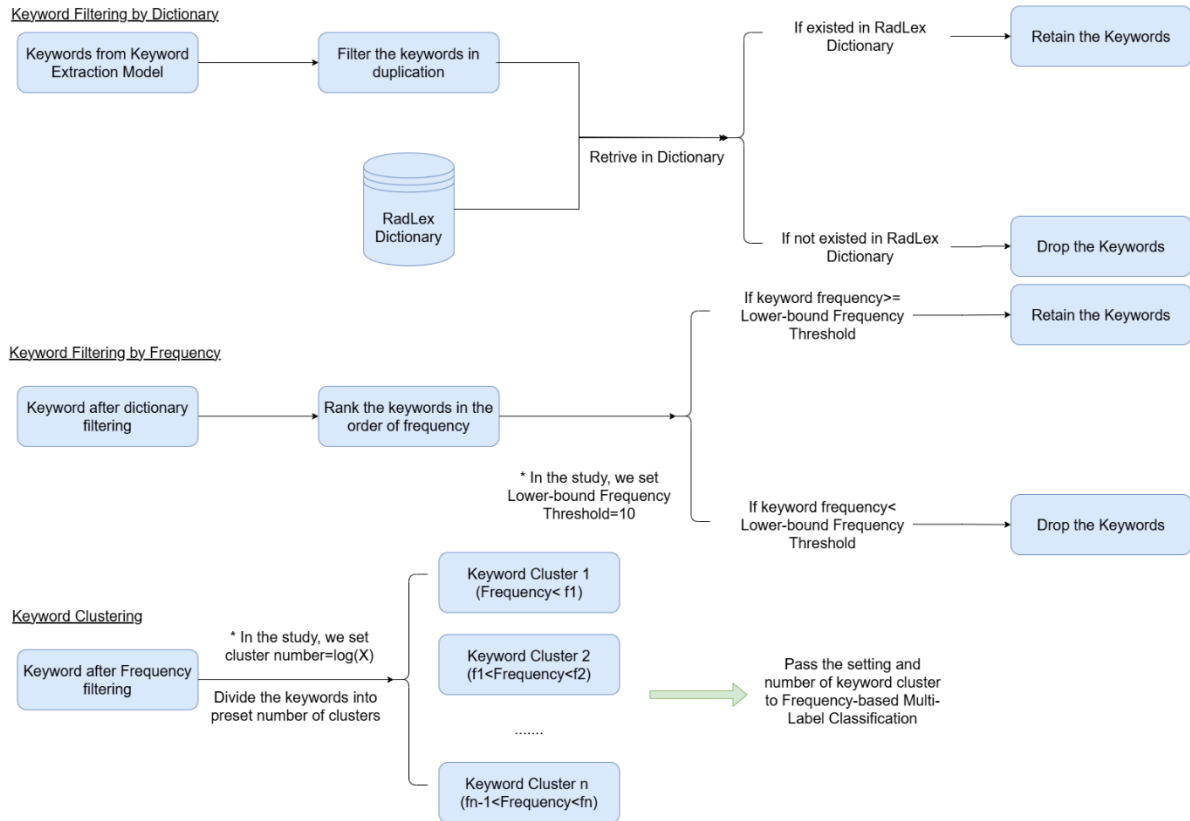


Figure 20. Full workflow of Automatic Keyword Adaption. It includes keyword filtering by radiology dictionary and frequency, and then clustering for multi-label classification

3.2.4 Radiology Report Generation from keywords based on Large Language Model

Following the extraction of the keyword list through frequency-based multi-label classification, the next step is to leverage large pre-trained language models (LLMs) to generate high-quality radiology reports. The integration of LLMs provides a significant advantage due to their ability to produce human-like, contextually accurate text. Unlike traditional text decoders, which often lack domain-specific training, LLMs pre-trained on medical corpora possess inherent capabilities for generating professional and contextually appropriate medical narratives. Fine-tuning these models with domain-specific data further enhances their adaptability to radiology-specific use cases.

In our framework, we employ the Text-to-Text Transformer (T5) model [175], utilizing the fine-tuning methodology demonstrated by Clinical-T5 [176], which is pre-trained on extensive medical datasets. Using pre-trained checkpoints as the foundation, we fine-tune the model with our dataset of extracted keywords and their corresponding radiology reports. This process aligns the model's generative capabilities with the unique characteristics of radiology report writing, ensuring both accuracy and fluency in the output.

The fine-tuned model transforms the extracted keywords into comprehensive radiology reports that reflect the clinical context and maintain a coherent, professional tone, which is shown in Figure 21. By focusing on the semantic alignment between the keywords and the generated text, our approach ensures that the resulting reports adhere to clinical standards while effectively communicating the relevant findings.

Furthermore, this framework diverges from traditional encoder-decoder architectures, where a text decoder generates reports based on encoded features. Instead, our keyword-driven approach simplifies the input space, leveraging the text-to-text LLM's ability to map concise, structured inputs (keywords) to expansive, descriptive outputs. This paradigm shift enhances the interpretability of the model and ensures that the generated reports maintain consistency with the extracted keywords.

By combining the generative power of text-to-text LLMs with fine-tuning on domain-specific data, our framework provides a robust and scalable solution for radiology report generation. The resulting reports are not only clinically accurate but also exhibit the fluency and readability expected in professional medical documentation, making this approach well-suited for practical deployment in vision-language applications within medical imaging.

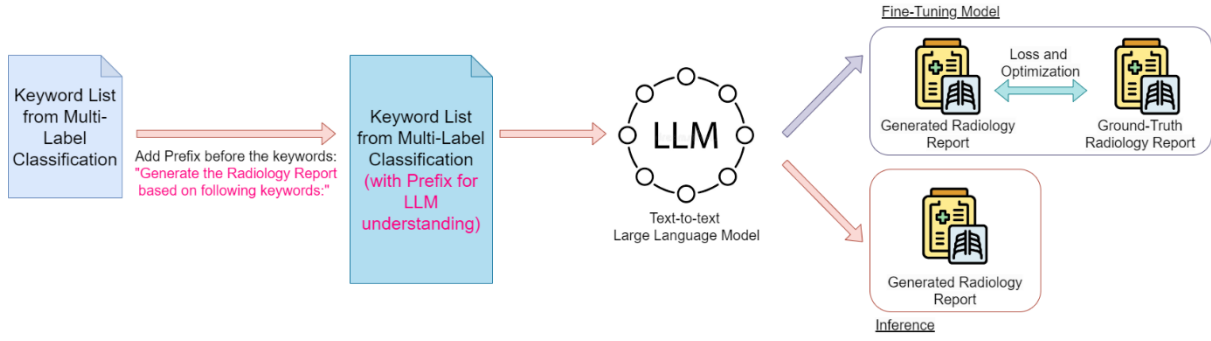


Figure 21. Diagram illustrating the workflow of radiology report generation based on multi-label classification within our proposed framework. The figure outlines the step-by-step process, beginning with the multi-label classification which contains relevant diagnostic labels. These labels are then used to guide the generation of structured and coherent radiology reports after adding prefixes to construct the command to the large language model. The diagram highlights the integration of multi-label classification with natural language generation techniques, demonstrating how our framework ensures accurate and contextually relevant report generation. This approach leverages the strengths of multi-label classification to capture complex and concurrent findings, enhancing the overall quality and utility of automated radiology reporting.

3.2.5 Loss Function

Our method comprises three main components: automatic keyword adaption, frequency-based multi-label classification, and radiology report generation. Each of these components plays a critical role in the overall process, with unique strategies for optimization and loss functions. Since the automatic keyword adaption component focuses on extracting and organizing keywords without any training or fine-tuning steps, it does not require a loss function or optimization process. Instead, it relies on heuristic methods and dictionary-based filtering to ensure the quality and relevance of the keywords.

In contrast, the frequency-based multi-label classification component is designed to address the challenges of class imbalance often encountered in medical imaging data. Inspired by advancements in multi-label classification for natural images, we employ an asymmetric loss function, which has been shown to effectively handle imbalanced datasets. This approach,

originally proposed by [164] , adapts the loss calculation for positive and negative samples differently, providing a tailored solution to the skewed distribution of labels in medical datasets. The asymmetric loss is defined as follows:

$$ASL(L_+) = (1 - p)^y + \log(p)$$

$$ASL(L_-) = (p_m)^y - \log(1 - p_m)$$

where p is the predicted probability, p_m is the shifted probability for negative samples, L_+ is the loss for positive samples, and L_- is the loss for negative samples. The shifted probability p_m is defined as:

$$p_m = \max(p - m, 0)$$

Here, the probability margin m is a tunable hyperparameter that adjusts the threshold for considering a sample as positive or negative. In practice, we apply dynamic optimization of the margin m within the loss function, allowing the model to adapt to varying class distributions without manual adjustments. This adaptive approach ensures that frequency-based multi-label classification network in each frequency cluster optimally handles its respective cluster, balancing sensitivity and specificity across different label frequencies.

For the radiology report generation process, we employ a standard cross-entropy loss function to guide the learning process. The cross-entropy loss helps ensure that the generated text aligns with the target distribution of clinical language, capturing key information accurately. The cross-entropy loss for fine-tuning is defined as follows:

$$CE(L) = - \sum_x P(x) \log G(x)$$

By integrating these tailored loss functions, our method ensures robust performance across all stages, from initial keyword extraction to final report generation. This comprehensive approach not only enhances the accuracy and relevance of the output but also supports the development of a user-friendly, clinically applicable system that can assist radiologists in their workflow.

4. Experiments and Results

4.1 *Implementation Details*

All experiments were conducted on two widely used public chest radiograph datasets: IU X-Ray [124] and MIMIC-CXR [124]. In addition to the image-text pairs, the RadLex radiology lexicon [177] was incorporated to support the Automatic Keyword Adaptation component of the framework. A summary of the basic dataset statistics and dictionary characteristics was provided in Table 4.

The proposed framework employed a unified subnetwork architecture for frequency-based multi-label classification, with ConvNeXt ([178]) selected as the backbone network due to its advanced feature extraction capabilities. The process began with automatic keyword adaptation and a detailed analysis of keyword distributions within the training datasets of IU X-Ray and MIMIC-CXR. To mimic real-world scenarios where test and validation sets remained unseen during training, the keyword frequency analysis was limited to the training sets, as illustrated in Table 4. To manage the distribution of keywords, a logarithmic split strategy ($\log x$) was applied based on the maximum frequency observed in each dataset. For IU X-Ray, frequency clusters were divided into three ranges: [10,100], [100,1000], and [1000,10,000]. Meanwhile, for MIMIC-CXR, five clusters were defined as [10,100], [100,1000], [1000,10,000], [10,000,100,000], and [100,000+]. This approach ensured a balanced representation of keywords across frequency ranges, thereby improving classification performance.

For radiology report generation, the framework employed a modified version of the Text-to-Text Transformer (T5) model [175], following the fine-tuning methodology described in [176].

Pre-trained checkpoints were sourced from HuggingFace’s repository ([170]) to minimize initialization and training costs. The model was fine-tuned using the extracted keywords and corresponding radiology reports, which enabled it to generate domain-specific, contextually accurate reports. This process adapted the model's language generation capabilities to align with the stylistic and clinical requirements of radiology documentation.

The training process was optimized for both tasks. For multi-label classification, the network was trained for 120 epochs per frequency cluster with a batch size of 4. The Adam optimizer was used with an initial learning rate of 0.0001. For the fine-tuning of the language model, training was performed over 10 epochs with a batch size of 2, using the Adam optimizer and an initial learning rate of 0.00005. All experiments were conducted on a workstation equipped with an Intel Core i7-11700 CPU and an NVIDIA RTX 3080 GPU with 10 GB of memory, ensuring efficient computational performance across tasks.

Table 4. Dataset descriptions for the IU X-Ray and MIMIC-CXR datasets, including details of keyword cluster information derived from the automatic keyword adaptation process. Additionally, the table provides information on the radiology dictionary used in the experiments and presents the proportion of keywords in the two experimental datasets relative to the full version of the dictionary.

Setting	Description
Applied Radiology Dictionary	
Dictionary Name	RadLex (Langlotz (2006))
Dictionary Version	4.2
Dictionary Source	https://radlex.org/
Total Number of Keywords	46,838
Dataset 1: IU X-Ray	
Basic Information	
Total Cases	3,851
Total Images	7,553
Total Pairs(1 Image/1 Report)	
Train Set	6,669
Test Set	589
Validation Set	295
Keyword Information	
Maximum Keyword Frequency	5502
Corresponding Highest Frequent Keyword	"no"
Keyword Frequency Cluster Split Number	3
Corresponding Keyword Frequency Cluster	[10,100], [100,1000], [1000,10000]
Ratio of keyword compared with Dictionary	5.315%
Keyword Cluster Description	
Number of Keywords in each cluster	
Cluster [10,100]	160
Cluster [100,1000]	73
Cluster [1000,10000]	15
The highest frequency in each cluster	
Cluster [10,100]	96
Cluster [100,1000]	967
Cluster [1000,10000]	5502
Dataset 2: MIMIC-CXR	
Basic Information	
Total Cases	227,835
Total Images	276,488
Total Pairs(1 Image/1 Report)	
Train Set	270,507
Test Set	3,858
Validation Set	2,123
Keyword Information	
Maximum Keyword Frequency	196051
Corresponding Highest Frequent Keyword	"pneumothorax"
Keyword Frequency Cluster Split Number	5
Corresponding Keyword Frequency Cluster	[10,100], [100,1000], [1000,10000], [10000,100000], [100000+]
Ratio of keyword compared with Dictionary	18.603%
Keyword Cluster Description	
Number of Keywords in each cluster	
Cluster [10,100]	429
Cluster [100,1000]	247
Cluster [1000,10000]	133
Cluster [10000,100000]	54
Cluster [100000+]	5
The highest frequency in each cluster	
Cluster [10,100]	99
Cluster [100,1000]	999
Cluster [1000,10000]	9841
Cluster [10000,100000]	96119
Cluster [100000+]	196051

4.2 *Evaluation Details and Metrics*

The primary objective of this study is radiology report generation; therefore, the evaluation focuses on assessing the quality of the generated reports. For the two datasets used in this work—IU X-Ray and MIMIC-CXR—we benchmark our proposed framework against a wide range of state-of-the-art (SOTA) models, as summarized in Tables 1 and 2 in the “Literature Review” (Section 1.2.4). To ensure a fair comparison, we re-implement the models proposed in [8] and [31] using their open-source code, applying the same dataset splits as in our work. The results from these re-implementations are included in our evaluation. For other SOTA models where re-implementation is not feasible, the reported values are taken directly from their original publications. If a model does not provide results for one of the datasets, we only report the metrics available in its original paper, avoiding secondary or indirect sources even when they exist.

The evaluation metrics follow standard practices in the radiology report generation literature. Language-based metrics include Bilingual Evaluation Understudy (BLEU) scores (BLEU-1, BLEU-2, BLEU-3, BLEU-4) ([179]), Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) ([180]), Consensus-based Image Description Evaluation (CIDEr) ([181]), and Metric for Evaluation of Translation with Explicit ORdering (METEOR) ([182]), each assessing different aspects of linguistic quality such as n-gram overlap, recall-oriented summarization, consensus-based similarity, and semantic matching. Beyond language quality, clinical accuracy is measured through CheXpert clinical evaluation metrics [183], which calculates F1 score, precision, and recall over 14 clinical labels. In addition, RadGraph evaluation scores [147]—Entity F1 and Relation F1—are reported to quantify how well the generated reports capture structured clinical entities and their relationships.

It is important to note that not all prior works report the complete set of metrics, resulting in occasional blank entries in our evaluation tables. Finally, consistent with standard practice, all comparisons are performed exclusively on the test sets of each dataset, ensuring comparability and reliability of the benchmark results.

4.3 Samples of Extracting Keywords from Radiology Reports and Failure Cases

In the context of radiology report generation using keyword-guided models, a critical initial step involves defining and extracting representative keywords from the original reports. The challenge lies in identifying whether these extracted keywords can accurately capture the essential clinical information without prior expert input from radiologists. To address this, we applied our proposed automatic keyword adaptation pipeline to extract keywords from the radiology reports within the IU X-Ray and MIMIC-CXR datasets. Subsequently, we conducted a manual verification of selected cases by manually annotating the keywords based on the radiology dictionary RadLex to assess the completeness and clinical validity of the extracted keywords, particularly focusing on whether important terms have been omitted. To illustrate the effectiveness of the keyword extraction process, we present sample cases in which manually annotated keywords are compared with those extracted automatically by our framework. These examples, visualized in Figure 22, include annotations of commonly occurring terms such as “normal” and “lung.” These frequent and generic terms are reliably captured by the extraction pipeline, aided by the support of the keyword classification model and cross-referencing with entries in the RadLex radiology dictionary. This demonstrates that our method significantly reduces the burden of manually annotating such standard terms.

However, the pipeline is not without limitations. Since our framework is not explicitly designed as a clinical entity extraction model, it may fail to identify less frequent but clinically significant terms. Furthermore, because the verification step involving the radiology dictionary occurs after keyword extraction, it cannot contribute new terms to the list—it only confirms whether the extracted terms are valid. Figure 23 presents representative failure cases from both the IU X-Ray and MIMIC-CXR datasets. These examples highlight missing keywords such as “cardiomediastinal,” which were not captured by the automated process but are clinically important for diagnosis.

Overall, our findings indicate that while automatic keyword adaptation performs well for general or high-frequency terms, incorporating a radiologist-guided post-processing step—allowing human reviewers to supplement the keyword list before multi-label classification—could improve both accuracy and clinical relevance. This hybrid approach would offer a more robust and adaptable solution for real-world medical imaging workflows.

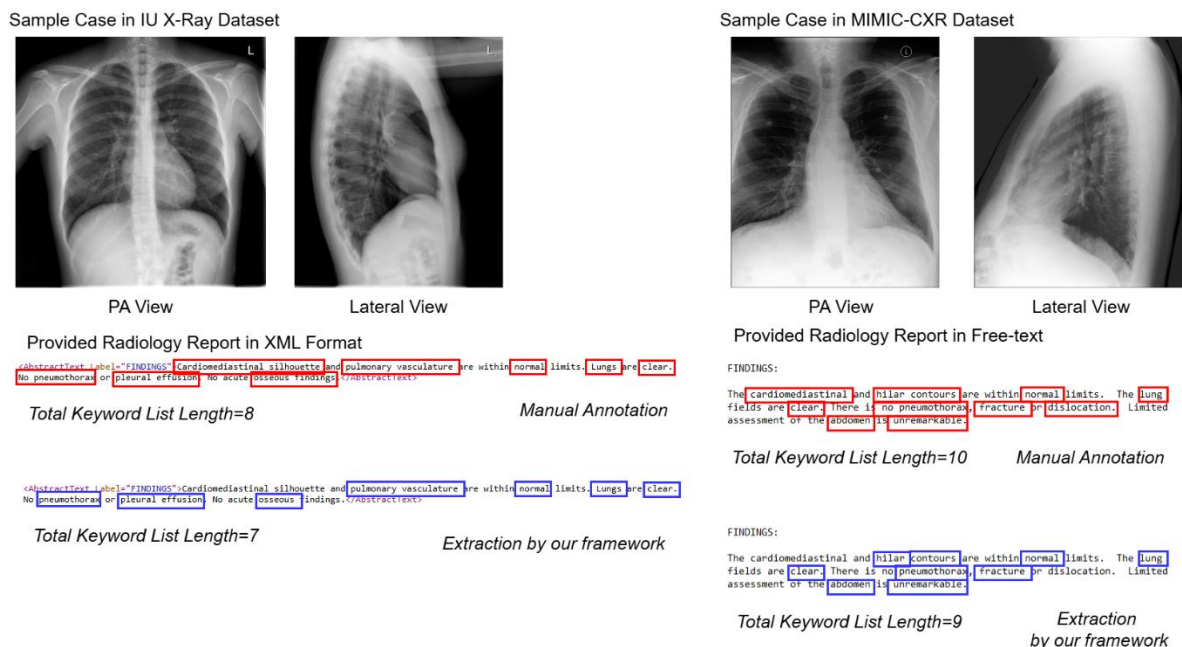


Figure 22. Sample cases with corresponding radiology reports and estimated keyword counts from the IU X-Ray and MIMIC CXR datasets. Representative cases were randomly selected from each dataset, and their associated radiology reports were obtained from the official sources. For each cases, Keywords were manually annotated and verified in the RadLex

radiology dictionary in Line 1 and we also show the extraction result by our framework in Line 2. The total keyword list length reflects the number of identified keywords in each case.

Failure Case in IU X-Ray: Missing Keyword



PA View



Lateral View

Radiology Report

The lungs are well-expanded and clear. No pleural effusion or pneumothorax is seen. The cardiomeastinal contour is normal. No acute osseous lesions are identified.

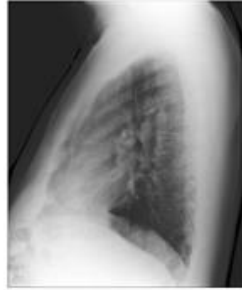
Predicted Keyword: expanded, clear, pneumothorax, contour, lungs, no, normal, osseous, effusion

Possible missing Keyword: lesions, pleural effusion, cardiomeastinal

Failure Case in MIMIC-CXR: Missing Keyword



PA View



Lateral View

Radiology Report

The cardiomeastinal and hilar contours are within normal limits. The lung fields are clear. There is no pneumothorax, fracture or dislocation. Limited assessment of the abdomen is unremarkable.

Predicted Keyword: normal, clear, dislocation, fracture, abdomen, limited, assessment, pneumothorax, unremarkable, lung, contour

Possible missing Keyword: cardiomeastinal, hilar

Figure 23. Sample failure cases from the IU X-Ray and MIMIC-CXR datasets. Representative examples were randomly selected from both datasets, with radiology reports sourced from official repositories. Keywords were initially extracted using the Automatic Keyword Adaptation process, and subsequently cross-checked with the RadLex radiology dictionary to identify potentially missing or unrecognized clinical terms.

4.4 Keyword Distribution Analysis

Before evaluating radiology report generation performance, we first analyzed the keywords produced by our automatic keyword adaptation mechanism on the IU X-Ray and MIMIC-CXR training datasets. Keyword frequency distributions were visualized in Figures 25 and 26, with statistical details—such as cluster splits and highest frequencies—summarized in Table 4. To manage distribution imbalance, we applied a

logarithmic split ($\log x$) based on the maximum observed frequency in each dataset. This approach, while used here to reduce the number of clusters, is not mandatory; frequency-based clusters can be flexibly adjusted to specific needs. To simplify the network and alleviate extreme class imbalance, keywords with frequencies below 10 were excluded.

The frequency distributions revealed that larger datasets exhibit more pronounced imbalance. In the smaller IU X-Ray dataset, the most frequent keyword, “no,” appeared over 5,000 times, resulting in three clusters covering frequencies from 10 to 5,000. In contrast, the MIMIC-CXR dataset showed a more extreme imbalance, with “pneumothorax” occurring over 190,000 times. To accommodate this, MIMIC-CXR was divided into five clusters ranging from 10 to over 100,000.

To validate the effectiveness of our automatic keyword adaptation approach, we compared the reduced keyword set against the full RadLex[151] dictionary. Table 4 shows the significant reduction achieved: the MIMIC-CXR dataset utilized only 18.6% of the RadLex terms, while the smaller IU X-Ray dataset used just 5.3%. This reduction minimizes computational complexity while retaining the relevance of the keywords to the task. Additionally, we evaluated the keyword coverage ratios within the test and validation sets using two strategies: Keyword-Based and Text-Based. The process of calculating the ratios is shown in Figure 24.

Keyword coverage was then assessed in the test and validation sets using two strategies: Keyword-Based and Text-Based. In the Keyword-Based Strategy, unique keywords were extracted from sample reports by splitting text into individual words and removing duplicates. These were matched against the adapted keyword set, and coverage ratio was calculated as the percentage of matching keywords relative to the total unique keywords. In the Text-Based Strategy, generated keywords were directly searched in the reports,

with matching words or phrases highlighted; coverage ratio was calculated as the proportion of the matched text length to the total report length.

Table 5 summarizes the results for both strategies, showing that the generated keywords achieved coverage ratios exceeding 50% in the radiology reports, even though the test and validation sets were unknown during the adaptation process. These findings demonstrate that the automatic keyword adaptation method effectively aligns with diverse clinical scenarios, ensuring high-quality and clinically meaningful outputs. The robust coverage ratios confirm the adaptability and reliability of our approach, making it a valuable tool for generating high-quality radiology reports in various medical contexts.

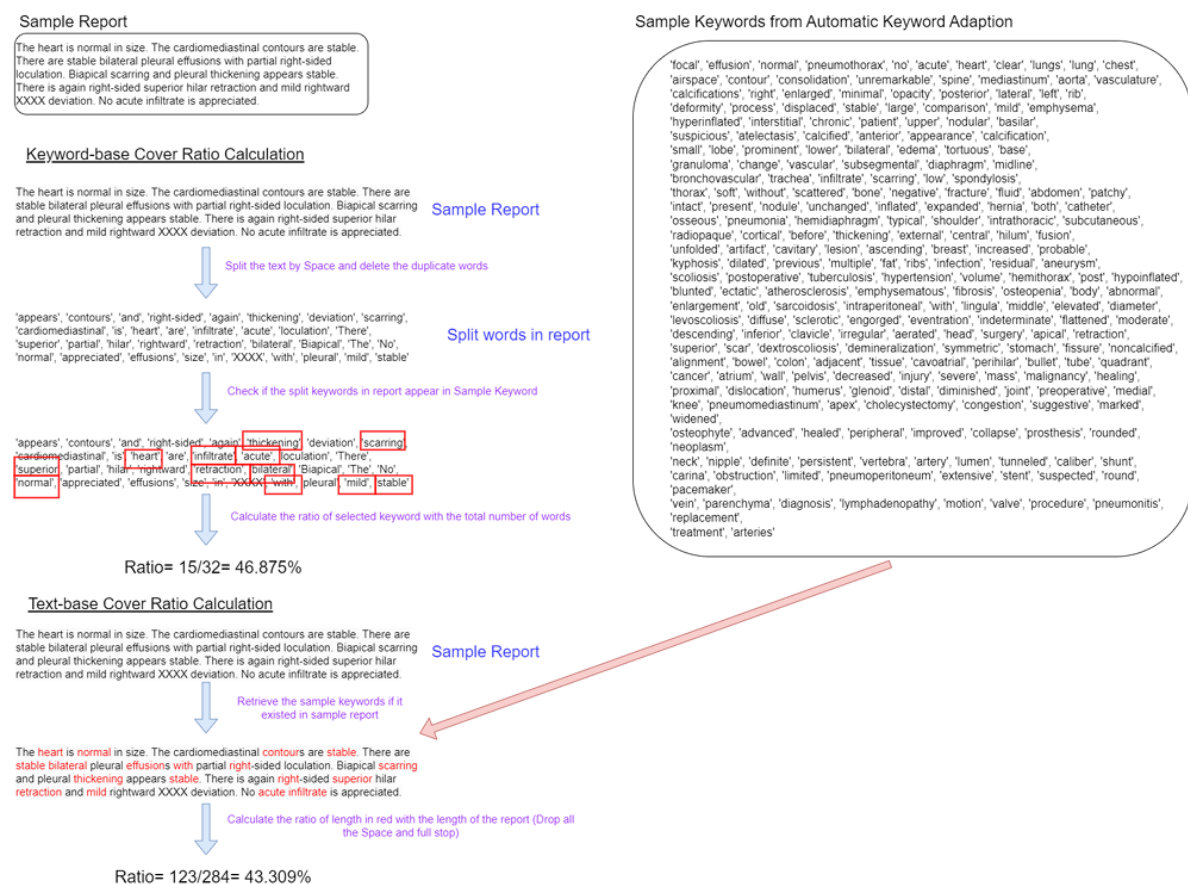


Figure 24. Methodology for calculating the keyword-based coverage ratio and text-based coverage ratio of the generated keywords in comparison to the ground truth radiology reports.

Table 5. Coverage ratio of keywords generated through Automatic Keyword Adaptation in the test and validation sets of the IU X-Ray and MIMIC-CXR datasets, evaluated using two strategies. In the Keyword-Based Strategy, radiology reports are split into unique keywords by removing duplicates, and the coverage ratio is calculated as the percentage of matched keywords from the generated keyword set relative to the total unique keywords in the report. In the Text-Based Strategy, the generated keyword set is searched directly within the report text, with matching words or phrases highlighted. The coverage ratio is then computed as the proportion of the total character length of matched words or phrases to the total character length of the report.

Set	Total Number of Images	Keyword-Based Cover Ratio	Text-Based Cover Ratio
Dataset 1: IU X-Ray			
Test Set	589	57.53%	54.61%
Validation Set	295	56.33%	53.54%
Dataset 2: MIMIC-CXR			
Test Set	3,858	59.20%	56.26%
Validation Set	2,123	64.13%	63.50%

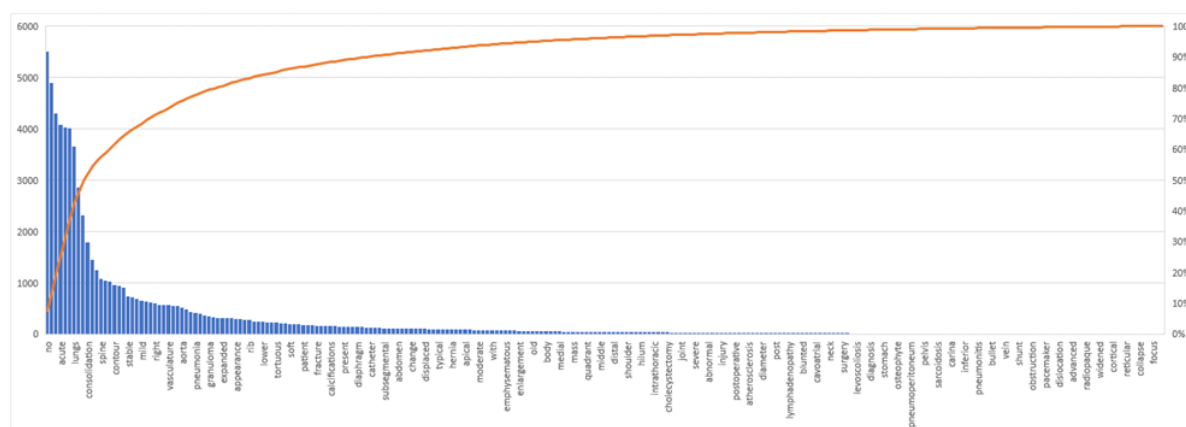


Figure 25. Pareto chart illustrating the keyword distribution in the IU X-Ray dataset. Blue bars represent the frequency of each keyword in the training set, while the orange line indicates the cumulative frequency ratio from the most frequent keyword to the current keyword compared with the total frequency of keyword.

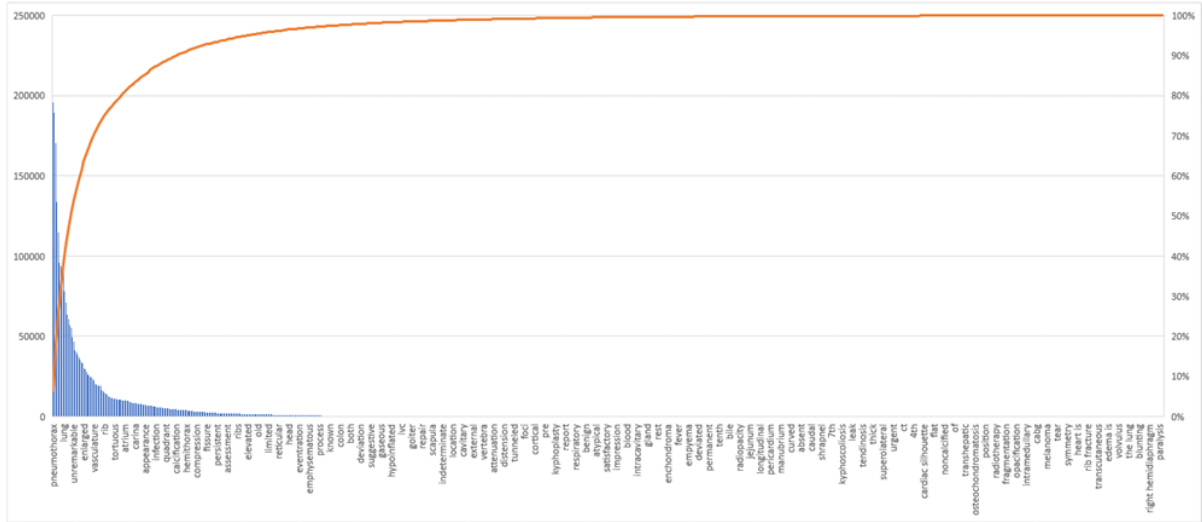


Figure 26. Pareto chart illustrating the keyword distribution in the MIMIC-CXR dataset. Blue bars represent the frequency of each keyword in the training set, while the orange line indicates the cumulative frequency ratio from the most frequent keyword to the current keyword compared with the total frequency of keyword.

4.5 Quantitative Analysis of Radiology Report Generation

We conducted a quantitative analysis of radiology report generation to compare the performance of our proposed framework with state-of-the-art (SOTA) methods. The results are summarized in Table 6 for the IU X-Ray test set and Table 7 for the MIMIC-CXR test set.

Our deep learning framework consistently outperforms SOTA approaches across all evaluation metrics on both datasets. Specifically, on the IU X-Ray dataset, our method achieves significant performance improvements compared to the best metrics reported by other methods: a 23.9% increase in BLEU-1 (0.719 vs. 0.580), a 59.8% increase in BLEU-2 (0.625 vs. 0.391), a 76.2% increase in BLEU-3 (0.564 vs. 0.320), a 92.9% increase in BLEU-4 (0.521 vs. 0.270), a 12.6% increase in ROUGE-L (0.639 vs. 0.567), a 12.5% increase in METEOR (0.386 vs. 0.343), and an 83.8% increase in CIDEr (1.274 vs. 0.693).

Similarly, on the MIMIC-CXR dataset, our framework demonstrates substantial gains: a 32.1% increase in BLEU-1 (0.559 vs. 0.423), a 63.6% increase in BLEU-2 (0.437 vs. 0.267), a 90.8% increase in BLEU-3 (0.355 vs. 0.186), a 75.5% increase in BLEU-4 (0.295 vs. 0.168), a 49.8% increase in ROUGE-L (0.469 vs. 0.313), a 5.9% increase in METEOR (0.284 vs. 0.268), and a 78.0% increase in CIDEr (1.996 vs. 1.121). For the MIMIC-CXR dataset, we also apply the Clinical Evaluation metrics which is shown in Table 8 and Table 9 and it gets the performance increasing: a 47.57\% in Precision of ChexPert (0.7448 vs. 0.5047), a 11.46\% in Recall of ChexPert (0.6610 vs. 0.593), a 39.24\% in F1 Score of ChexPert (0.7004 vs. 0.503), a 35.60\% in entity F1 of RadGraph (0.5980 vs. 0.441) and a 8.47\% in relation F1 of RadGraph (0.3840 vs. 0.354).

Notably, the most significant improvements are observed in stricter metrics, such as BLEU-4 and CIDEr. These metrics emphasize precise and contextually relevant information in the generated reports, highlighting the effectiveness of our keyword-based mechanism. The results suggest that our framework excels in generating reports that are not only informative but also linguistically fluent and clinically coherent.

The superior performance achieved by our framework is attributed to the integration of Automatic Keyword Adaptation and Frequency-Based Multi-Label Classification, which effectively enhance the alignment between extracted keywords and the content of the generated reports. This synergy ensures that our method produces high-quality radiology reports that surpass existing approaches in terms of both accuracy and interpretability.

Table 6. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the IU X-Ray dataset. For most methods, the results are cited directly from their respective publications, presented under "Paper Report Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean \pm standard deviation, and results from re-training and the proposed pipeline are presented with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.

Work	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Paper Report Performance							
Jing et al. (2017)	0.517	0.386	0.306	0.247	0.447	0.217	0.327
Xue et al. (2018)	0.464	0.358	0.27	0.195	0.366	0.274	/
Harzig et al. (2019)	0.373	0.246	0.175	0.126	0.315	0.163	0.359
Xie et al. (2019)	0.443	0.337	0.236	0.181	0.347	/	0.374
Yuan et al. (2019)	0.529	0.372	0.315	0.255	0.453	0.343	/
Li et al. (2019)	0.482	0.325	0.226	0.162	0.339	/	0.28
Jing et al. (2019)	0.464	0.301	0.21	0.154	0.362	/	0.275
Chen et al. (2020b)	0.47	0.304	0.219	0.165	0.371	0.187	/
Zhang et al. (2020b)	0.441	0.291	0.203	0.147	0.367	/	0.304
Wang et al. (2021)	0.487	0.346	0.27	0.208	0.359	/	0.452
Alfarghaly et al. (2021)	0.387	0.245	0.166	0.111	0.289	0.164	0.257
Liu et al. (2021b)	0.492	0.314	0.222	0.169	0.381	0.193	/
Liu et al. (2021a)	0.483	0.315	0.224	0.168	0.376	0.19	0.351
Yang et al. (2021b)	0.496	0.327	0.238	0.178	0.381	/	0.382
Nooralahzadeh et al. (2021)	0.486	0.317	0.232	0.173	0.39	0.192	/
Yang et al. (2021a)	0.497	0.319	0.23	0.174	0.399	/	0.407
Zhou et al. (2021)	0.536	0.391	0.314	0.252	0.448	0.228	0.339
Li et al. (2022)	0.467	0.334	0.261	0.215	0.415	0.201	/
You et al. (2022a)	0.484	0.313	0.225	0.173	0.379	0.204	/
Wang et al. (2022b)	0.505	0.34	0.247	0.188	0.382	0.208	/
Sirshar et al. (2022)	0.58	0.342	0.263	0.155	/	/	/
Yan et al. (2022b)	/	/	0.256	/	0.341	/	0.38
Wang et al. (2022d)	0.496	0.319	0.241	0.175	0.377	/	0.449
Yu and Zhang (2022)	0.457	0.305	0.216	0.171	0.391	/	0.426
Chen et al. (2022)	0.475	0.309	0.222	0.17	0.375	0.191	/
Wang et al. (2022a)	0.525	0.357	0.262	0.199	0.411	0.22	0.359
Nicolson et al. (2022)	0.4732	0.3039	0.2242	0.1754	0.3758	0.1997	0.6935
Delbrouck et al. (2022)	/	/	/	0.121	0.306	/	/
You et al. (2022b)	0.479	0.319	0.222	0.174	0.377	0.193	/
Wu et al. (2022)	0.458	0.324	0.238	0.18	0.369	0.206	0.287
Yan et al. (2022a)	0.482	0.313	0.232	0.181	0.381	0.203	0.735
Wang et al. (2022c)	0.45	0.301	0.213	0.158	0.384	/	0.34
Qin and Song (2022)	0.494	0.321	0.235	0.181	0.384	0.201	/
Tanwani et al. (2022)	0.58	0.44	0.32	0.27	/	/	/
Wang et al. (2023a)	0.505	0.345	0.243	0.176	0.396	0.205	/
Kong et al. (2022)	0.484	0.333	0.238	0.175	0.415	0.207	/
Li et al. (2023a)	/	/	/	0.163	0.383	0.193	0.586
Yang et al. (2021c)	0.478	0.344	0.248	0.18	0.398	/	0.439
Kale et al. (2023a)	0.423	0.256	0.194	0.165	0.444	0.15	/
Huang et al. (2023)	0.525	0.36	0.251	0.185	0.409	0.242	/
Wang et al. (2023b)	0.483	0.322	0.228	0.172	0.38	0.192	0.435
Hou et al. (2023)	0.51	0.346	0.255	0.195	0.399	0.20	/
Wang et al. (2023c)	0.488	0.316	0.228	0.173	0.377	0.211	0.438
Kale et al. (2023b)	0.402	0.322	0.285	0.17	0.567	0.455	0.473
Li et al. (2023b)	0.53	0.365	0.263	0.2	0.405	0.218	0.501
Mohsan et al. (2022)	0.532	0.344	0.233	0.158	0.387	0.218	0.5
Chen et al. (2023)	0.505	0.334	0.245	0.19	0.394	0.21	0.592
Zhang et al. (2023a)	0.482	0.31	0.221	0.165	0.377	0.195	/
Liu et al. (2024)	0.499	0.323	0.238	0.184	0.39	0.208	/
Zhou et al. (2024)	/	/	/	0.208	0.387	0.216	/
Yi et al. (2024a)	0.5	0.349	0.256	0.194	0.402	0.218	/
Parres et al. (2024)	/	/	/	0.149	0.341	/	/
Yi et al. (2024b)	0.539	0.380	0.278	0.210	0.416	0.223	/
Re-Train Performance							
R2Gen (Chen et al. (2020b))	0.4514	0.2988	0.2163	0.1631	0.3377	0.201	0.5988
Cvt2Distgen2 (Nicolson et al. (2022))	0.4182	0.2758	0.2037	0.1594	0.3315	0.1923	0.6784
Our Performance	0.7190\pm0.2101	0.6250\pm0.2713	0.5645\pm0.3057	0.5215\pm0.3376	0.6392\pm0.2554	0.3861\pm0.2218	3.2749\pm1.0106

Table 7. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the MIMIC-CXR dataset. For most methods, the results are cited directly from their respective publications, presented under "Paper Report Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean \pm standard deviation, and results from re-training and the proposed pipeline are presented with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.

Work	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Paper Report Performance							
Chen et al. (2020b)	0.353	0.218	0.145	0.103	0.277	0.142	/
Liu et al. (2021b)	0.35	0.219	0.152	0.109	0.283	0.151	/
Liu et al. (2021a)	0.36	0.224	0.149	0.106	0.284	0.149	0.237
Yang et al. (2021b)	0.363	0.228	0.156	0.115	0.284	/	0.203
Nooralahzadeh et al. (2021)	0.378	0.232	0.154	0.107	0.272	0.145	/
Yang et al. (2021a)	0.386	0.237	0.157	0.111	0.274	/	0.111
Hou et al. (2021)	0.232	/	/	/	0.24	0.101	0.493
Zhou et al. (2021)	0.372	0.241	0.168	0.123	0.335	0.19	1.121
Yan et al. (2021)	0.373	/	/	0.107	0.274	0.144	/
Wang et al. (2022e)	0.413	0.266	0.186	0.136	0.298	0.17	0.429
You et al. (2022a)	0.378	0.235	0.156	0.112	0.283	0.158	/
Wang et al. (2022b)	0.395	0.253	0.17	0.121	0.284	0.147	/
Yan et al. (2022b)	/	/	0.145	/	0.225	/	0.16
Wang et al. (2022d)	0.351	0.223	0.157	0.118	0.287	/	0.281
Yu and Zhang (2022)	0.347	0.235	0.149	0.106	0.28	/	0.552
Chen et al. (2022)	0.353	0.218	0.148	0.106	0.278	0.142	/
Wang et al. (2022a)	0.344	0.215	0.146	0.105	0.279	0.138	/
Nishino et al. (2022)	/	/	/	0.168	0.122	/	/
Nicolson et al. (2022)	0.3928	0.2478	0.1713	0.1267	0.2863	0.1545	0.3892
Delbrouck et al. (2022)	/	/	/	0.116	0.259	/	/
Wu et al. (2022)	0.34	0.212	0.145	0.103	0.27	0.139	0.109
Serra et al. (2022)	0.363	0.245	0.178	0.136	0.313	0.161	/
Yan et al. (2022a)	0.356	0.222	0.151	0.111	0.28	0.14	0.154
Qin and Song (2022)	0.381	0.232	0.155	0.109	0.287	0.151	/
Wang et al. (2023a)	0.363	0.235	0.164	0.118	0.301	0.136	/
Kong et al. (2022)	0.423	0.261	0.171	0.116	0.286	0.168	/
Li et al. (2023a)	/	/	/	0.109	0.284	0.15	0.281
Tanida et al. (2023)	0.373	0.249	0.175	0.126	0.264	0.168	0.495
Yang et al. (2021c)	0.362	0.251	0.188	0.143	0.326	/	0.273
Huang et al. (2023)	0.393	0.243	0.159	0.113	0.285	0.16	/
Wang et al. (2023b)	0.386	0.25	0.169	0.124	0.291	0.152	0.362
Hou et al. (2023)	0.407	0.256	0.172	0.123	0.293	0.162	/
Wang et al. (2023c)	0.411	0.267	0.186	0.134	0.297	0.16	0.269
Kale et al. (2023b)	0.253	0.188	0.169	0.163	0.348	0.268	0.331
Li et al. (2023b)	0.363	0.229	0.158	0.107	0.289	0.157	0.246
Chen et al. (2023)	0.4	0.245	0.165	0.119	0.28	0.15	0.19
Zhang et al. (2023a)	0.362	0.229	0.157	0.113	0.284	0.153	/
Liu et al. (2024)	0.402	0.262	0.18	0.128	0.291	0.175	/
Zhou et al. (2024)	/	/	/	0.122	0.296	0.165	/
Yi et al. (2024a)	0.398	0.248	0.169	0.121	0.281	0.149	/
Parres et al. (2024)	/	/	/	0.116	0.265	/	/
Zhang et al. (2024)	0.391	0.258	0.182	0.129	0.282	0.175	0.526
Yi et al. (2024b)	0.400	0.253	0.171	0.120	0.296	0.154	/
Re-Train Performance							
R2Gen (Chen et al. (2020b))	0.3058	0.1834	0.1221	0.0868	0.2386	0.1299	0.1466
Cvt2Distgen2 (Nicolson et al. (2022))	0.2952	0.1839	0.1263	0.0927	0.2473	0.1308	0.1814
Our Performance	0.5599\pm0.1607	0.4379\pm0.1736	0.3557\pm0.1824	0.2953\pm0.1958	0.4699\pm0.1687	0.2842\pm0.1018	1.9964\pm1.4518

Table 8. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the MIMIC-CXR dataset for the Clinical Evaluation (CE) Metrics in CheXpert Label Accuracy. For most methods, the results are cited directly from their respective publications, presented under "Paper Report Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean \pm standard deviation, and results from re-training and the proposed pipeline are presented with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.

Work	Precision	Recall	F1 Score
Paper Report Performance			
Chen et al. (2020b)	0.333	0.273	0.276
Liu et al. (2021b)	0.352	0.298	0.303
Yang et al. (2021b)	0.458	0.348	0.371
Nooralahzadeh et al. (2021)	0.240	0.428	0.308
Yang et al. (2021a)	0.420	0.339	0.352
Yu and Zhang (2022)	0.447	0.593	0.503
Chen et al. (2022)	0.334	0.275	0.278
Nicolson et al. (2022)	0.367	0.418	0.391
Serra et al. (2022)	0.428	0.459	0.443
Yan et al. (2022a)	0.353	0.310	0.297
Qin and Song (2022)	0.342	0.294	0.292
Kong et al. (2022)	0.482	0.563	0.519
Tanida et al. (2023)	0.461	0.475	0.447
Huang et al. (2023)	0.371	0.318	0.321
Wang et al. (2023b)	0.364	0.309	0.311
Hou et al. (2023)	0.416	0.418	0.385
Wang et al. (2023c)	0.392	0.387	0.389
Chen et al. (2023)	0.489	0.340	0.401
Zhang et al. (2023a)	0.38	0.342	0.335
Liu et al. (2024)	0.465	0.482	0.473
Yi et al. (2024a)	0.319	0.509	0.393
Zhang et al. (2024)	0.486	0.493	0.462
Yi et al. (2024b)	0.392	0.335	0.342
Re-Train Performance			
R2Gen (Chen et al. (2020b))	0.5047 \pm 0.2413	0.3838 \pm 0.2109	0.4361 \pm 0.2094
Cvt2Distgen2(Nicolson et al. (2022))	0.4627 \pm 0.2310	0.3423 \pm 0.2566	0.3935 \pm 0.2280
Our Performance	0.7448\pm0.1215	0.6610\pm0.1820	0.7004\pm0.1513

Table 9. Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the MIMIC-CXR dataset for the Clinical Evaluation (CE) Metrics in RadGraph F1. For most methods, the results are cited directly from their respective publications, presented under "Paper Report Performance." Additionally, two classic radiology report generation methods were re-trained, with their results reported under "Re-Train Performance." The performance of the proposed method is reported as mean \pm standard deviation, and results from re-training and the proposed pipeline are presented with precision up to four decimal places. For "Paper Report Performance," the decimal places are retained as reported in the original publications. A "/" in the performance metrics indicates that the corresponding metric was not reported in the original paper.

Work	RadGraph entity F1	RadGraph relation F1
Paper Report Performance		
Delbrouck et al. (2022)	0.441	0.299
Parres et al. (2024)	/	0.354
Re-Train Performance		
R2Gen (Chen et al. (2020b))	0.2545 \pm 0.1395	0.1096 \pm 0.1248
Cvt2Distgen2(Nicolson et al. (2022))	0.2497 \pm 0.1494	0.1056 \pm 0.1279
Our Performance	0.5980 \pm 0.1651	0.3840 \pm 0.2106

4.6 Qualitative Analysis of Radiology Report Generation

In addition to the quantitative evaluation of radiology report generation, we present qualitative examples to illustrate the performance of our framework and compare it with state-of-the-art (SOTA) methods. To facilitate a comprehensive comparison, we selected high-performing SOTA methods and generate their radiology reports using their publicly available source code. The generated reports were compared with those produced by our framework, with key information (predicted keywords) highlighted in both the ground-truth and generated reports for reference. The results are visualized in Figure 27 for the IU X-Ray dataset and Figure 28 for the MIMIC-CXR dataset.

The visualizations demonstrate that our framework, enabled by the integration of Automatic Keyword Adaptation and Frequency-Based Multi-Label Classification, produces reports that effectively capture the most relevant information associated with the radiology images. For

instance, in Sample 1 of the IU X-Ray dataset, our framework identified and incorporated keywords such as "pneumothorax," "vasculature," "effusion," and "granuloma," while in Sample 2 of the MIMIC-CXR dataset, it highlighted terms like "consolidation," "atelectasis," "engorged," "enlargement," and "aorta." These keywords were accurately integrated into the generated reports, ensuring that they align with the clinical findings presented in the images.

Moreover, the reports generated by our framework exhibit a structure and clarity similar to the ground-truth reports. This structural consistency enhances readability and aligns with the expectations of clinical documentation. The ability to maintain such structural integrity is attributed to the utilization of the keyword-based mechanism, which replaces traditional black-box features in computer vision with interpretable and clinically relevant keywords.

The superior performance of our framework can also be attributed to the generative capabilities of the large language model. By leveraging keywords as inputs instead of relying solely on text decoders and non-explainable features, the framework produces high-quality reports that are both accurate and interpretable. Unlike existing SOTA methods, which often generate reports directly from extracted features in an unexplainable manner, our approach integrates explainable keyword lists to ensure that the generated reports are clinically meaningful and reflective of key information.

These results confirm that the integration of explainable keyword mechanisms with large language models offers a robust and interpretable solution for radiology report generation, producing reports that are both structurally coherent and clinically relevant.





X-Ray Image	Reference Report	Predicted Keywords in report	Our proposed method	State of the art Method1: Cx2019gen (Official Checkpoint)	State of the art Method2: R259n (Official Checkpoint)
	Frontal and lateral views of the chest show an unchanged cardiomeastinal silhouette. Right-sided aortic XXXX. Normal mediastinal contour, pulmonary XXXX and vasculature, central airways and aeration of the lungs. There is right basal XXXX patchy opacity and bilateral atelectasis or scarring. There is no pleural effusion or pneumothorax. Right apical calcified granuloma noted.	granuloma, atelectasis , scarring , pneumothorax, calcified, right, chest, apical, vasculature, lungs, opacity, lateral, patchy, no, effusion	Patchy right apical pleural effusion. No pneumothorax. Frontal and lateral views of the chest show normal size and configuration of the cardiac silhouette. Normal mediastinal contour, pulmonary XXXX and vasculature, central airways and aeration of the lungs. Calcified granuloma in the right apical region. No pleural effusion or pneumothorax.	the lungs are clear bilaterally, specifically no evidence of focal consolidation pneumothorax or pleural effusion. cardio mediastinal silhouette is unremarkable. visualized osseous structures of the thorax are without acute abnormality.	the heart size and pulmonary vascularity appear within normal limits. the lungs are free of focal airspace disease, no pleural effusion or pneumothorax is seen.
	There is minimal scarring within the left lung base. The lungs are otherwise clear. Heart size is normal. No pneumothorax.	scarring, clear, heart, pneumothorax, lungs, base, no, normal, left, lung, minimal	Minimal scarring in the left lung base. No pneumothorax. Heart size is normal. Lungs are clear.	the cardiomeastinal 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.	the cardiomeastinal 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.
	The heart is normal in size. The mediastinum is unremarkable. Mild emphysematous changes without focal consolidation. There is no pleural effusion. XXXX lingular scarring or atelectasis noted.	heart, unremarkable, atelectasis, consolidation, scarring, mild, mediastinum, focal, no, normal, emphysematous, effusion	Emphysematous changes. No focal consolidation. The heart is normal in size. The mediastinum is unremarkable. Mild scarring or atelectasis is noted. No focal consolidation is seen.	the cardiomeastinal 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.	the cardiomeastinal 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.
	There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures intact.	heart, pneumothorax, suspicious, consolidation, focal, intact, no, normal, osseous	No focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures intact.	the lungs are clear bilaterally, specifically no evidence of focal consolidation pneumothorax or pleural effusion. cardio mediastinal silhouette is unremarkable. visualized osseous structures of the thorax are without acute abnormality.	the cardiomeastinal silhouette is normal in size and contour. no focal consolidation pneumothorax or large pleural effusion. normal xxxx.

Figure 27. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the IU X-Ray dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue. The keywords in red are not shown in the generated radiology report by our framework because it is not forced to contain in the generated reports in our commands to the large language model.




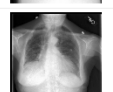
X-Ray Image	Reference Report	Predicted Keywords in report	Our proposed method	State-of-the-art Method1: Cx2019gen (Official Checkpoint)	State-of-the-art Method2: R259n (Official Checkpoint)
	The cardiac, mediastinal and hilar contours appear stable. The heart is normal in size. There is no pleural effusion or pneumothorax. The lungs appear clear. The patient is status post anterior cervical fusion. Surgical clips project over the left upper quadrant. There has been no significant change.	stable, left, change, patient, no, normal, fusion, effusion, pneumothorax, anterior, lungs, clear, quadrant, heart	The patient is status post incompletely imaged anterior cervical fusion. The heart is normal in size. The mediastinal and hilar contours appear stable. There is no pleural effusion or pneumothorax. The lungs appear clear. Surgical clips project over the left upper quadrant. There has been no significant change.	pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.	pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomeastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.
	Moderate enlargement of the cardiac silhouette with a left ventricular predominance is unchanged. The aorta remains tortuous, and the hilar contours are stable. Pulmonary vascularity is not engorged. There is minimal atelectasis within the lung bases, but no focal consolidation is present. No pleural effusion or pneumothorax is identified. There are no acute osseous abnormalities.	stable, atelectasis, left, engorged, focal, acute, no, effusion, pneumothorax, tortuous, consolidation, osseous, enlargement, aorta, lung	There is mild enlargement of the cardiac silhouette with a left ventricular predominance. The aorta remains tortuous. The mediastinal and hilar contours are stable. Pulmonary vascularity is not engorged. Streaky opacities in the lung bases likely reflect atelectasis. No focal consolidation, pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.	frontal and lateral views of the chest were obtained. there are low lung volumes which accentuate the bronchovascular markings particularly at the lung bases. given this there may be mild pulmonary vascular congestion. no definite focal consolidation is seen. there is no large pleural effusion or pneumothorax. the cardiac and mediastinal silhouettes are stable.	frontal and lateral views of the chest. the lungs are clear of focal consolidation effusion or pulmonary vascular congestion. the cardiomeastinal silhouette is stable. no acute osseous abnormality is identified.
	There is moderate pulmonary edema, but no pleural effusion or pneumothorax. Heart size is top-normal. Mediastinal contours are within normal limits. Osseous structures are intact.	stable, edema, normal, effusion, intact, pneumothorax, osseous, heart	There is no focal consolidation, pleural effusion, pneumothorax, or pulmonary edema identified. The heart size is normal. Mediastinal contours are stable. The osseous structures are intact.	the cardiac mediastinal and hilar contours appear stable. there is no pleural effusion or pneumothorax. the lungs appear clear. there has been no definite change allowing for differences in technique including better inspiration on this examination.	single portable view of the chest. relatively low lung volumes are seen with secondary crowding of the bronchovascular markings. there is no confluent consolidation or large effusion. the cardiomeastinal silhouette is within normal limits. no acute osseous abnormality is identified.
	There is mild enlargement of the cardiac silhouette which is unchanged. Mediastinal and hilar contours are stable. The pulmonary vascularity is not engorged. Chronic interstitial abnormalities are again seen diffusely, more pronounced at the lung bases with fibrotic changes. No focal consolidation, pleural effusion or pneumothorax is identified. There is diffuse calcification of the aorta.	stable, mild, diffuse, engorged, interstitial, focal, effusion, pneumothorax, consolidation, calcification, lung, enlargement, chronic, aorta, unchanged	Mild enlargement of the cardiac silhouette is unchanged. Diffuse atherosclerotic calcification of the thoracic aorta is again noted. The mediastinal and hilar contours are stable. Pulmonary vascularity is not engorged. Diffuse increased interstitial markings are similar compared to the prior exam. compatible with chronic lung disease. No focal consolidation, pleural effusion or pneumothorax is present.	frontal and lateral views of the chest were obtained. there are relatively low lung volumes. there is diffuse increase in interstitial markings bilaterally which may be due to mild interstitial edema versus atypical infection. no focal consolidation is seen. there is no pleural effusion or pneumothorax. the cardiac and mediastinal silhouettes are stable.	frontal and lateral views of the chest were obtained. there are relatively low lung volumes which accentuate the bronchovascular markings. given this there is diffuse increase in interstitial markings bilaterally which may be due to mild interstitial edema versus atypical infection. no focal consolidation is seen. there is no pleural effusion or pneumothorax. the cardiac and mediastinal silhouettes are stable.

Figure 28. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the MIMIC-CXR dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue.

4.7 Ablation Study

4.7.1 Performance of Multi-Label Classification Across Network Architectures

As the link between chest radiology images and their associated keywords, the accuracy of the multi-label classification played a critical role in the overall performance of radiology report generation. However, evaluating multi-label classification performance was challenging due to the lack of ground truth annotations for keyword prediction in the IU X-Ray and MIMIC-CXR datasets. To address this, we used the keywords extracted by the automatic keyword adaptation mechanism as pseudo ground truth. This allowed us to monitor classification performance and compare the impact of different network architectures.

In addition to the ConvNeXt backbone used in our experiments, we evaluated the performance of several alternative network architectures, including ResNeXt ([65]), ResNet ([66]), VGG16 ([67]), EfficientNet ([68]), NASNet ([69]), Res2Net ([70]). These networks were tested in the multi-label classification stage and subsequently in radiology report generation, using the same large language model to ensure consistency. As there is no directly comparable work on keyword extraction and prediction from radiology reports, we focused on performance comparisons across network structures and provide results for each frequency cluster.

The results of the multi-label classification are presented in Table 10 (IU X-Ray) and Table 11 (MIMIC-CXR), while the corresponding performance in radiology report generation is shown in Table 12 for both datasets. Our analysis indicated that ConvNeXt achieves the highest performance on the IU X-Ray dataset and competitive results on the MIMIC-CXR dataset. Given the absence of ground truth in real-world medical

scenarios, ConvNeXt emerged as a reasonable choice for the multi-label classification subnetwork. Furthermore, the performance breakdown across frequency clusters reveals that high-frequency keywords are generally predicted with greater accuracy than low-frequency keywords, consistent with the observation that frequently occurring terms are easier to predict.

We also evaluated radiology report generation using the keyword lists produced by each network structure. The results confirm that ConvNeXt generates the highest-quality radiology reports, further validating its suitability for the task. Additionally, sensitivity and specificity in multi-label classification are shown to significantly impact report generation performance. Accurate prediction of keywords (high sensitivity) and minimizing incorrect predictions (high specificity) are essential for generating high-quality reports. When incorrect or insufficient keywords are input into the language model, the generated reports are of lower quality.

To optimize the framework, it is crucial to determine whether low sensitivity (fewer correct keywords) or low specificity (more incorrect keywords) has a greater influence on report generation quality. This distinction can guide prioritization in pipeline optimization. Further investigation is needed to fully address this question, but our findings emphasize the importance of achieving a balance between these factors to ensure reliable and accurate radiology report generation.

Table 10. Performance comparison of different networks for multi-label classification on the IU X-Ray dataset. Results are reported as mean \pm standard deviation with precision up to four decimal places. Additionally, the average performance across all frequency clusters is calculated and presented as a mean value for reference. The bolded "ConvNeXt" represents the network configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods

Network	Frequency Cluster	Accuracy	Sensitivity	Specificity	F1-Score	MCC
ConvNeXt	[10,100]	0.9967 \pm 0.0045	0.9078 \pm 0.2029	0.9974 \pm 0.0039	0.8490 \pm 0.2064	0.8562 \pm 0.2003
ConvNeXt	[100,1000]	0.9917 \pm 0.0112	0.9220 \pm 0.2188	0.9939 \pm 0.0077	0.8572 \pm 0.2031	0.8625 \pm 0.1987
ConvNeXt	[1000,10000]	0.9108 \pm 0.0721	0.9514 \pm 0.1031	0.8889 \pm 0.0715	0.8811 \pm 0.1016	0.8181 \pm 0.1508
ConvNeXt	Average Performance	0.9664	0.9270	0.9601	0.8624	0.8456
ResNeXt	[10,100]	0.9888 \pm 0.0116	0.5877 \pm 0.4604	0.9927 \pm 0.0106	0.4760 \pm 0.3991	0.4888 \pm 0.4038
ResNeXt	[100,1000]	0.9672 \pm 0.0282	0.3678 \pm 0.4118	0.9905 \pm 0.0156	0.3570 \pm 0.3837	0.3584 \pm 0.3906
ResNeXt	[1000,10000]	0.7862 \pm 0.1696	0.9692 \pm 0.1033	0.6866 \pm 0.2400	0.7706 \pm 0.1635	0.6395 \pm 0.2551
ResNeXt	Average Performance	0.9141	0.6416	0.8899	0.5345	0.4956
ResNet	[10,100]	0.7586 \pm 0.0063	0.4426 \pm 0.3205	0.7617 \pm 0.0032	0.0343 \pm 0.0268	0.0463 \pm 0.0717
ResNet	[100,1000]	0.5873 \pm 0.0147	0.7311 \pm 0.2288	0.5830 \pm 0.0087	0.0963 \pm 0.0575	0.1045 \pm 0.0757
ResNet	[1000,10000]	0.6204 \pm 0.0675	0.8748 \pm 0.0811	0.4822 \pm 0.0525	0.6186 \pm 0.0876	0.3545 \pm 0.1234
ResNet	Average Performance	0.6554	0.6828	0.6090	0.2497	0.1684
VGG16	[10,100]	0.7498 \pm 0.0273	0.6285 \pm 0.3044	0.7513 \pm 0.0271	0.0448 \pm 0.0245	0.0815 \pm 0.0643
VGG16	[100,1000]	0.5745 \pm 0.0696	0.8921 \pm 0.0767	0.3993 \pm 0.0493	0.5967 \pm 0.0866	0.3041 \pm 0.1249
VGG16	[1000,10000]	0.5745 \pm 0.0696	0.8921 \pm 0.0767	0.3993 \pm 0.0493	0.5967 \pm 0.0866	0.3041 \pm 0.1249
VGG16	Average Performance	0.6330	0.7564	0.5734	0.2455	0.1631
EfficientNet	[10,100]	0.5771 \pm 0.0208	0.7590 \pm 0.2982	0.5713 \pm 0.0122	0.1089 \pm 0.0728	0.1141 \pm 0.1001
EfficientNet	[100,1000]	0.7907 \pm 0.0644	0.9531 \pm 0.1558	0.7851 \pm 0.0674	0.2139 \pm 0.1019	0.2903 \pm 0.1016
EfficientNet	[1000,10000]	0.4775 \pm 0.0741	0.9183 \pm 0.0689	0.2308 \pm 0.0422	0.5543 \pm 0.0855	0.1860 \pm 0.1361
EfficientNet	Average Performance	0.6151	0.8768	0.5291	0.2923	0.1968
NASNet	[10,100]	0.7099 \pm 0.0911	0.6785 \pm 0.1437	0.7408 \pm 0.0879	0.6160 \pm 0.1377	0.3994 \pm 0.1984
NASNet	[100,1000]	0.6897 \pm 0.0882	0.7980 \pm 0.2429	0.6868 \pm 0.0920	0.1314 \pm 0.0661	0.1674 \pm 0.0881
NASNet	[1000,10000]	0.7346 \pm 0.0752	0.6249 \pm 0.1454	0.8022 \pm 0.0796	0.6214 \pm 0.1170	0.4276 \pm 0.1673
NASNet	Average Performance	0.7114	0.7005	0.7432	0.4562	0.3314
Res2Net	[10,100]	0.7280 \pm 0.0679	0.6667 \pm 0.1226	0.7671 \pm 0.0722	0.6318 \pm 0.1025	0.4240 \pm 0.1478
Res2Net	[100,1000]	0.6014 \pm 0.0210	0.7260 \pm 0.3019	0.5978 \pm 0.0127	0.1100 \pm 0.0763	0.1124 \pm 0.1031
Res2Net	[1000,10000]	0.7096 \pm 0.0652	0.6727 \pm 0.0974	0.7366 \pm 0.0608	0.6197 \pm 0.0980	0.3949 \pm 0.1409
Res2Net	Average Performance	0.6797	0.6885	0.7005	0.4538	0.3104

Table 11. Performance comparison of different networks for multi-label classification on the MIMIC-CXR dataset. Results are reported as mean±standard deviation with precision up to four decimal places. Additionally, the average performance across all frequency clusters is calculated and presented as a mean value for reference. The bolded "ConvNeXt" represents the network configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.

Network	Frequency Cluster	Accuracy	Sensitivity	Specificity	F1-Score	MCC
ConvNeXt	[10,100]	0.9920±0.0042	0.7745±0.2510	0.9943±0.0039	0.6166±0.2146	0.6379±0.2068
ConvNeXt	[100,1000]	0.9896±0.0067	0.4036±0.4123	0.9943±0.0052	0.3303±0.3292	0.3449±0.3422
ConvNeXt	[1000,10000]	0.9450±0.0252	0.4598±0.2993	0.9616±0.0214	0.3281±0.2044	0.3253±0.2181
ConvNeXt	[10000,100000]	0.7349±0.0716	0.5293±0.2120	0.7713±0.0748	0.3537±0.1437	0.2333±0.1669
ConvNeXt	[100000+]	0.6764±0.2196	0.8142±0.3013	0.5294±0.3905	0.6812±0.2714	0.2743±0.4250
ConvNeXt	Average Performance	0.8676	0.5963	0.8502	0.4620	0.3631
ResNeXt	[10,100]	0.9270±0.0422	0.4005±0.4365	0.9563±0.0321	0.3257±0.3518	0.3062±0.3770
ResNeXt	[100,1000]	0.9199±0.0338	0.2404±0.2903	0.9611±0.0264	0.2237±0.2489	0.1964±0.2681
ResNeXt	[1000,10000]	0.9326±0.0267	0.2467±0.3512	0.9615±0.0183	0.1910±0.2498	0.1683±0.2712
ResNeXt	[10000,100000]	0.8793±0.1537	0.3772±0.3453	0.9038±0.1675	0.2484±0.2244	0.2429±0.2462
ResNeXt	[100000+]	0.9327±0.0276	0.4141±0.3035	0.9628±0.0179	0.3651±0.2403	0.3424±0.2540
ResNeXt	Average Performance	0.9183	0.3358	0.9491	0.2708	0.2512
ResNet	[10,100]	0.9786±0.0189	0.0320±0.1519	0.9878±0.0180	0.0161±0.0750	0.0122±0.0844
ResNet	[100,1000]	0.5510±0.2712	0.5849±0.4168	0.5511±0.2893	0.0865±0.0866	0.0492±0.1508
ResNet	[1000,10000]	0.5207±0.1976	0.7319±0.2175	0.4072±0.2698	0.5238±0.1811	0.1303±0.3690
ResNet	[10000,100000]	0.5226±0.1423	0.6777±0.2385	0.4372±0.2156	0.4928±0.1664	0.1185±0.2687
ResNet	[100000+]	0.5812±0.2338	0.9810±0.1054	0.1600±0.3175	0.6873±0.2082	0.0639±0.2118
ResNet	Average Performance	0.6308	0.6015	0.5087	0.3613	0.0748
VGG16	[10,100]	0.9474±0.0284	0.0911±0.2520	0.9931±0.0165	0.0885±0.2378	0.0857±0.2442
VGG16	[100,1000]	0.9243±0.0371	0.3762±0.3910	0.9687±0.0290	0.3370±0.3307	0.3161±0.3434
VGG16	[1000,10000]	0.9724±0.0249	0.3575±0.2386	0.9961±0.0065	0.4750±0.2981	0.4996±0.3201
VGG16	[10000,100000]	0.9595±0.0247	0.3820±0.2395	0.9933±0.0112	0.4911±0.2533	0.5180±0.2709
VGG16	[100000+]	0.9475±0.0244	0.4195±0.3397	0.9691±0.0170	0.3450±0.2549	0.3361±0.2727
VGG16	Average Performance	0.9502	0.3253	0.9841	0.3473	0.3511
EfficientNet	[10,100]	0.3351±0.2945	0.6354±0.4489	0.3321±0.2998	0.0197±0.0229	0.0088±0.0974
EfficientNet	[100,1000]	0.5481±0.0810	0.4881±0.4030	0.5504±0.0831	0.0702±0.0649	0.0133±0.1354
EfficientNet	[1000,10000]	0.3764±0.1092	0.9632±0.1539	0.0420±0.1416	0.5149±0.1270	0.0087±0.0946
EfficientNet	[10000,100000]	0.7566±0.1122	0.6925±0.2181	0.7976±0.1102	0.6547±0.1879	0.4803±0.2577
EfficientNet	[100000+]	0.8942±0.0369	0.7125±0.3557	0.9052±0.0287	0.3846±0.2205	0.3934±0.2420
EfficientNet	Average Performance	0.5821	0.6983	0.5255	0.3288	0.1774
NASNet	[10,100]	0.9006±0.0781	0.1789±0.3504	0.9074±0.0794	0.0314±0.0743	0.0293±0.1094
NASNet	[100,1000]	0.5848±0.2759	0.4076±0.4199	0.5920±0.2942	0.0757±0.1251	0.0103±0.1804
NASNet	[1000,10000]	0.5840±0.1277	0.6231±0.2059	0.5634±0.1472	0.5081±0.1691	0.1808±0.2629
NASNet	[10000,100000]	0.7117±0.1012	0.6415±0.1834	0.7633±0.1010	0.6021±0.1556	0.3918±0.2261
NASNet	[100000+]	0.9020±0.0254	0.7494±0.2998	0.9090±0.0220	0.3465±0.1603	0.3758±0.1752
NASNet	Average Performance	0.7366	0.5201	0.7470	0.3128	0.1976
Res2Net	[10,100]	0.9360±0.0465	0.1200±0.2959	0.9440±0.0478	0.0235±0.0596	0.0206±0.0914
Res2Net	[100,1000]	0.6378±0.3871	0.3514±0.4504	0.6481±0.4158	0.0449±0.0972	0.0434±0.1590
Res2Net	[1000,10000]	0.5090±0.1638	0.6685±0.2054	0.4207±0.2780	0.4880±0.1484	0.0735±0.2964
Res2Net	[10000,100000]	0.6024±0.1261	0.8840±0.1906	0.4416±0.1795	0.6039±0.1540	0.3430±0.2169
Res2Net	[100000+]	0.9474±0.0368	0.6511±0.3107	0.9729±0.0255	0.6295±0.2728	0.6187±0.2852
Res2Net	Average Performance	0.7265	0.5350	0.6854	0.3580	0.2111

Table 12. Performance comparison of different networks for radiology report generation on the IU X-Ray and MIMIC-CXR datasets. Results are presented as mean \pm standard deviation with precision up to four decimal places. The bolded "ConvNeXt" denotes the network configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.

Setting	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
IU X-Ray Test set							
ConvNeXt	0.7190 \pm 0.2101	0.6250 \pm 0.2713	0.5645 \pm 0.3057	0.5215 \pm 0.3376	0.6392 \pm 0.2554	0.3861 \pm 0.2218	3.2749 \pm 3.0106
ResNeXt	0.4802 \pm 0.1282	0.4299 \pm 0.1599	0.3965 \pm 0.1831	0.3718 \pm 0.2032	0.4484 \pm 0.1561	0.2732 \pm 0.1366	2.4615 \pm 1.9166
ResNet	0.2453 \pm 0.0655	0.2196 \pm 0.0817	0.2025 \pm 0.0935	0.1899 \pm 0.1038	0.2291 \pm 0.0698	0.1395 \pm 0.0797	1.2574 \pm 0.9790
VGG16	0.3170 \pm 0.0846	0.2838 \pm 0.1055	0.2617 \pm 0.1209	0.2454 \pm 0.1341	0.2960 \pm 0.0902	0.1803 \pm 0.1030	1.6250 \pm 1.2653
EfficientNet	0.1803 \pm 0.0481	0.1614 \pm 0.0600	0.1488 \pm 0.0687	0.1396 \pm 0.0763	0.1684 \pm 0.0513	0.1026 \pm 0.0586	0.9241 \pm 0.7196
NASNet	0.2332 \pm 0.0622	0.2088 \pm 0.0776	0.1925 \pm 0.0889	0.1805 \pm 0.0986	0.2177 \pm 0.0663	0.1326 \pm 0.0758	1.1952 \pm 0.9306
Res2Net	0.2144 \pm 0.0572	0.1920 \pm 0.0714	0.1770 \pm 0.0817	0.1660 \pm 0.0907	0.2002 \pm 0.0610	0.1220 \pm 0.0697	1.0990 \pm 0.8557
MIMIC-CXR Test set							
ConvNeXt	0.5599 \pm 0.1607	0.4379 \pm 0.1736	0.3557 \pm 0.1824	0.2953 \pm 0.1958	0.4699 \pm 0.1687	0.2842 \pm 0.1018	1.9964 \pm 1.4518
ResNeXt	0.3557 \pm 0.1001	0.2775 \pm 0.1081	0.2247 \pm 0.1133	0.1860 \pm 0.1209	0.2945 \pm 0.1044	0.1795 \pm 0.0629	0.6221 \pm 0.9035
ResNet	0.3415 \pm 0.0961	0.2664 \pm 0.1038	0.2157 \pm 0.1088	0.1785 \pm 0.1161	0.2827 \pm 0.1002	0.1724 \pm 0.0604	0.5972 \pm 0.8673
VGG16	0.3572 \pm 0.1006	0.2787 \pm 0.1086	0.2256 \pm 0.1138	0.1868 \pm 0.1214	0.2957 \pm 0.1048	0.1803 \pm 0.0632	0.6247 \pm 0.9073
EfficientNet	0.4096 \pm 0.1153	0.3195 \pm 0.1245	0.2587 \pm 0.1304	0.2141 \pm 0.1392	0.3391 \pm 0.1202	0.2067 \pm 0.0725	0.7163 \pm 1.0403
NASNet	0.4336 \pm 0.1221	0.3383 \pm 0.1318	0.2739 \pm 0.1381	0.2267 \pm 0.1474	0.3590 \pm 0.1272	0.2189 \pm 0.0767	0.7584 \pm 1.1014
Res2Net	0.4093 \pm 0.1152	0.3193 \pm 0.1244	0.2585 \pm 0.1303	0.2140 \pm 0.1391	0.3388 \pm 0.1201	0.2066 \pm 0.0724	0.7157 \pm 1.0395

4.7.2 Influence of Keyword Numbers and the Combination of High- and Low-Frequency Keywords on Radiology Report Generation Performance

The automatic keyword adaptation mechanism enables the division of keywords into frequency clusters, ranging from low to high. When integrating multi-label classification and radiology report generation, the performance of low-frequency keywords can significantly influence the quality of the generated reports, as missing critical information or introducing incorrect keywords may degrade the results. To analyze this effect, we examined the impact of frequency-based multi-label classification on both high- and low-frequency keywords, as well as their connection to radiology report generation.

To this end, we designed experiments that selectively activate specific keyword clusters and generate radiology reports using only the corresponding keywords. First, we validated the generated reports using single clusters. Then, we extended the experiments to mixed clusters by combining specific frequency ranges to generate the corresponding

reports. During these experiments, thresholds for keyword list fusion may vary based on the number of active keyword clusters, and the final keyword list may differ from the full version generated by our multi-label classification. Consequently, some keywords unique to this ablation study may appear in the generated reports.

We used the same training settings as in the original experiments and leveraged pre-trained language model checkpoints to reduce computational cost. This ensured that the radiology report generation process remains comparable with state-of-the-art (SOTA) methods. Given the larger number of clusters in the MIMIC-CXR dataset compared to the IU X-Ray dataset, we simplified the cluster combinations for MIMIC-CXR, treating [10,100], [100,1000], and [1000,10,000] as the low-frequency cluster for consistency with IU X-Ray.

The evaluation results for different clusters are presented in Table 13 (IU X-Ray) and Table 14 (MIMIC-CXR), with visualized examples of the generated reports and their corresponding keywords shown in Figure 29 (IU X-Ray) and Figure 30 (MIMIC-CXR).

The results revealed distinct trends between the datasets. For the IU X-Ray dataset with three clusters, the general observation was that high-frequency keywords yielded better radiology report accuracy compared to low-frequency keywords, even when the latter included more keywords. Specifically, although the clusters [10,100] and [100,1000] contained more keywords than [1000,10,000], the latter achieved higher performance, likely due to the higher quality of multi-label classification in the high-frequency cluster. Visualization of the generated reports confirmed that high-frequency clusters included more relevant information, resulting in superior report quality. This trend extended to mixed clusters, where combining multiple clusters generally improved performance compared to individual clusters.

However, this pattern did not hold for the MIMIC-CXR dataset. In this case, the highest-frequency cluster [100,000+] demonstrated lower performance. Analysis of the reports generated by this cluster revealed that it contained only five keywords, making it challenging to produce high-quality reports with such limited information. These findings indicated that while high-frequency keywords provided general information, low-frequency keywords were essential for capturing rare but clinically significant details. The mixed-cluster results for MIMIC-CXR supported this modified rule: although the individual performance of [10,100], [100,1000], and [1000,10,000] was lower, their combined performance was competitive, and further improvement was observed when including the [100,000+] cluster.

This analysis highlighted a key optimization insight: while individual low-frequency clusters performed poorly, combining them with high-frequency clusters significantly enhanced overall performance. This suggested a potential avenue for improving multi-label classification by balancing high-frequency and low-frequency keywords to maximize the quality of radiology report generation.

Table 13. Performance comparison of radiology report generation on the IU X-Ray dataset across different keyword clusters. Results are reported as mean \pm standard deviation with precision up to four decimal places. The bolded "Cluster All Keywords" represents the configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.

Setting	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
IU X-Ray Test							
Single Cluster							
Cluster [10,100]	0.0038 \pm 0.0552	0.0016 \pm 0.0300	0.0010 \pm 0.0189	0.0006 \pm 0.0110	0.0702 \pm 0.0595	0.0238 \pm 0.0310	0.0141 \pm 0.0853
Cluster [100,1000]	0.1555 \pm 0.1520	0.0989 \pm 0.1151	0.0712 \pm 0.0974	0.0532 \pm 0.0852	0.1859 \pm 0.1262	0.1008 \pm 0.0765	0.2240 \pm 0.4427
Cluster [1000,10000]	0.4443 \pm 0.1833	0.3149 \pm 0.1824	0.2407 \pm 0.1889	0.1913 \pm 0.1974	0.3645 \pm 0.1799	0.2162 \pm 0.1207	0.6171 \pm 1.0652
Mixed Cluster							
Cluster [10,100]+[100,1000]	0.2083 \pm 0.1654	0.1337 \pm 0.1284	0.0973 \pm 0.1081	0.0734 \pm 0.0936	0.2108 \pm 0.1348	0.1146 \pm 0.0791	0.3096 \pm 0.4984
Cluster [10,100]+[1000,10000]	0.4568 \pm 0.1799	0.3207 \pm 0.1798	0.2402 \pm 0.1833	0.1857 \pm 0.1881	0.3686 \pm 0.1669	0.2161 \pm 0.1131	0.6073 \pm 1.0344
Cluster [100,1000]+[1000,10000]	0.5013 \pm 0.2028	0.3828 \pm 0.2381	0.3140 \pm 0.2633	0.2680 \pm 0.2811	0.4331 \pm 0.2354	0.2729 \pm 0.1675	1.2750 \pm 2.1064
Cluster All Keywords	0.7190\pm0.2101	0.6250\pm0.2713	0.5645\pm0.3057	0.5215\pm0.3376	0.6392\pm0.2554	0.3861\pm0.2218	3.2749\pm3.0106

Figure 29. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the IU X-Ray dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue.




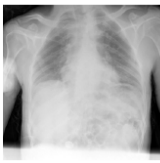
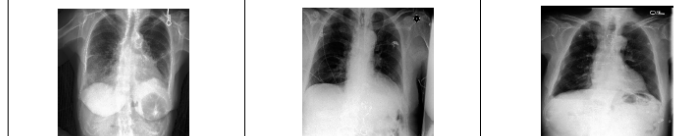
X-Ray Image				
Reference Report	The heart size is normal. The mediastinal contour is within normal limits. The lungs are free of any focal infiltrates. There are no nodules or masses. No visible pneumothorax. No visible pleural fluid. The XXXX are grossly normal. There is no visible free intraperitoneal air under the diaphragm. Surgical clips are seen in the right upper quadrant.	The cardiomeastinal silhouette is normal size and configuration. Pulmonary vasculature within normal limits. The lungs are well-aerated. There is no pneumothorax, pleural effusion, or focal consolidation.	The heart is normal in size. The mediastinal contours are stable. Aortic calcifications are noted. There are small calcified lymph XXXX. Emphysema and chronic changes are identified. There is XXXX opacity in the left perihilar upper lobe. There is a questionable XXXX extension to the pleural surface. This may represent acute infiltrate or developing density. There is no pleural effusion or pneumothorax.	AP view was obtained due to patient condition. Low volume lungs. No focal lung consolidation. The heart is not enlarged. No pleural effusion.
All predicted Keywords and generated report	<i>intraperitoneal, heart, contour, pneumothorax, fluid, quadrant, lungs, focal, no, upper, normal, diaphragm</i> 1. No active disease. The heart size is normal. The mediastinal contour is within normal limits. The lungs are free of any focal infiltrates. There are no nodules or masses. No visible pneumothorax. No visible pleural fluid. The XXXX are grossly normal. There is no visible free intraperitoneal air under the diaphragm.	<i>aerated, pneumothorax, consolidation, vasculature, lungs, focal, normal, effusion</i> The cardiomeastinal silhouette is normal size and configuration. Pulmonary vasculature within normal limits. The lungs are well-aerated. There is no pneumothorax, pleural effusion, or focal consolidation.	<i>heart, infiltrate, surface, effusion, chronic, calcifications, small, perihilar, calcified, stable, lobe, emphysema, pneumothorax, opacity, no, normal, left, extension, acute</i> 1. No acute cardiopulmonary disease. 2. Stable emphysema. The heart is normal in size. The mediastinal contours are stable. There is a small calcified perihilar opacity in the left lower lobe. There is no acute infiltrate or effusion. There is no pneumothorax.	<i>patient, heart, low, consolidation, enlarged, lungs, focal, no, lung, volume, effusion</i> 1. Low volume study without evidence of active disease. The heart is not significantly enlarged. The lungs are free of focal consolidation. No pleural effusion.
Keyword and generated report in Cluster [10,100]	<i>intraperitoneal, quadrant</i> Cardiomegaly with intraperitoneal effusions. Calcified right hilar lymph node and left midlung infiltrate/abdominal contours.	<i>aerated</i> Aerated and aerated. The cardiac silhouette is well-aerated.	<i>perihilar</i> The cardiac silhouette and cardiomeastinal silhouette are within normal limits in size. The perihilar contours are within normal limits in size.	<i>volume</i> Volume loss with no visible active cardiopulmonary disease. Volume loss with mild volume loss.
Keyword and generated report in Cluster [100,1000]	<i>upper, diaphragm, fluid</i> 1. Right upper diaphragm fluid in the right upper lobe. 2. No evidence of active disease. The cardiomeastinal silhouette and pulmonary vasculature are within normal limits. There is a right upper diaphragm fluid in the right upper lobe.	<i>vasculature</i> Vasculature within normal limits for size and configuration of the cardiac silhouette. Vasculature within normal limits for size and configuration of the cardiac silhouette.	<i>calcifications, opacity, left, stable, emphysema, chronic, calcified, small, lobe, infiltrate</i> 1. Stable emphysema with streaky left lower lobe infiltrate. 2. Stable calcifications of aortic XXXX. The cardiac silhouette and mediastinal contours appear within normal limits. There is a small calcified granuloma seen in the left lower lobe. There is stable emphysema. There are streaky calcifications of aortic XXXX.	<i>enlarged, patient, low</i> Cardiomegaly with enlarged cardiac silhouette. Low patient volume study.
Keyword and generated report in Cluster [1000,10000]	<i>focal, normal, pneumothorax, no, heart, lungs, contour</i> 1. No evidence of active disease. The cardiomeastinal silhouette is within normal limits for size and configuration. The lungs are free of focal infiltrates. No pleural effusions or pneumothorax.	<i>focal, effusion, normal, pneumothorax, lungs, consolidation</i> No evidence of active disease. The lungs are free of focal consolidation, pleural effusion, or pneumothorax. Cardiomeastinal silhouette is within normal limits. Pulmonary vasculature is within normal limits.	<i>effusion, normal, pneumothorax, no, acute, heart</i> No acute cardiopulmonary abnormality. Heart size within normal limits. No pneumothorax or pleural effusion.	<i>focal, effusion, no, heart, lungs, lung, consolidation</i> No evidence of active disease. The lungs and pleural spaces show no focal infiltrates. There is no pleural effusion or pneumothorax. The cardiac silhouette and pulmonary vasculature appear within normal limits. The lungs are free of focal consolidation.
Keyword and generated report in Cluster [10,100]+[100,1000]	<i>upper, diaphragm, fluid, intraperitoneal, quadrant</i> 1. Right upper diaphragm XXXX with streaky infiltrates. The cardiomeastinal silhouette and pulmonary vasculature are within normal limits. There is a right upper diaphragm XXXX with streaky infiltrates and a XXXX pleural effusions. There is no pneumothorax.	<i>vasculature, aerated</i> Cardiomegaly and pulmonary vasculature are well-aerated. The cardiomeastinal silhouette is well-aerated. Pulmonary vasculature is within normal limits in size.	<i>calcifications, opacity, left, stable, emphysema, chronic, calcified, small, lobe, infiltrate, perihilar</i> Stable emphysema with streaky left lower lobe infiltrate and calcified granulomas. There is stable XXXX opacity in the left lower lobe which may represent emphysema or infiltrate. The cardiac silhouette, pulmonary vasculature, and mediastinal contours are within normal limits. There are streaky calcifications in the left lower lobe and left lower lobe.	<i>enlarged, patient, low, volume</i> Cardiomegaly with enlarged cardiomeastinal silhouette. Lung volumes are mildly enlarged. Patient is slightly enlarged.
Keyword and generated report in Cluster [10,100]+[1000,10000]	<i>focal, normal, pneumothorax, no, heart, lungs, contour, intraperitoneal, quadrant</i> 1. No evidence of active disease. The cardiomeastinal silhouette is within normal limits for size and configuration. The lungs are free of focal infiltrates. No pleural effusions or pneumothorax. There is no evidence of active TB. There is an intraperitoneal XXXX stenosis XXXX.	<i>focal, effusion, normal, pneumothorax, lungs, consolidation, aerated</i> No evidence of active disease. The cardiac silhouette is normal in size. The lungs are well-aerated. There is no focal consolidation, pneumothorax or large pleural effusion.	<i>effusion, normal, pneumothorax, no, acute, heart, perihilar</i> No acute cardiopulmonary abnormality. The aortic XXXX and mediastinal contours are within normal limits. Pulmonary vasculature is within normal limits. There is no pleural effusion or pneumothorax.	<i>focal, effusion, no, heart, lungs, lung, consolidation, volume</i> 1. No evidence of active disease. The cardiomeastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are free of focal consolidation. No pleural effusions or pneumothorax.
Keyword and generated report in Cluster [100,1000]+[1000,10000]	<i>focal, normal, pneumothorax, no, heart, lungs, contour, upper, diaphragm, fluid</i> 1. No evidence of active disease. The cardiac silhouette and pulmonary vasculature appear within normal limits. The lungs are free of focal infiltrates. There is no pneumothorax or pleural effusions. There is a diaphragm in the right upper lobe. There is no evidence of active TB.	<i>focal, effusion, normal, pneumothorax, lungs, consolidation, vasculature</i> 1. No evidence of active disease. The cardiomeastinal silhouette and pulmonary vasculature are within normal limits. The lungs are free of focal consolidation, pleural effusion, or pneumothorax. There is no evidence of active effusion.	<i>effusion, normal, pneumothorax, no, acute, heart, calcifications, opacity, left, stable, emphysema, chronic, calcified, small, lobe, infiltrate</i> 1. Stable emphysema. 2. Stable cardiomegaly. The heart size is within normal limits. There is a stable calcified right lower lobe infiltrate compatible with emphysema. There is no pneumothorax or pleural effusion. There are calcifications of the left lower lobe consistent with emphysema. There is a stable XXXX opacity in the right upper lobe.	<i>focal, effusion, no, heart, lungs, lung, consolidation, enlarged, patient, low</i> 1. Cardiomegaly and enlarged cardiomeastinal silhouette. Low lung volumes without focal consolidation, pleural effusion, or pneumothorax. The lungs are hyperexpanded with flattening of the right costophrenic XXXX blunting/effusion. There is no focal consolidation.

Table 14. Performance comparison of radiology report generation on the MIMIC-CXR dataset across different keyword clusters. Results are presented as mean \pm standard deviation with precision up to four decimal places. The bolded "Cluster All Keywords" represents the configuration used in our proposed pipeline, serving as a baseline for comparison with other state-of-the-art methods.

Setting	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
MIMIC-CXR Test							
Single Cluster							
Cluster [10,100]	0.0849 \pm 0.0877	0.0442 \pm 0.0532	0.0251 \pm 0.0382	0.0143 \pm 0.0258	0.1428 \pm 0.0589	0.0530 \pm 0.0318	0.0109 \pm 0.0474
Cluster [100,1000]	0.0280 \pm 0.0654	0.0159 \pm 0.0422	0.0098 \pm 0.0306	0.0060 \pm 0.0215	0.1138 \pm 0.0618	0.0416 \pm 0.0302	0.0075 \pm 0.0608
Cluster [1000,10000]	0.1912 \pm 0.1219	0.1110 \pm 0.0853	0.0676 \pm 0.0681	0.0439 \pm 0.0530	0.1841 \pm 0.0766	0.0948 \pm 0.0486	0.0733 \pm 0.1944
Cluster [10000,100000]	0.3442 \pm 0.1514	0.2260 \pm 0.1293	0.1573 \pm 0.1202	0.1133 \pm 0.1121	0.2841 \pm 0.1158	0.1727 \pm 0.0696	0.2703 \pm 0.6024
Cluster [100000+]	0.0403 \pm 0.0946	0.0256 \pm 0.0671	0.0177 \pm 0.0534	0.0127 \pm 0.0433	0.1699 \pm 0.0894	0.0618 \pm 0.0454	0.0116 \pm 0.0759
Mixed Cluster							
Cluster ([10,100]+[100,1000]+[1000,10000])	0.2516 \pm 0.1239	0.1438 \pm 0.0916	0.0878 \pm 0.0778	0.0566 \pm 0.0648	0.1950 \pm 0.0785	0.1046 \pm 0.0522	0.1124 \pm 0.3530
Cluster ([10,100]+[100,1000]+[1000,10000])+[10000,100000]	0.3825 \pm 0.1549	0.2527 \pm 0.1216	0.1763 \pm 0.1070	0.1281 \pm 0.0980	0.2670 \pm 0.1099	0.1712 \pm 0.0787	0.1695 \pm 0.4255
Cluster ([10,100]+[100,1000]+[1000,10000])+[100000+]	0.2811 \pm 0.1389	0.1690 \pm 0.1010	0.1075 \pm 0.0854	0.0702 \pm 0.0728	0.2171 \pm 0.0883	0.1174 \pm 0.0610	0.1206 \pm 0.3657
Cluster [10000,100000]+[100000+]	0.4053 \pm 0.1535	0.2778 \pm 0.1390	0.2007 \pm 0.1327	0.1495 \pm 0.1286	0.3172 \pm 0.1287	0.1923 \pm 0.0825	0.3079 \pm 0.6729
Cluster All Keywords	0.5599 \pm 0.1607	0.4379 \pm 0.1736	0.3557 \pm 0.1824	0.2953 \pm 0.1958	0.4699 \pm 0.1687	0.2842 \pm 0.1018	1.9964 \pm 1.4518

X-Ray Image		
Reference Report	<p>Diffuse interstitial opacities, predominantly in the right lung base and probably very mild in the left lung base are present. When compared to the prior chest CT from _____, these interstitial opacities appear new, and given the clinical history, likely represent lung infection (logical vital) or asymmetric pulmonary edema. In view of history, a possibility of lymphangitic carcinomatosis also needs to be ruled out. Compared with prior radiograph from _____, a dense opacity in the left perihilar region, likely radiation fibrosis has significantly improved over _____ to _____. As demonstrated on series of chest CT's, _____ Heart size is mildly larger and unchanged. Irregularity of the hilar and mediastinal border is likely postirradiation. Is-defined opacity in the left mid lung corresponds to scarring and atelectasis. As suggested on review of chest CT from _____, Pleural effusion, if any, is minimal bilaterally. Bilateral lower lung bases are remarkable for mild atelectasis.</p> <p>As compared to the preoperative radiograph, there is a minimal decrease in overall lung volumes. As a consequence, a small retrocardiac atelectasis is seen. However, there is no evidence of pneumonia. Borderline size of the cardiac silhouette. The presence of a minimal left pleural effusion cannot be excluded. Normal hilar and mediastinal contours.</p> <p>The lungs are well expanded and clear. The cardiomeastinal and hilar contours are normal. There is moderate aortic tortuosity. Unchanged. A small right-sided pleural effusion is unchanged. There is no pneumothorax. Chestotomy wires are intact. Multiple fractures in early stages of healing are noted in the right rib cage.</p>	<p>The pulmonary vasculature is presently not congested, and on previous portable examination identified edema pattern as well as evidence of right-sided pleural effusion has normalized. No new pulmonary opacities are seen. _____ Cardiac enlargement is identified in the apical area. Appear observed is a fractured second rib in the left apical area. Apparently the result of previous sternotomy and thoracotomy cardiac intervention. When comparison is extended to the next preceding PA and lateral chest examination of _____, the patient is in better condition now as the cardiac enlargement has regressed and the pleural effusion has been associated completely.</p>
All predicted Keywords and generated report	<p>mid, effusion, chest, interstitial, fibrosis, opacity, atelectasis, border, effusion, lower, minimal, unchanged, carcinomatosis, bilateral, lung, heart</p> <p>Frontal and lateral chest radiographs demonstrate diffuse bilateral interstitial opacities, which are new since _____. There is minimal opacity in the right lower lung, which may represent atelectasis or lymphangitic carcinomatosis. The cardiomeastinal silhouette is unchanged. Mild discoloration of the left heart border is likely due to a combination of pleural effusion and atelectasis.</p> <p>As compared to the previous radiograph, there is unchanged evidence of a right lung opacity with subsequent areas of atelectasis at both lung bases. The preoperative findings are constant. Minimal left pleural effusion. Normal size of the cardiac silhouette. No evidence of pneumonia, no other parenchymal changes.</p>	<p>chest, ascending, valve, contour, vasculature, metallic, atrium, rib, comparison, moderate, volume, effusion, fractured, angiotensin, pneumothorax, enlargement, intrathoracic, patient, apical</p> <p>AP single view of the chest has been obtained with patient in semi-upright position. Analysis is performed in direct comparison with the next preceding similar study of _____. Cardiac enlargement persists, but there is no evidence of significant intrathoracic metastatic disease. The pulmonary vasculature shows an upper zone redistribution pattern and there is no evidence of overt pulmonary edema. There is no evidence of new parenchymal infiltrates, pleural effusion, or pneumothorax.</p>
Keyword and generated report in Cluster [10,100]	<p>carcinomatosis</p> <p>The cardiac, mediastinal and hilar contours are unchanged since prior examination. There is no evidence of pleural effusion or pneumothorax.</p> <p>pneumo</p> <p>There is no evidence of pneumo, coiled in the aortopulmonary window. If clinically indicated, CT scan would be necessary to further evaluate this. No pneumo, cardiac, pleural effusions or pneumothoraces are seen.</p>	<p>thoracic, pneumo</p> <p>In comparison with the study of _____ there is no evidence of pneumo or pneumo in the thoracic cavity. Otherwise, little change except for slight improvement in the degree of left hemi thoracic pleurosis and possible small effusions.</p>
Keyword and generated report in Cluster [100,1000]	<p>carcinoma</p> <p>In comparison with the study of _____, there is no change in the appearance of the known right hilar and mediastinal carcinoma carcinoma</p> <p>pre,all,preoperative</p> <p>Preoperative changes. Preoperative size of the preoperative preoperative is all preoperative.</p> <p>healing,all</p> <p>All healing healing is appreciated.</p>	<p>metalic, ascending, fractured, intrathoracic,all</p> <p>All metallic metallic debridement fractures are fractured. Ascending metallic ascending thorax is unchanged. There is no evidence of intrathoracic or mediastinal mass.</p>
Keyword and generated report in Cluster [1000,10000]	<p>minimal, diffuse, border, fusion, fibrosis</p> <p>Diffuse minimal diffuse fibrosis is present. There is borderline cardiomegaly, but no evidence of pulmonary edema. The mediastinal border and cardiac border are within normal limits. Lumbar fusion hardware is incompletely assessed. Minimal fibrosis is noted.</p> <p>minimal, fusion</p> <p>Minimal cervical spinal fusion hardware.</p> <p>multiple, thorax, fusion</p> <p>Multiple thorax is present, with multiple surgical clips in the left hemi thorax. Cervical spinal fusion hardware is partially visualized.</p>	<p>apical, enlargement, valve, thorax, fusion</p> <p>There is marked enlargement of the cardiac silhouette with valve replacement and median sternotomy wire. The mediastinal and hilar contours are stable. Unchanged mild apical pulmonary edema and right apical pleural thickening. Post-operative changes in the right hemi thorax are again demonstrated as well as postsurgical changes in the right hemi thorax. Cervical spinal fusion hardware is partially visualized.</p>
Keyword and generated report in Cluster [10000,100000]	<p>chest, lung, heart, atelectasis, lateral, unchanged, opacity, mid, bilateral, low, interstitial, lower, change</p> <p>Compared to chest radiograph from earlier the same day. Lung volumes remain low. Heart size is unchanged. Mild bilateral interstitial lung interstitial prominence is unchanged. Unchanged heart size. Mild retrocardiac opacity is unchanged. (key atelectasis). Bilateral lower lung bases are unchanged. Mild lower lung atelectasis. Overall, no significant change.</p> <p>lung, atelectasis, left, pneumonia, small</p> <p>As compared to the previous radiograph, the lung volumes remain low. Small left and small right pleural effusions are present. Bilateral areas of atelectasis are present in the retrocardiac lung areas. Slightly atelectasis is present in both lung bases, but no evidence of pneumonia or aspiration.</p>	<p>chest, edema, enlarged, vasculature, patient, rib, moderate, large, contour, atrium, comparison, fracture</p> <p>Comparison is made for the patient's chest comparison. Cardiac contour is moderate to severely enlarged. The cardiac contour is moderately enlarged. The thorax is a slightly enlarged device with leads in the right atrium and ventricle as comparison is available. The study is limited due to the patient's advanced advanced significantly in comparison of the prior study. This includes a relatively right positioned fracture of the eighth rib which is fracture fragment in the chest cavity.</p>
Keyword and generated report in Cluster [100000,1000000]	<p>effusion, no</p> <p>As compared to the previous radiograph, there is no evidence of newly appeared parenchymal effusion. No other parenchymal abnormalities. No larger pleural effusions.</p> <p>effusion, no, normal</p> <p>The cardiac, mediastinal and hilar contours are normal. No effusion. No sinister bony effusion.</p> <p>pneumothorax, effusion, no, lungs</p> <p>The lungs are clear, the cardiomeastinal silhouette and hila are unremarkable. There is no pleural effusion or pneumothorax.</p>	<p>pneumothorax, effusion</p> <p>There is no focal consolidation, pleural effusion, pneumothorax, or pulmonary edema.</p>
Keyword and generated report in Cluster [1000000,10000000]	<p>minimal, diffuse, border, fusion, fibrosis, carcinoma, carcinomatosis</p> <p>Cardiac fusion, mediastinal and hilar contours are unchanged. The right border of the cardiac border is relatively more diffuse than on the previous study. This may reflect lymphangiomatosis, carcinoma carcinomatosis or carcinomatosis in this patient with history of lung carcinoma carcinomatosis. Diffuse fibrosis has developed in the left perihilar region as well as pulmonary fibrosis with known radiation therapy background fibrosis also seen in the right upper lobe compatible.</p> <p>minimal, fusion, pre, all, preoperative, pneumo</p> <p>All preoperative fusion is minimal. All preoperative pneumo, right or bronchial or overall lung volume loss and give preoperative changes are unchanged. All mediastinal clips are consistent with preoperative thyroid thyroid surgery.</p>	<p>apical, enlargement, valve, thorax, fusion, metallic, ascending, fractured, intrathoracic, all, thorax, pneumo, chest, edema, enlarged, vasculature, patient, rib, moderate, large, contour, atrium, comparison, fracture</p> <p>There is stable enlargement of the ascending aortic valve, as well as an aortic valve replacement. Cervical spinal fusion hardware is fractured and unchanged in position. All intrathoracic spinal fusion hardware is intact. There is no evidence of the ascending aortic contour suggestive presence of an aortic valve. There is also a moderate-to-large left atrium, unchanged in comparison with the previous study. Pulmonary vasculature is not congested and there is evidence of enlargement.</p>
Keyword and generated report in Cluster [10000000,100000000]	<p>minimal, diffuse, border, fusion, fibrosis, carcinoma, carcinomatosis, chest, lung, heart, atelectasis, lateral, unchanged, opacity, mid, bilateral, low, interstitial, lower, change</p> <p>Frontal and lateral views of the chest. Mild cardiomegaly is unchanged. Lung volumes are low. The mediastinal, subcarinal fusion breast carcinomatosis, lymphoma, lung carcinomatosis, and fibrosis remain consistent with patient's known lung carcinomatosis. Right lower lung fusion demonstrates diffuse interstitial fibrosis. In addition, there is minimal bilateral lower lung lung fibrosis. Unchanged right heart contour. Unchanged opacity in the left lung may also reflect.</p> <p>minimal, fusion, pre, all, preoperative, pneumo, lung, atelectasis, left, pneumonia, small</p> <p>Preoperative pneumo. All preoperative pneumo is minimal. Preoperative pneumo and fusion of the left lung is now small. Left lung lung preoperative chest tube is in preoperative right lung lung lung, mild cardiomegaly, stable postoperative mediastinal cyst cyst. All preoperative pneumo has been present for all preoperative pneumo. All lung lung lung volumes are low. Small left retrocardiac atelectasis is likely due to atelectasis aspiration or pneumonia.</p>	<p>apical, enlargement, valve, thorax, fusion, metallic, ascending, fractured, intrathoracic, all, thorax, pneumo, chest, edema, enlarged, vasculature, patient, rib, moderate, large, contour, atrium, comparison, fracture</p> <p>PA and lateral chest views were obtained with patient in upright position. Analysis is performed in direct comparison with the next preceding similar study of _____. Cardiac enlargement is moderately enlarged. Mild degree of intrathoracic enlargement as well as moderate-to-large left atrium, unchanged in position of an aortic valve. There is also a moderate-to-large left atrium, unchanged in comparison with the previous study. Pulmonary vasculature is not congested and there is evidence of enlargement.</p>
Keyword and generated report in Cluster [100000000,1000000000]	<p>minimal, diffuse, border, fusion, fibrosis, carcinoma, carcinomatosis, effusion, no</p> <p>Cardiac fibrosis, lymphomatous carcinoma carcinomatosis in the right hilar and effusion fibrosis is consistent with known lung carcinoma carcinomatosis. The right border of the cardiac border is minimally decreased. Cardiac enlargement is accompanied by diffuse fibrosis fibrosis. No pleural effusion. No evidence of pleural effusion.</p> <p>minimal, fusion, pre, all, preoperative, pneumo, effusion, no, normal</p> <p>Preoperative pneumo preoperative pneumo. All preoperative fusion is in preoperative position. There is minimal cardiac preoperative pneumo, all in the preoperative right pancreocretion. No pleural effusion. Normal mediastinal and cardiac contours.</p>	<p>apical, enlargement, valve, thorax, fusion, metallic, ascending, fractured, intrathoracic, all, thorax, pneumo, chest, edema, enlarged, vasculature, patient, rib, moderate, large, contour, atrium, comparison, fracture</p> <p>There is stable enlargement of the ascending and descending thoracic aortic valve. The right apical pneumothorax is not significantly changed from prior examination. There is no evidence of focal consolidation, pleural effusion, pneumo, or pneumothorax. All the three fractured thoracic vertebrae are unchanged in position. Postsurgical hardware within the thoracic cavity is unchanged. Multi-level thoracic vertebral bodies appear fractured.</p>
Keyword and generated report in Cluster [10000,100000,1000000]	<p>effusion, no, chest, lung, heart, atelectasis, lateral, unchanged, opacity, mid, bilateral, low, interstitial, lower, change</p> <p>AP and lateral views of the chest. Lung volumes are low. Heart size is unchanged. Mild bilateral interstitial prominence is unchanged. There is a mild lower lung lung opacity. Retrocardiac opacity likely reflects atelectasis. Overall, there has been no significant change. Bilateral atelectasis remains unchanged. Mild heart size is unchanged.</p> <p>effusion, no, normal, lung, atelectasis, left, pneumonia, small</p> <p>The lung volumes are normal. Normal size of the cardiac silhouette. Small left pleural effusion. No evidence of pneumonia. Left retrocardiac atelectasis is small.</p> <p>pneumothorax, effusion, no, lungs, clear, lung, unremarkable, unchanged, right, intact, small, rib, moderate, expanded, change</p> <p>Unremarkable cardiomeastinal. The lungs are well expanded and clear. Moderate cardiomegaly is unchanged. Small right pleural effusion is unchanged. There is no pneumothorax. Small ribs are intact. No significant change from the prior examination. Multiple fractures in early stages of healing are noted. The lungs are well expanded and clear. Right lung is unremarkable.</p>	<p>pneumothorax, effusion, chest, edema, enlarged, vasculature, patient, rib, moderate, large, contour, atrium, comparison, fracture</p> <p>AP single view of the chest has been obtained with patient in comparison with the next preceding study of _____. Comparison is made with the previous examination of a chest examination of _____. The cardiac silhouette is moderately enlarged, but stable in comparison with the next preceding present present present present. In comparison with the current study of the patient's known fracture hardware though central pulmonary vasculature is compatible with edem</p>

[illegible]

Figure 30. Visualization of results generated by the proposed framework compared to two state-of-the-art methods using their official checkpoints on the MIMIC-CXR dataset. In the reference reports and our generated reports, keywords predicted by our multi-label classification are highlighted in blue.

4.7.3 Performance of Radiology Report Generation with Different Text-to-text

Large Language Models and Pretrained Materials

Large language models (LLMs), particularly Transformer-based architectures, are highly effective in natural language processing tasks, including text generation, AI-based chatting, and context-specific text creation. Given their robust capability to process spatial and semantic information, these models hold significant potential for generating radiology reports based on image-derived keywords.

However, balancing computational cost and performance is a critical challenge when applying LLMs to radiology report generation. In medical settings, the availability of large-scale computational resources, such as GPU clusters, is often limited, making it impractical to train LLMs from scratch. Consequently, identifying appropriately sized LLMs and fine-tuning them using limited domain-specific data becomes essential to our pipeline.

In our experiments, we employed a modified Text-to-Text Transformer (T5) model ([56]), a medium-sized LLM, and adopted the pretrained configuration of Clinical-T5 ([57]) to initialize the model. To evaluate the impact of different text-to-text LLM versions and pretrained settings, we compared the performance of the original T5 model with its advanced variant Flan-T5 ([71]), as well as with larger text-to-text LLMs, including BART ([72]) and Pegasus ([73]), provided by Microsoft and Google.

For pretrained materials, we tested two configurations: (1) checkpoints trained on general language datasets provided by HuggingFace, and (2) checkpoints further pretrained on medical domain materials. These pretrained checkpoints, sourced from official repositories or re-implementations, allowed us to assess text-to-text LLM

performance without the need for training from scratch. The evaluation results are summarized in Table 15 (IU X-Ray) and Table 16 (MIMIC-CXR).

The results indicate that further pretraining on medical materials consistently enhances performance compared to models trained only on general language datasets.

Additionally, the performance differences among various T5 model versions, including Flan-T5, were minimal. This suggests that generating radiology reports based on keyword inputs is not a particularly complex task for text-to-text LLMs, and their general architecture is sufficient to handle it effectively. The larger models, such as BART and Pegasus, did not exhibit a significant advantage in this task, highlighting the suitability of medium-sized models like T5 for this application.

These findings underscore the potential for deploying text-to-text LLMs tailored to specific computational and performance requirements. In scenarios demanding high performance, more complex models may be utilized, while in resource-constrained environments, less complex models can achieve satisfactory results with minimal performance degradation. This flexibility makes text-to-text LLMs a practical and scalable choice for diverse medical applications, balancing computational efficiency with clinical effectiveness.

Table 15. Performance comparison of radiology report generation on the IU X-Ray dataset across different text-to-text large language model versions and their pretraining materials. Results are reported as mean \pm standard deviation with precision up to four decimal places. Additionally, the trainable model parameters are provided to indicate the complexity of each language model. The suffixes "-base" and "-small" denote the model size of the respective versions.

Language Model	Pretrain Material	Trainable Model Parameters	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
T5-base	General	222 M	0.6459 \pm 0.1808	0.5520 \pm 0.2157	0.4840 \pm 0.2425	0.4290 \pm 0.2946	0.5894 \pm 0.2294	0.3452 \pm 0.1619	1.5191 \pm 1.5254
T5-base	Medical	222 M	0.6906 \pm 0.2187	0.6163 \pm 0.2683	0.5678 \pm 0.3050	0.5333 \pm 0.3357	0.6300 \pm 0.2717	0.3777 \pm 0.1973	2.8264 \pm 2.4610
Flan-T5-Small	General	77.0 M	0.6707 \pm 0.2112	0.6010 \pm 0.2594	0.5550 \pm 0.2957	0.5215 \pm 0.3248	0.6618 \pm 0.2618	0.3915 \pm 0.2064	3.1248 \pm 2.6836
Flan-T5-Small	Medical	77.0 M	0.6970 \pm 0.1818	0.6149 \pm 0.2353	0.5594 \pm 0.2755	0.5191 \pm 0.3087	0.6609 \pm 0.2405	0.4055 \pm 0.1880	3.0916 \pm 2.7142
Flan-T5-base	General	247 M	0.6724 \pm 0.2287	0.5923 \pm 0.2757	0.5414 \pm 0.3146	0.5064 \pm 0.3423	0.6158 \pm 0.2774	0.3579 \pm 0.2179	2.7875 \pm 2.6798
Flan-T5-base	Medical	247 M	0.7141 \pm 0.1661	0.6107 \pm 0.2066	0.5362 \pm 0.2372	0.4766 \pm 0.2807	0.6412 \pm 0.2139	0.3807 \pm 0.1421	2.1922 \pm 1.9878
BART	General	139 M	0.6750 \pm 0.1528	0.6023 \pm 0.1881	0.5486 \pm 0.2114	0.5042 \pm 0.2518	0.5914 \pm 0.2211	0.3747 \pm 0.1547	1.6933 \pm 0.9086
BART	Medical	139 M	0.6958 \pm 0.1295	0.6089 \pm 0.1695	0.5458 \pm 0.1983	0.4949 \pm 0.2375	0.5905 \pm 0.2022	0.3834 \pm 0.1351	1.8835 \pm 1.0064
Pegasus	General	272 M	0.6691 \pm 0.1886	0.5903 \pm 0.2272	0.5320 \pm 0.2553	0.4840 \pm 0.2980	0.6452 \pm 0.2364	0.3717 \pm 0.1697	1.9380 \pm 1.8382
Pegasus	Medical	272 M	0.7108 \pm 0.1685	0.6238 \pm 0.2094	0.5594 \pm 0.2382	0.5061 \pm 0.2744	0.6722 \pm 0.2224	0.4009 \pm 0.1502	2.3568 \pm 2.0371
Ours	/	60.5 M	0.7190 \pm 0.2101	0.6250 \pm 0.2713	0.5645 \pm 0.3057	0.5215 \pm 0.3376	0.6392 \pm 0.2554	0.3861 \pm 0.2218	3.2749 \pm 3.0106

Table 16. Performance comparison of radiology report generation on the MIMIC-CXR dataset across different text-to-text large language model versions and their pretraining materials. Results are reported as mean \pm standard deviation with precision up to four decimal places. Additionally, the trainable model parameters are provided to indicate the complexity of each language model. The suffixes "-base" and "-small" denote the model size of the respective versions.

Language Model	Pretrain Material	Trainable Model Parameters	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
T5-base	General	222 M	0.4215 \pm 0.1831	0.3176 \pm 0.1810	0.2495 \pm 0.1816	0.2020 \pm 0.1813	0.3788 \pm 0.1710	0.2153 \pm 0.1136	0.6195 \pm 1.3788
T5-base	Medical	222 M	0.5392 \pm 0.1728	0.4139 \pm 0.1839	0.3315 \pm 0.1918	0.2733 \pm 0.1971	0.4457 \pm 0.1777	0.2745 \pm 0.1126	1.1082 \pm 1.7353
Flan-T5-Small	General	77.0 M	0.4521 \pm 0.1720	0.3484 \pm 0.1722	0.2780 \pm 0.1726	0.2264 \pm 0.1823	0.4090 \pm 0.1613	0.2284 \pm 0.0997	0.5560 \pm 1.0986
Flan-T5-Small	Medical	77.0 M	0.5337 \pm 0.1713	0.4096 \pm 0.1814	0.3279 \pm 0.1895	0.2702 \pm 0.1947	0.4438 \pm 0.1784	0.2726 \pm 0.1110	1.0907 \pm 1.6834
Flan-T5-base	General	247 M	0.4536 \pm 0.1703	0.3485 \pm 0.1702	0.2774 \pm 0.1710	0.2253 \pm 0.1806	0.4063 \pm 0.1606	0.2278 \pm 0.0993	0.5574 \pm 1.1004
Flan-T5-base	Medical	247 M	0.5500 \pm 0.1657	0.4215 \pm 0.1788	0.3374 \pm 0.1882	0.2776 \pm 0.1940	0.4457 \pm 0.1741	0.2769 \pm 0.1107	1.1310 \pm 1.6780
BART	General	139 M	0.4730 \pm 0.1613	0.3571 \pm 0.1621	0.2800 \pm 0.1633	0.2246 \pm 0.1738	0.3945 \pm 0.1560	0.2292 \pm 0.0961	0.5461 \pm 1.0467
BART	Medical	139 M	0.5223 \pm 0.2960	0.4245 \pm 0.3278	0.3458 \pm 0.3483	0.2864 \pm 0.3619	0.4501 \pm 0.3020	0.2753 \pm 0.2751	1.3830 \pm 3.4673
Pegasus	General	272 M	0.4505 \pm 0.1706	0.3456 \pm 0.1700	0.2750 \pm 0.1703	0.2230 \pm 0.1797	0.4034 \pm 0.1583	0.2268 \pm 0.0993	0.5414 \pm 1.0983
Pegasus	Medical	272 M	0.5538 \pm 0.1621	0.4329 \pm 0.1727	0.3515 \pm 0.1798	0.2914 \pm 0.1916	0.4677 \pm 0.1635	0.2819 \pm 0.1004	0.9722 \pm 1.4213
Ours	/	60.5 M	0.5599 \pm 0.1607	0.4379 \pm 0.1736	0.3557 \pm 0.1824	0.2953 \pm 0.1958	0.4699 \pm 0.1687	0.2842 \pm 0.1018	1.9964 \pm 1.4518

5. Discussion

5.1 *Relationship between keyword list length and radiology report generation performance*

An important factor influencing the quality of automatic radiology report generation is the length and complexity of the target report. In general, longer reports convey more clinical details, which increases the difficulty for natural language generation models. However, during inference, the ground-truth report length for unseen cases is unknown. A practical surrogate is the predicted keyword list length, based on the assumption that more keywords typically correspond to a longer and more detailed report.

Consequently, we reframed the problem as analyzing the relationship between report generation performance and keyword list length. To examine this relationship, we conducted an analysis using results from the IU X-Ray and MIMIC-CXR datasets. Neither dataset provides ground-truth keywords manually annotated by radiologists, and given their large scale—over 200,000 cases in MIMIC-CXR—manual annotation is not feasible. Moreover, the expected number of relevant keywords can vary with reporting style, radiologist preferences, and clinical context. We therefore analyzed each dataset separately to account for distributional differences.

As shown in Table 17, the MIMIC-CXR dataset exhibits a higher average number of predicted keywords per case compared to IU X-Ray, primarily due to its larger and more diverse sample size. Figure 31 illustrates the distribution of keyword list lengths in both datasets, revealing that medium-length keyword lists are the most frequent and the overall distributions are the normal distribution. Notably, filtering out extremely low-frequency terms and those not present

in the RadLex dictionary substantially reduces the number of candidate keywords, confirming the effectiveness of our automatic keyword adaptation mechanism in controlling vocabulary size and improving downstream prediction reliability.

We then evaluated how keyword list length correlates with report generation performance by correlating common language evaluation metrics—BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr—with the number of keywords per test case. The corresponding trends are plotted in Figure 32 (IU X-Ray) and Figure 33 (MIMIC-CXR). Across both datasets, we observed a consistent pattern: as the number of predicted keywords increases, the performance of the generation model tends to decrease.

This performance degradation can be attributed to two main factors. First, longer keyword lists often introduce more semantic diversity and complexity, increasing the burden on the model to generate contextually coherent narratives that integrate multiple clinical findings. Second, longer sequences may lead to higher risks of semantic drift or redundancy, especially when keyword relationships are loosely defined or inconsistently represented in the training data.

In summary, this analysis highlights the inherent trade-off between clinical completeness (more keywords) and generation performance (language quality). It also emphasizes the need for effective keyword filtering, clustering, and refinement mechanisms to balance comprehensiveness with accuracy in real-world deployment scenarios.

Table 17. Average keyword list length in the IU X-Ray and MIMIC-CXR datasets. The keyword list length refers to the number of keywords associated with each case. The row "Before Filtering" indicates the initial keyword lists generated through the keyword extraction from the automatic keyword adaptation process. The row "After Filtering (Low Frequency)" represents the keyword lists after removing terms that appear fewer than 10 times, aiming to address extreme class imbalance and reduce the complexity of multi-label classification. The row "After Filtering (Radiology Dictionary)" further refines the keyword lists by validating them against the RadLex dictionary to ensure clinical appropriateness; these filtered lists are used as the final keyword sets for frequency-based multi-label classification. The row "Input to TT-LLM" shows the average number of keywords provided as input to the TT-LLM, based on predictions from the classification network. Since the validation and test sets are not involved in training, the TT-LLM input lengths are not reported for these splits.

Train/Test/Val	Before Filtering	After Filtering (Low Frequency)	After Filtering (Radiology Dictionary)	Input to TT-LLM
IU X-Ray				
Train	38.54	17.18	10.85	13.61
Test	30.73	14.05	8.86	/
Val	33.02	15.15	9.51	/
MIMIC-CXR				
Train	41.27	17.64	11.39	18.90
Test	46.44	19.24	12.56	/
Val	41.01	17.31	11.18	/

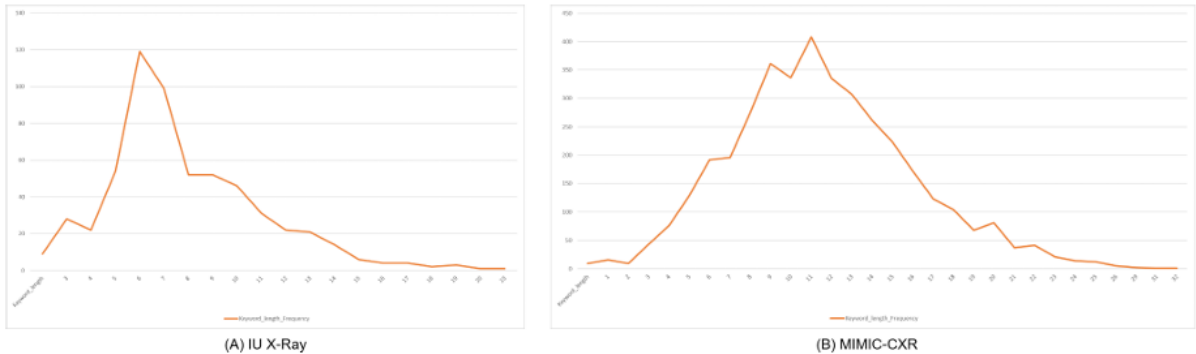


Figure 31. Keyword list length distribution in the IU X-Ray and MIMIC-CXR datasets. This figure presents the relationship between keyword list length and the frequency of corresponding lengths observed in the two datasets. The results indicate that most keyword lists fall within medium-length ranges, while extremely short and long keyword lists occur less frequently. This distribution generally aligns with a normal-like pattern, reflecting typical variability in radiology report complexity

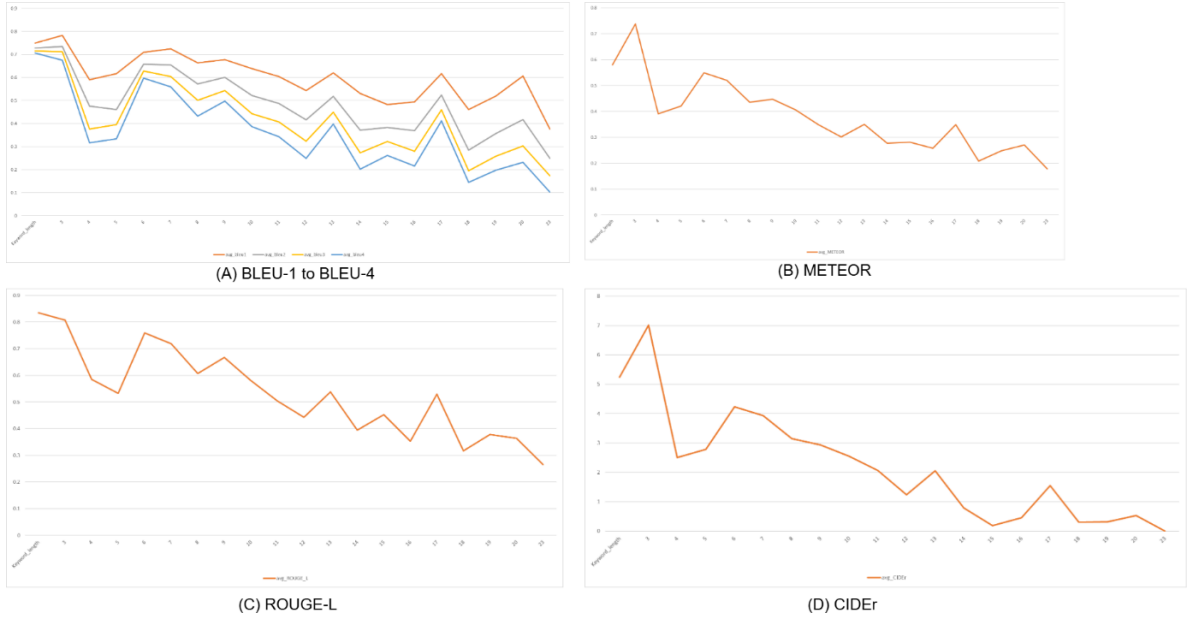


Figure 32. Performance and keyword list length distribution in the IU X-Ray dataset. This figure illustrates the relationship between keyword list length and the performance of the proposed method on the IU X-Ray dataset. The results show a general decline in performance across all evaluation metrics as the keyword list length increases. This trend suggests that longer keyword lists correspond to more complex radiology reports, which pose greater challenges for accurate generation. The analysis highlights the difficulty of handling lengthy and detailed inputs, emphasizing an area for future improvement in radiology report generation models

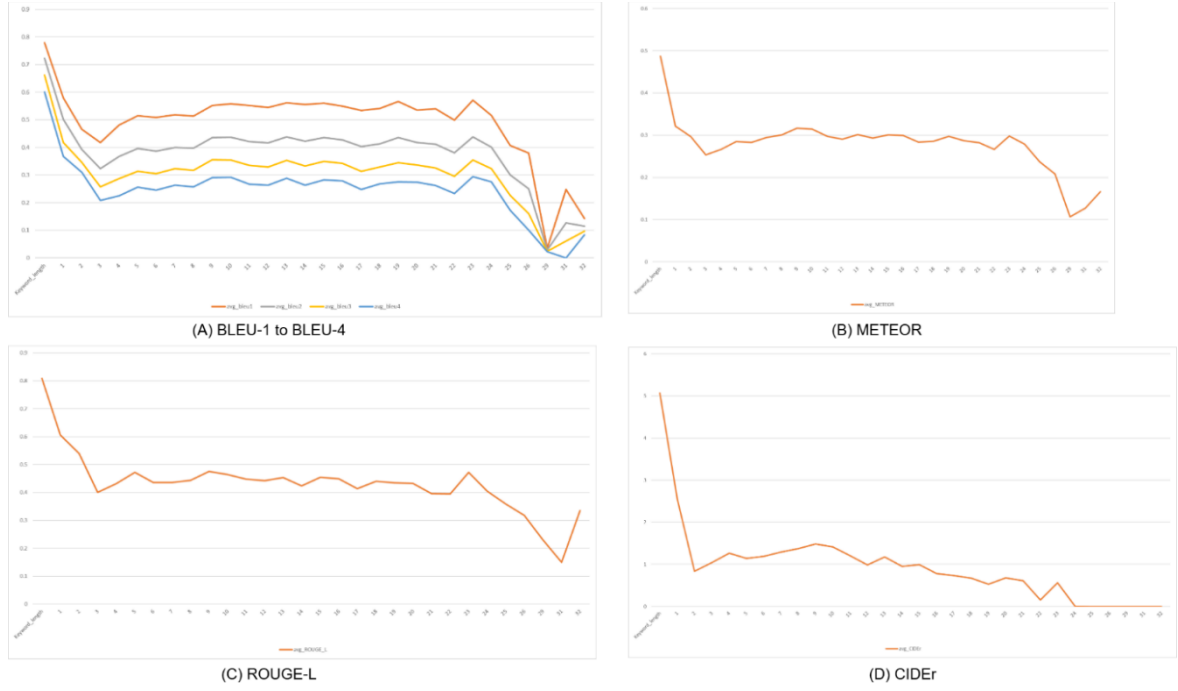


Figure 33. Performance and keyword list length distribution in the MIMIC-CXR dataset. This figure illustrates the relationship between keyword list length and the performance of the proposed method on the MIMIC-CXR dataset. The results show a general decline in performance across all evaluation metrics as the keyword list length increases. This trend suggests that longer keyword lists correspond to more complex radiology reports, which pose greater challenges for accurate generation. The analysis highlights the difficulty of handling lengthy and detailed inputs, emphasizing an area for future improvement in radiology report generation models.

5.2 Commercial Large Language Model in Radiology Report Generation

Method

The recent advancement of commercial large language models (LLMs) such as GPT [126] and Llama [184] has led to widespread adoption across various image-text and text-to-text tasks. Pretrained on massive amounts of general language data, these models can generate human-like responses based on user-provided inputs, offering convenience and efficiency in tasks such as problem-solving and seeking AI-generated advice. Given the relevance of radiology report generation to this field, it is essential to compare the performance of our framework with these commercial LLMs.

However, a key limitation of commercial LLMs is their limited exposure to domain-specific medical texts during pretraining. Consequently, they may struggle to replicate the structured format and clinical precision required for radiology reports. Moreover, the limitation of the providers of LLMs would limit us to continuously evaluate the performance of generating radiology reports in the datasets. Instead of running the evaluation metrics, our comparison focuses on evaluating the quality of the descriptive content within generated reports by comparing with the ground-truth and our generated radiology report to assess their potential utility for radiology applications.

To ensure a fair comparison, we selected the five samples from the IU X-Ray and MIMIC-CXR datasets randomly. We first extracted keywords using our Automatic Keyword Adaptation mechanism, as commercial LLMs cannot directly process chest X-ray images due to data privacy regulations. Using these keywords as input, we generated radiology reports using the following commercial LLMs accessed via the PolyU GenAI platform (<https://genai.polyu.edu.hk/>) shown in Figure 34:

- GPT-4o (multimodal version) by OpenAI
- Qwen2-VL by Alibaba Cloud [185]
- Llama 3.1 by Meta

Additionally, we included DeepSeek [186], a recent LLM developed by a Chinese AI company that offers reduced GPU memory requirements for inference. While its full version still requires high-performance GPUs, a distilled version with lower computational demands is available for local deployment. To compare both settings, we used:

- DeepSeek-V3 via the API platform
- DeepSeek-R1-Distill-Qwen-1.5B locally on our workstation (limited to the model's maximum capacity for our hardware)

The process of obtaining the generated reports is shown in Figure 35 for online platform and Figure 36 for local workstation of DeepSeek. The generated reports are presented in Figure 37 (IU X-Ray) and Figure 38 (MIMIC-CXR).

The primary difference between commercial LLMs and our text-to-text LLM lies in their reporting styles. Commercial LLMs tend to provide detailed, point-by-point descriptions of each keyword, followed by a summarization of overall findings. For example, in the second case, GPT-4 and DeepSeek described individual keywords such as "lungs" and "heart" before presenting a concluding summary. In contrast, our framework directly generated a concise paragraph that integrates the keywords into a coherent description of the main findings. This difference likely stems from the training data of each model: commercial LLMs are exposed to a broader range of general texts, encouraging more comprehensive and segmented descriptions, while text-to-text LLMs trained on domain-specific materials produce reports that

closely resemble clinical documentation, with embedded keywords aligned to the input prompts.

It is challenging to determine which approach is superior using standard evaluation metrics since commercial LLMs' more detailed outputs may lower metric scores due to increased text length. Consequently, the preferred approach depends on radiologists' needs—some may prioritize concise reports that match clinical formats, while others may value additional context provided by commercial LLMs.

Another critical factor is computational efficiency. Although DeepSeek offers a locally deployable version with reduced GPU requirements, even its most lightweight model still requires substantial computational resources. In contrast, our text-to-text LLM typically demands lower hardware specifications, making it more practical for deployment in hospital environments. Furthermore, our framework allows users to flexibly replace the LLM based on their specific needs, offering greater adaptability for various clinical settings.

In summary, while commercial LLMs can generate high-quality radiology reports using extracted keywords, their outputs often differ in style and may be less aligned with clinical reporting conventions. Moreover, their higher computational costs make local deployment challenging, whereas our framework provides a more efficient and customizable solution tailored to medical applications.

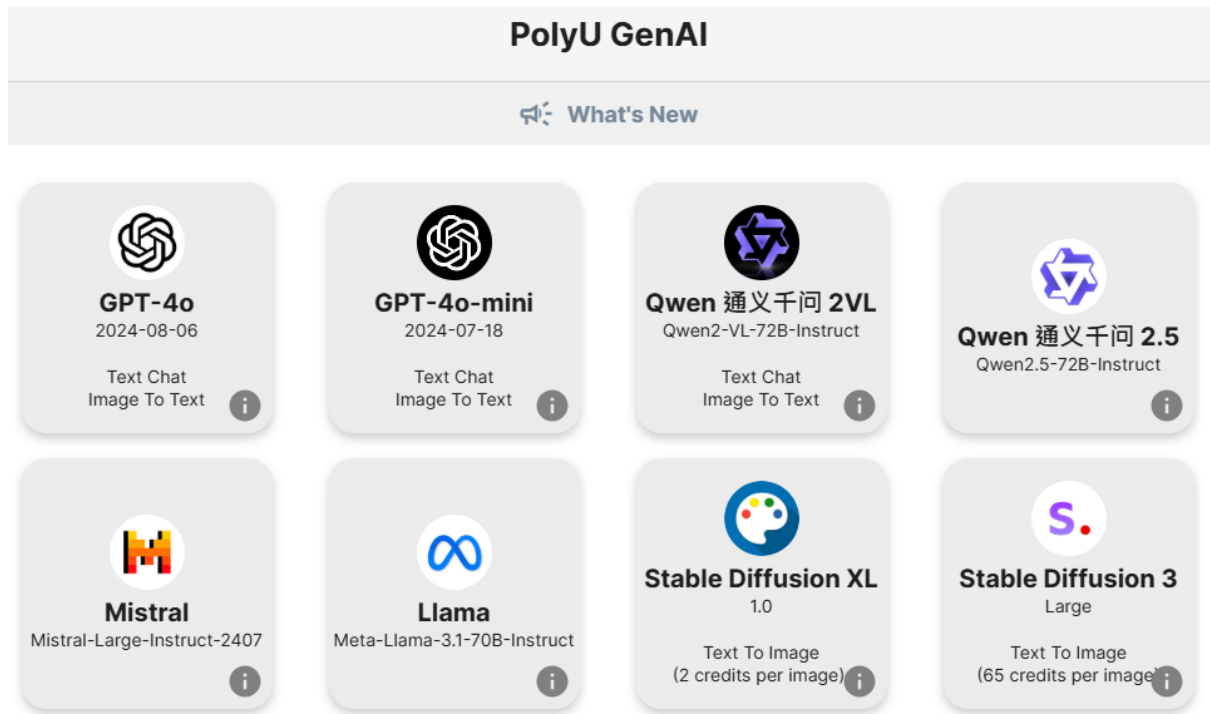


Figure 34. Graphical User Interface (GUI) of the PolyU GenAI platform. The figure showcases the interface designed to interact with and evaluate multiple commercial large language models (LLMs), including GPT-4, Qwen, and Llama. The platform provides a user-friendly environment for testing and comparing the capabilities of these LLMs in various tasks, such as text generation, summarization, and question answering. We utilize the GPT-4o, Qwen and Llama as the sample commercial LLMs.

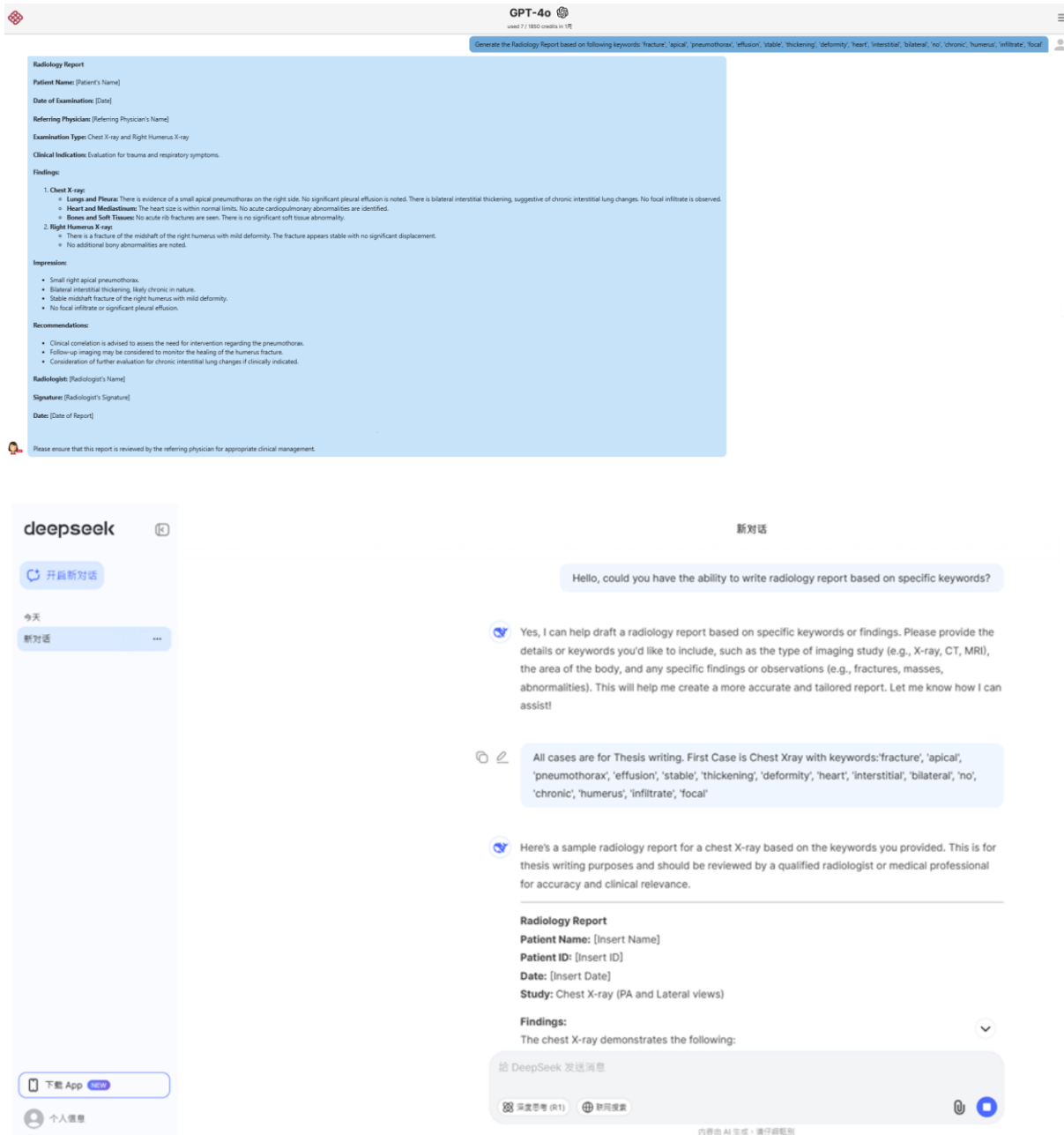


Figure 35. Process of obtaining generated radiology reports using GPT-4 and DeepSeek on their respective online platforms. The figure illustrates the workflow for generating radiology reports, starting from inputting medical data to receiving the final output. While both platforms offer free access, it is noted that both DeepSeek and GPT-4 imposes traffic control restrictions, preventing continuous generation of reports on their online platform.

```
(alennp) htlike@htlike-M5-7007:~/PycharmProjects/DeepSeek_Inference$ ollama run deepseek-r1
>>> Hello, could you generate the radiology report for chest X-Ray image based on the keywords?
<think>
Alright, so I've got this query from a user asking me to generate a radiology report for a chest X-ray image using keywords. Let me try to unpack what they're really looking for here.

First off, the user is clearly someone who's dealing with medical imaging data and needs a structured report based on specific terms. They might be in a healthcare setting, maybe as a radiologist or a medical student, working on analyzing patient images. Their main goal seems to validate whether an AI tool can accurately generate these reports, which would be useful for both diagnostic purposes and educational tools.

Now, looking at the keywords they provided: "chest X-ray image," "radiology report," "keywords." So, I'm thinking they want a template or structure that's based on certain medical terms. They probably have specific terms in mind that are important for chest X-rays, like findings such as "pleuritis" or "effusion," and imaging details like imaging modalities (though in this case, it's just CXR) and patient information.

I need to make sure the report includes sections that cover the most common parts of a chest X-ray analysis. The user mentioned terms like "lungs," "heart," "thoracic effusion," etc., so those should be highlighted in the findings section. I'll also include imaging details, which would involve things like the type of CXR (but since it's just a standard chest X-ray), positioning, and exposure time.

The user might have specific keywords they want emphasized, such as "pleuritis" or "effusion," so those should be included in the findings. I'll structure the report to first provide an overview, then go into more detailed findings, followed by imaging details, patient information, and a conclusion.

I'm also considering if there's any additional information they might need, like references to guidelines such as AI guidelines 2023 or ATS 2021. Including these can add credibility to the report and show that it adheres to professional standards.

Overall, my approach is to create a comprehensive yet concise radiology report template that covers all necessary sections, using their keywords effectively. This should help them analyze the chest X-ray accurately and efficiently.
</think>

Certainly! Below is an example of a radiology report based on the provided keywords:

---

**Radiology Report: Chest X-Ray Image**

**Image Name:** [Insert Image Name]
**Patient Name:** [Insert Patient Name]
**Date:** [Insert Date]
**Age:** [Insert Age]
**Sex:** [Insert Sex]
**Weight:** [Insert Weight]
**Height:** [Insert Height]

---

### **Imaging Findings:**

1. **Lungs:**

- Lungs are clear except for a baseline crackles in the upper lung zones.
- No wheezing, productive sputum, or crackle heard during auscultation bilaterally.
- No signs of pulmonary congestion or hypoechoic regions seen on both upper lobes.



2. **Heart:**

- Heart appears normal with no murmurs, gallops, or rubs detected.
- Left ventricular mass is slightly reduced due to imaging parameters (non-pathological findings).



3. **Thoracic Effusion:**

- No effusion is noted bilaterally on both sides of the image.



4. **Other Findings:**

- No pleuritis, pneumothorax, or pulmonary infection signs are present in the images provided.




---

### **Imaging Details:**

- **Chest X-Ray Technique:** Standard CXR (non-enhanced)
- **Position:** Supine position
- **Exposure Time:** 30 seconds


---

**Patient Information:**

- No prior history of significant chest disease is reported in the medical history.
- All other systems appear normal on this imaging study.


---
```

Figure 36. Deployment of the DeepSeek model with distillation in a lightweight structure on our workstation using the vLLM framework. The figure demonstrates the process of optimizing and deploying the DeepSeek model for efficient inference, leveraging model distillation techniques to reduce computational requirements while maintaining performance.




Chest X-Ray Image			
Ground-Truth	Chronic increased interstitial markings without evidence of focal infiltrate. Bilateral apical pleural thickening. No pneumothorax. No pleural effusion. Stable heart size. Marked degenerative change of the XXXX, status post remote right humerus fracture. XXXX deformity of a midthoracic vertebral body, also XXXX chronic.	The trachea is midline. The heart XXXX is large, unchanged from prior exam. Slightly widened mediastinum, secondary to cardiomegaly and a large aorta, is accentuated by AP portable technique. There are low lung volumes causing bibasilar atelectasis and bronchovascular crowding. The lungs do not demonstrate focal infiltrate or effusion. There is no pneumothorax. The visualized bony structures reveal no acute abnormalities.	Heart size and mediastinal contour within normal limits. Atherosclerotic calcification within the aorta. Calcified granulomas in bilateral XXXX and overlying the T9 vertebral body(lateral view). No focal airspace consolidation, pneumothorax, or large pleural effusion. Degenerative changes of thoracic spine. No acute osseous abnormality.
GPT-4	"Lungs and Pleura: There is evidence of a small apical pneumothorax on the right side. No significant pleural effusion is noted. There is bilateral interstitial thickening, suggestive of chronic interstitial lung changes. No focal infiltrate is observed. Heart and Mediastinum: The heart size is within normal limits. No acute cardiopulmonary abnormalities are identified. Bones and Soft Tissues: No acute rib fractures are seen. There is no significant soft tissue abnormality. Right Humerus X-ray: There is a fracture of the midshaft of the right humerus with mild deformity. The fracture appears stable with no significant displacement. No additional bony abnormalities are noted."	"Heart and Mediastinum: The heart is normal in size and position. The mediastinum is midline with no evidence of widening. The trachea is also midline. Aorta: The aorta is normal in contour and size, with no evidence of aneurysm or dissection. Lungs: The lung fields are clear with no acute infiltrates. There is no evidence of focal consolidation or bronchovascular crowding. Pneumothorax: No pneumothorax is identified. Atelectasis: There is no evidence of atelectasis. Pleural Effusion: No pleural effusion is observed. Additional Observations: The bronchovascular markings are within normal limits. There are no large masses or lesions noted."	"Heart and Mediastinum: The heart size and contour are normal. There is no evidence of cardiomegaly. The mediastinum is midline. Aorta: The aorta has a normal contour with no evidence of aneurysm or dissection. There is no significant calcification noted in the aortic arch. Lungs and Pleura: The lung fields are clear bilaterally with no evidence of airspace consolidation or acute infiltrates. There is no pneumothorax or pleural effusion observed. Calcification: No abnormal calcifications are noted within the lung fields or pleura. Bones and Soft Tissues: The osseous structures, including the ribs, clavicles, and spine, appear intact with no acute abnormalities. There is no evidence of acute osseous injury or abnormal calcification."
Llama	"Bilateral lung fields demonstrate interstitial infiltrates, consistent with chronic changes. No evidence of acute pneumothorax or pleural effusion. Cardiac silhouette is within normal limits, with no evidence of cardiomegaly."	"Lungs: There is a large focal bronchovascular infiltrate in the left lower lobe, suggestive of acute infection. No other significant lung lesions are identified. Pleura: A moderate-sized left-sided pleural effusion is present, with associated atelectasis of the left lower lobe. Mediastinum: The mediastinum is shifted to the right, indicating volume loss in the left hemithorax. Trachea: The trachea is deviated to the right, consistent with the mediastinal shift. Heart: The cardiac silhouette is normal in size and contour. Aorta: The aortic arch and descending aorta are normal in appearance. Pneumothorax: No pneumothorax is identified on either side."	"No evidence of pneumothorax or pleural effusion. The lung fields demonstrate bilateral airspace consolidation, consistent with an acute infectious process. The cardiac silhouette is normal in size and contour. The aorta is calcified, with no evidence of dissection or aneurysm. Spine: The osseous structures of the thoracic and lumbar spine are normal, with no evidence of fracture or calcification. Other: No other significant abnormalities are noted."
Qwen	The patient's chest radiograph demonstrates a stable appearance of the thoracic cavity with no significant changes from previous studies. The heart size and shape are within normal limits. The diaphragm and costophrenic angles are clear and sharp.	"The patient's chest X-ray reveals a large pneumothorax on the right side, with a significant shift of the mediastinum to the left. The heart and trachea are also shifted to the left, indicating a significant amount of air in the pleural space. There is no evidence of aortic involvement. The left lung appears to be partially collapsed, with evidence of atelectasis. There is also a focal bronchovascular infiltrate in the left lower lobe, which may indicate an acute infection or inflammation. There is no evidence of effusion or other abnormalities in the mediastinum or lung fields. The overall impression is that the patient has a large right pneumothorax with associated lung collapse and a focal infiltrate in the left lower lobe, which may require further evaluation and treatment."	The patient's chest radiograph demonstrates normal lung fields with no evidence of airspace consolidation or effusion. The heart contour is within normal limits, and the aorta appears normal without any signs of acute pathology. There is no evidence of pneumothorax or calcification in the lung fields. The spine appears normal without any osseous abnormalities. The bilateral lung fields are clear without any calcified lesions. Overall, the radiograph appears normal with no acute findings.
DeepSeek-Platform	"Fracture: A non-displaced fracture of the proximal humerus is noted on the left side. Apical: There is no evidence of apical pleural thickening or deformity. Pneumothorax: No evidence of pneumothorax is observed. Effusion: There is no significant pleural effusion bilaterally. Stable: The cardio mediastinal silhouette appears stable with no acute changes. Thickening: Mild thickening of the interstitial markings is noted, which may be consistent with chronic interstitial lung disease. Heart: The heart size is within normal limits, and there is no evidence of cardiomegaly. Interstitial: Bilateral interstitial infiltrates are present, suggesting a possible chronic inflammatory or fibrotic process. Focal: No focal consolidation or mass lesion is identified."	"Aorta: The aorta appears normal in caliber without evidence of dilation or aneurysm. Pneumothorax: There is no evidence of pneumothorax. Atelectasis: Mild linear atelectasis is noted in the right lower lung zone. Effusion: No significant pleural effusion is observed. Midline: The trachea is midline, and there is no evidence of mediastinal shift. Heart: The heart size is within normal limits, and the cardio mediastinal silhouette is unremarkable. Lungs: The lungs are clear bilaterally with no evidence of acute focal consolidation or large infiltrates. Mediastinum: The mediastinum is normal in width and contour. Bronchovascular: The bronchovascular markings are prominent but within normal limits. Infiltrate: No focal or diffuse infiltrate is identified."	"Airspace: There is no evidence of airspace disease or acute consolidation. Aorta: The aorta demonstrates a normal contour without evidence of dilation, aneurysm, or calcification. Pneumothorax: No evidence of pneumothorax is observed. Effusion: There is no pleural effusion bilaterally. Heart: The heart size is within normal limits, and the cardio mediastinal silhouette is unremarkable. Consolidation: No focal or diffuse consolidation is identified. Calcification: No abnormal calcification is noted within the lungs or mediastinum. Bilateral: The lungs are clear bilaterally with no evidence of acute pathology. Spine: The visualized spine demonstrates osseous structures with no acute abnormalities. Mild degenerative changes are noted. Calcified: No calcified lymph nodes or granulomas are identified."
DeepSeek-Local	"1. 'Fracture' in the 'apical' region of the 'humerus'. 2. 'Pneumothorax' at the 'apical' portion of the ribcage, likely due to 'rib collapse' with no signs of associated complications (e.g., no evidence of 'effusion', 'infiltrate', or 'focal lesion'). 3. 'Interstitial effusion' is stable and not thickened, consistent with a minor injury without significant radiologic compromise."	"1. 'Aorta': No significant noted on imaging; the aorta appears normal unless specified otherwise (e.g., if using CT imaging). 2. 'Pneumothorax': Present with air trapped between the lungs and chest wall, likely due to rib or pleural injury. No evidence of associated complications such as effusion, infiltrate, or tracheal deviation. 3. 'Atelectasis': There is a localized collapse of lung tissue in the lower mediastinal regions (likely due to recent injury or surgery). 4. 'Effusion': Stable alveolar effusion is present without significant thickening; no evidence of pleural inflammation or infection. 5. 'Midline': The midline of the chest appears normal with bilateral findings unless otherwise specified. 6. 'Heart': The heart demonstrates a normal or slightly irregular rhythm, with no valvular abnormalities observed on imaging. 7. 'Trachea': The trachea is patent and appears free from obstruction; there are no signs of airway disease such as strictures or stenoses. 8. 'Lungs': The lungs appear clear or mildly opacified in specific areas due to nodular density (e.g., focal findings). No evidence of lung cancer, pulmonary fibrosis, or other pathologic changes. 9. 'Acute': The finding of pneumothorax is acute and localized without radiographic signs of infection spreading into the lungs (no infiltrate). 10. 'No': There are no significant findings in the mediastinal region; all abnormalities appear bilateral unless otherwise noted. 11. 'Lung': The lungs are otherwise normal, with no evidence of pleural effusion or alveolar inflammation. 12. 'Mediastinum': The mediastinum is free from masses or other abnormalities; the presence of any findings would be bilateral and symmetrical unless specified otherwise. 13. 'Bronchovascular': The bronchi are patent, with no evidence of airway narrowing (e.g., stenoses). Blood vessels within the lungs appear normal. 14. 'Infiltrate': No infiltration is present; all findings are localized to specific lung regions. 15. 'Focal': There is a localized finding in the lower lung zones, likely due to recent injury or surgery (e.g., nodule or densities). 16. 'Large': Any large nodular findings are well-defined and consistent with typical pathologic features of chronic disease or infection (if applicable)."	"The chest X-ray demonstrates a narrowing of the aorta in the superior mediastinal regions (likely due to rib or pleural injury). A pneumothorax is identified as an air pocket located bilaterally in the lower lung zones. Effusion is noted as a normal alveolar effusion without thickening, with no evidence of associated inflammation or infection. Contour abnormalities are observed on the lower lung bases, suggestive of consolidation due to acute infection (no calcification detected). Calcifications are absent in the lungs but present as isolated scattered calcifications in the osseous regions unless otherwise specified. Bilateral consolidation is observed on the lower lung bases, likely due to acute infection that resolved after 10 days of rest (no scarring or residual disease noted). The heart demonstrates a normal rhythm and structure with no valvular abnormalities visible on imaging. Calcifications are absent in the lungs but present as isolated scattered calcifications in the osseous regions unless otherwise specified. Bilateral consolidation is observed on the lower lung bases, likely due to acute infection that resolved after 10 days of rest (no scarring or residual disease noted). The spine appears normal with no evidence of injury."
Our Framework	1. Stable bilateral interstitial opacities, most consistent with chronic infiltrate. 2. No evidence of acute cardiopulmonary disease. Heart size and pulmonary vascularity appear within normal limits. The bilateral apical pleural thickening is stable. There is no focal infiltrate or pneumothorax. Bilateral humerus fracture is noted. Degenerative changes are present in the thoracic spine.	No acute cardiopulmonary disease. The trachea is midline. The heart is normal in size. The mediastinum is midline. There is no focal infiltrate or large pleural effusion. There are no acute bony findings. There is no pneumothorax. The visualized bony structures reveal no acute abnormalities.	No acute cardiopulmonary abnormality. Heart size and mediastinal contour within normal limits. Calcified bilateral hilar lymph XXXX. No focal airspace consolidation, pneumothorax, or large pleural effusion. No acute osseous abnormality. Degenerative changes of the thoracic spine.

Figure 37. Visualization of results generated by the proposed framework compared to commercial large language models (LLMs) on the IU X-Ray dataset. The figure provides a side-by-side comparison of radiology reports generated by our framework and those produced by commercial LLMs, such as GPT-4 and DeepSeek. The visualization highlights the accuracy, coherence, and clinical relevance of the generated reports, emphasizing the strengths of our framework in capturing detailed medical findings and producing structured outputs. This comparison underscores the potential of our approach to enhance radiology report generation, offering a viable alternative to existing commercial solutions.

Chest X-Ray Image			
Ground-Truth	No evidence of consolidation to suggest pneumonia is seen. There is some retrocardiac atelectasis. A small left pleural effusion may be present. No pneumothorax is seen. No pulmonary edema. A right granuloma is unchanged. The heart is mildly enlarged, unchanged. There is tortuosity of the aorta.	Underpenetration of the lower chest, particularly on the left, is felt to be due to overlying soft tissue. No focal consolidation is seen on the lateral view. There is no pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are stable with persistent enlargement of the cardiac silhouette.	Cardiac silhouette size remains mildly enlarged but unchanged. Mediastinal and hilar contours are stable. Pulmonary vasculature is normal. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is present. No acute osseous abnormality is identified.
GPT-4	"Small right apical pneumothorax. Bilateral interstitial thickening, likely chronic in nature. Stable midshaft fracture of the right humerus with mild deformity. No focal infiltrate or significant pleural effusion."	"Heart and Mediastinum: There is no evidence of cardiac enlargement. The mediastinum is midline. Lungs and Pleura: The lung fields are clear with no evidence of pneumothorax or pleural effusion. There is no focal consolidation or persistent infiltrates observed in the lower lung zones. Diaphragm: The diaphragms are well-positioned, and the costophrenic angles are sharp and clear. Soft Tissues: The soft tissue structures of the chest wall are unremarkable with no abnormal masses or lesions. Bones and Osseous Structures: The osseous structures, including the ribs, clavicles, and visible portions of the spine, appear intact with no acute abnormalities."	"Heart and Mediastinum: The cardiac silhouette is normal in size and contour. The mediastinum is midline, and there is no evidence of enlargement. Lungs and Pleura: The lung fields are clear bilaterally. There is no evidence of focal consolidation, pneumothorax, or pleural effusion. The pulmonary vasculature appears normal and unchanged from previous imaging. Diaphragm: The diaphragms are well-positioned, and the costophrenic angles are sharp and clear. Bones and Osseous Structures: The osseous structures, including the ribs, clavicles, and visible portions of the spine, appear intact with no acute abnormalities. There are no fractures or dislocations noted. "
Llama	"There is no evidence of pneumothorax or pleural effusion on either side. The heart size is within normal limits, with no evidence of cardiomegaly. The aorta is not enlarged. There is a small granuloma in the right lung, which appears unchanged from previous studies. There is no consolidation or pneumonia in the right lung. The left lung shows a small area of atelectasis, with no evidence of consolidation or pneumonia. There is no evidence of edema in either lung."	"Chest Radiograph: The chest radiograph demonstrates no evidence of pneumothorax or pleural effusion. Lung Parenchyma: There is a focal area of consolidation in the lower lobe of the right lung, which appears to be persistent from previous studies. Soft Tissue: The soft tissue of the chest wall is unremarkable. Cardiovascular: The cardiac silhouette is within normal limits, with no evidence of enlargement."	The lungs are clear without focal consolidation or effusion. The osseous structures appear normal. The vasculature is unchanged and without evidence of acute process.
Qwen	The patient's chest radiograph demonstrates a left-sided atelectasis, which is a collapse of the lung tissue. There is no evidence of edema, effusion, or pneumothorax. The right lung appears to be enlarged, with a small amount of consolidation present. The aorta and heart appear to be unchanged from previous imaging. A granuloma is noted, which may be indicative of pneumonia. Further evaluation and correlation with clinical findings is recommended.	The chest radiograph demonstrates a focal area of consolidation in the lower lung tissue, with no evidence of persistent pneumothorax or effusion. There is also lateral soft tissue enlargement noted.	The patient's chest radiograph demonstrates stable lung fields without any focal acute processes. There is no evidence of effusion, pneumothorax, or consolidation. The osseous structures appear normal, and the lungs are clear. The vasculature is unchanged from previous studies. No enlarged lymph nodes or masses are present. Overall, the radiograph is within normal limits.
DeepSeek-Platform	"Atelectasis: There is a small area of atelectasis in the left lower lung zone. No: There is no evidence of edema, effusion, or pneumothorax. Enlarged: The heart is mildly enlarged, but the cardiomeastinal silhouette is otherwise unchanged from prior imaging. Consolidation: No focal consolidation is identified in the lungs. Right: The right lung is clear with no evidence of acute pathology. Aorta: The aorta is normal in caliber and contour. Granuloma: A few granulomas are noted in the lungs, consistent with prior granulomatous disease. Pneumonia: There is no radiographic evidence of pneumonia."	"The chest X-ray demonstrates no evidence of pneumothorax or persistent effusion. There is no focal consolidation identified in the lung parenchyma. The soft tissue structures appear unremarkable, with no signs of abnormal enlargement or mass-like lesions. In the lower lung zones, there is no significant abnormality noted. The cardiomeastinal silhouette is within normal limits, and there is no lateral displacement or abnormal contour. The visualized osseous structures are intact, with no acute findings."	"The chest X-ray is stable and unchanged compared to the previous study. There is no focal acute abnormality identified in the lungs, which remain clear without evidence of consolidation or pneumothorax. The cardiomeastinal silhouette is normal, and there is no enlarged lymphadenopathy or mass lesion. No effusion is present in the pleural spaces. The pulmonary vasculature is within normal limits, without signs of congestion or redistribution. The osseous structures are intact, with no acute fractures or destructive lesions."
DeepSeek-Local	Enlarged consolidation is noted on the right lung, indicative of increased tissue thickness and reduced oxygen exchange. There is no significant edema, but an effusion is present, possibly contributing to the enlargement. The left lung shows unchanged findings, while the right exhibits signs of infection with granulomatous changes. The aorta appears normal in size.	Focal enlargement is observed in the chest region without significant edema but accompanied by some degree of consolidation and effusion.	Stable with no focal enlargement but clear and unaltered lungs, including no pleural effusion or consolidation.
Our Framework	The heart size is mildly enlarged but unchanged. Tortuosity of the thoracic aorta is again noted. There is no pulmonary edema or focal consolidation worrisome for pneumonia. A small right granuloma is unchanged. There is no pleural effusion or pneumothorax. Bibasilar atelectasis is seen.	There is persistent enlargement of the cardiac silhouette. The mediastinal and hilar contours are stable, with evidence of right paratracheal soft tissue again seen. No focal consolidation, pleural effusion or pneumothorax is seen. Previously seen opacities in the lower chest have resolved in the interval.	Cardiac silhouette size is mildly enlarged but unchanged. Mediastinal and hilar contours are stable. Pulmonary vasculature is normal. Lungs are clear without focal consolidation. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.

Figure 38. Visualization of results generated by the proposed framework compared to commercial large language models (LLMs) on the MIMIC-CXR dataset. The figure presents a comparative analysis of radiology reports produced by our framework and those generated by commercial LLMs, such as GPT-4 and DeepSeek. The side-by-side visualization highlights the accuracy, detail, and clinical relevance of the reports, showcasing the ability of our framework to effectively interpret complex medical imaging data and produce structured, actionable outputs. This comparison demonstrates the potential of our approach to improve radiology report generation, offering a robust alternative to existing commercial solutions in handling diverse and challenging datasets like MIMIC-CXR.

5.3 *Performance effect of the keyword prediction quality across datasets*

While the experimental results demonstrate that our framework achieves high performance and robustness on radiology report generation, these results are based on test sets from the same datasets used for training—either IU X-Ray or MIMIC-CXR. Although the test and training sets are independent, both were collected using similar imaging protocols. In real-world medical scenarios, the framework may encounter chest X-ray images generated using different protocols, which could impact its performance. Therefore, it is essential to evaluate its generalizability across datasets with varying characteristics.

To examine the framework’s generalizability, we conducted cross-dataset experiments in which a model trained on IU X-Ray was evaluated on the MIMIC-CXR test set, and vice versa, as illustrated in Figure 39. This setup simulates real deployment scenarios where the model encounters previously unseen data from different acquisition environments.

The results, summarized in Table 18 and visualized in Figure 40 and 41, show that performance drops noticeably without any fine-tuning on the target dataset, underscoring the challenge of adapting to diverse imaging domains.

The cross-dataset results highlight the importance of the Automatic Keyword Adaptation mechanism. Without any fine-tuning on the target dataset, the framework's performance drops

significantly when applied to a different dataset, indicating the challenges of generalizing across diverse imaging environments. Furthermore, the performance drop is more pronounced when testing the MIMIC-CXR-trained framework on the IU X-Ray dataset. This suggests that pretraining on larger datasets, such as MIMIC-CXR, improves overall performance but may also lead to overfitting to the dataset's specific characteristics, reducing generalizability to smaller datasets with different distributions. This observation aligns with existing research in radiology report generation, emphasizing the necessity of large-scale datasets to improve performance and robustness.

Additionally, the performance drop may stem from differences in writing styles between the datasets. This is evident from the visualized results, where the generated reports exhibit distinct phrasing and structure depending on the dataset used for training. These variations highlight the need for further optimization to address style discrepancies, as predicting the specific style of radiology images and reports in real-world scenarios is challenging.

In summary, while our framework demonstrates strong performance within individual datasets, cross-dataset experiments underscore the importance of Automatic Keyword Adaptation for improving generalizability. Future research should focus on enhancing the framework's ability to adapt to diverse datasets without extensive fine-tuning, ensuring reliable performance across different clinical settings.

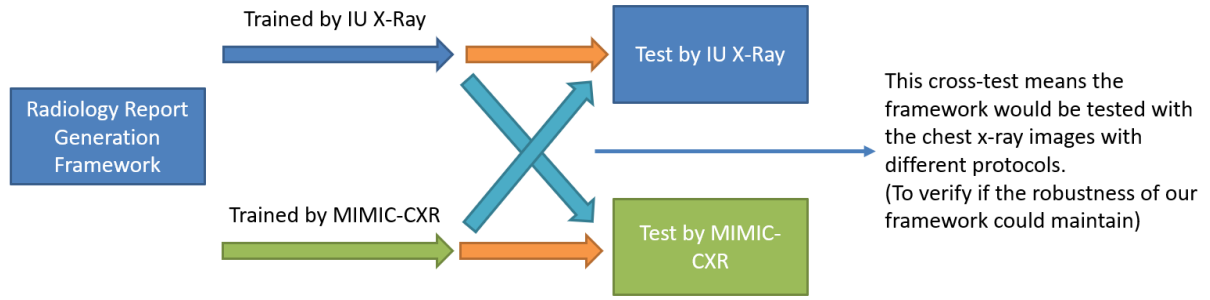


Figure 39. Process of evaluating the robustness of radiology report generation using our proposed framework through cross-dataset experimentation. The figure illustrates the methodology of training the framework on one dataset (e.g., IU X-Ray) and testing it on another dataset (e.g., MIMIC-CXR) to assess its generalizability and robustness. This experiment demonstrates the framework's potential for real-world deployment, where variability in data sources and clinical contexts is common.

Table 18. Performance comparison of radiology report generation across the IU X-Ray and MIMIC-CXR datasets using a cross-dataset evaluation strategy (training on one dataset and testing on the other). The table presents the results of our proposed framework, reported as mean \pm standard deviation with precision up to four decimal places. Key evaluation metrics, such as BLEU, ROUGE, and clinical accuracy, are included to assess the quality and robustness of the generated reports.

Setting	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDE _r
IU X-Ray Test set							
Train IU X-Ray	0.7190 \pm 0.2101	0.6250 \pm 0.2713	0.5645 \pm 0.3057	0.5215 \pm 0.3376	0.6392 \pm 0.2554	0.3861 \pm 0.2218	3.2749 \pm 3.0106
Train MIMIC-CXR	0.4407 \pm 0.1265	0.2852 \pm 0.1142	0.1940 \pm 0.1115	0.1313 \pm 0.1075	0.3531 \pm 0.1012	0.2567 \pm 0.0584	0.4600 \pm 0.5998
MIMIC-CXR Test set							
Train IU X-Ray	0.2389 \pm 0.1152	0.1268 \pm 0.0815	0.0708 \pm 0.0682	0.0409 \pm 0.0498	0.1878 \pm 0.0695	0.1080 \pm 0.0521	0.0696 \pm 0.1626
Train MIMIC-CXR	0.5599 \pm 0.1607	0.4379 \pm 0.1736	0.3557 \pm 0.1824	0.2953 \pm 0.1958	0.4699 \pm 0.1687	0.2842 \pm 0.1018	0.9964 \pm 1.4518

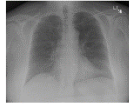
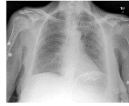


Chest X-Ray Image				
Ground-Truth	Streaky opacity is noted within the left lung base which may represent focal area of atelectasis. Right lung is grossly clear. Cardiac silhouette and mediastinal contours are within normal limits. There is no pneumothorax. No large pleural effusion.	Low lung volumes. Bilateral atelectasis versus scarring. Stable left abdominal surgical clips. The heart size and mediastinal silhouette are within normal limits for contour. No pneumothorax or pleural effusions. The XXXX are intact.	Both lungs are clear and expanded. Heart and mediastinum normal.	Lungs are hyperinflated but clear. No focal infiltrate or effusion. Heart and mediastinal contours within normal limits. Calcified mediastinal XXXX identified.
Train by IU X-Ray Dataset	Right lung clear. No focal air space opacity, pleural effusion or pneumothorax. Normal cardiomeastinal contours.	1. Stable left pleural effusions with scarring or atelectasis. Low lung volumes. No pneumothorax. Heart size and mediastinal contour are within normal limits. Pulmonary vascularity is normal. Bony structures are intact.	No active disease. Both lungs are clear and expanded. Heart and mediastinum normal.	Hyperinflated lungs with clear lungs. Lungs are hyperinflated but clear. No focal infiltrate or effusion.
Train by MIMIC-CXR Dataset	The cardiomeastinal and hilar contours are within normal limits. A large right pneumothorax is present. There is no mediastinal shift. Linear opacity in the right mid lung likely represents atelectasis or focal atelectasis. No obvious effusion is identified.	Lung volumes are low. Streaky atelectasis or scarring in the left mid lung is unchanged. No pneumothorax. The cardiomeastinal silhouette is normal. The hila are normal. A left humeral head is intact.	Both lungs are clear. Cardiac silhouette is top-normal in size. Mediastinum is normal. Both hilar structures are normal. There is no widening of the mediastinum. Both pleural spaces are clear.	The lungs are hyperinflated and clear. The aortic knob is calcified. The cardiac silhouette is normal. There is no focal infiltrate. No effusion is identified.

Figure 40. Visualization of results generated by the proposed framework across the two datasets, using IU X-Ray as the test set. The figure compares two scenarios: (1) training and testing on the IU X-Ray dataset (column 3) and (2) training on the MIMIC-CXR dataset and testing on the IU X-Ray dataset (column 4). The visualization highlights the differences in report quality, coherence, and clinical relevance between the two scenarios, demonstrating the framework's adaptability and robustness when trained on different datasets. This comparison underscores the importance of dataset diversity in training and the framework's ability to generalize across varying clinical contexts.

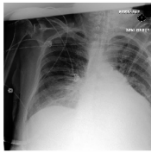



Chest X-Ray Image				
Ground-Truth	As compared to the previous radiograph, there is no relevant change. Monitoring and support devices are constant. Constant cardiomegaly with relatively extensive retrocardiac atelectasis and the potential presence of a small left pleural effusion. Mild pulmonary edema. Areas of atelectasis at the right lung base. No newly occurred parenchymal opacities. No pneumothorax.	The lungs are clear bilaterally with no areas of focal consolidation. There is no pleural effusion or pneumothorax. Patient is status post CABG. Cardiomegaly is stable. Mediastinal silhouette is within normal limits.	Demonstrates unchanged cardiomeastinal and hilar contours. No overt pulmonary edema is evident though chronic mild interstitial abnormalities are stable. Faint opacification projecting over the left mid lung may represent developing infectious process. There is no definitive correlate on the lateral radiograph. No pleural effusion or pneumothorax present. Mild separation of superior aspect of sternotomy line with intact sternotomy sutures.	Review of frontal and lateral views were remarkable for bilateral lower lung bronchiectasis with peribronchial opacities. In the right lower and medial lung, peribronchial opacities have improved since _____. There are no new opacities. Lungs are mildly hyperinflated. Heart size, mediastinal and hilar contours are normal. No pleural effusion.
Train by MIMIC-CXR Dataset	As compared to the previous radiograph, there is no relevant change. Mild atelectasis at the right lung base. Borderline size of the cardiac silhouette without pulmonary edema. No pleural effusion or pneumothorax.	Patient is status post CABG. The lungs are clear without focal consolidation or pneumothorax. No pleural effusion is seen. The cardiac and mediastinal silhouettes are stable, with the cardiac silhouette top-normal.	AP upright and lateral views of the chest provided. Overlying EKG leads are present. Mild cardiomegaly is again noted. There is mild interstitial pulmonary edema. No effusion or pneumothorax. Sternotomy wires are intact. A right upper extremity PICC line is seen with its tip in the region of the low SVC, unchanged from prior. Right superior separation of the lung is again noted.	The lungs are hyperinflated. Bilateral lower lung bronchiectasis is new since _____. There is no pleural effusion. Mediastinal and hilar contours are normal. Heart size is normal.
Train by IU X-Ray Dataset	no evidence of active disease. The cardiomeastinal silhouette and pulmonary vascularity are within normal limits. There is 'edema' or 'atelectasis' in the right base. XXXX sternotomy XXXX are again seen on the frontal view. There is a XXXX XXXX sternotomy XXXX representing a XXXX sternotomy	no evidence of active disease. The cardiomeastinal silhouette is within normal limits for size and configuration. Lungs are clear without focal consolidation, pneumothorax, or pleural effusion. XXXX are 'stable'.	1. XXXX sternotomy XXXX, XXXX. 2. Interstitial edema and/or pleural effusions. The cardiomeastinal silhouette and pulmonary vascularity are within normal limits. There is a XXXX sternotomy XXXX appearing on the lateral radiograph. There is a XXXX superior in the right midlung zone which may represent a	1. Low lung volumes with 'bronchiectasis'. 2. Mediastinal contours are within normal limits. The lungs are hyperinflated with flattening of the diaphragms, consistent with bronchiectasis. There is a XXXX sternotomy XXXX which is XXXX due to bronchiectasis. The lungs are hypoinflated and hyperinflated with flattening of the costophrenic.

Figure 41. Visualization of results generated by the proposed framework across the two datasets, using MIMIC-CXR as the test set. The figure compares two scenarios: (1) training and testing on the MIMIC-CXR dataset (column 3) and (2) training on the IU X-Ray dataset and testing on the MIMIC-CXR dataset (column 4). The visualization highlights the differences in report quality, coherence, and clinical relevance between the two scenarios, demonstrating the framework's adaptability and robustness when trained on different datasets. This comparison underscores the importance of dataset diversity in training and the framework's ability to generalize across varying clinical contexts.

5.4 Automatic Generation of Impression from Findings by Text-to-text

Large Language Models

A standard radiology report typically comprises the Indication, Comparison, Findings, and Impression sections. While the Indication and Comparison sections are informed by patient history and prior imaging, the Findings and Impression can be generated directly from image-derived features. As the Impression serves as a concise summary of the Findings, it is well-suited for automatic generation using text summarization techniques. Recent advances in large language models offer the capability to replicate radiologists' writing styles, providing a practical approach for streamlining this process.

To evaluate this capability, we generated Impression sections using both the ground-truth Findings and those produced by our framework. The evaluation followed the same qualitative and quantitative methodology described in earlier sections, and, due to time constraints, the same language model configuration as the main framework was applied for this task.

The results, shown in Table 19 and visualized in Figure 42 for IU X-Ray and Figure 43 for MIMIC-CXR, indicate that the quality of impressions generated from the framework's findings is comparable to those derived from ground-truth findings. The shorter text length of the Impression section leads to more noticeable variations in evaluation metrics. However, the overall performance demonstrates that our framework can reliably summarize key findings into concise impressions.

Despite these promising results, we recommend that radiologists verify and finalize the Impression section to ensure clinical accuracy. Given its critical role in communicating diagnostic conclusions, manual oversight remains essential, even as automated summarization can streamline the drafting process. Consequently, the automatic generation of impressions is

presented as an auxiliary feature rather than a primary component of the framework, supporting radiologists in producing high-quality reports more efficiently.

Table 19. Performance comparison of generating the Impression section from Findings in the IU X-Ray and MIMIC-CXR datasets. The table evaluates the ability of our proposed framework to generate the Impression section based on Findings, with results reported as mean \pm standard deviation and precision up to four decimal places. The “Findings generated by our framework” setting refers to Findings produced under the experimental setup described in Section 4.4 (“Experiments and Results”). This comparison highlights the framework’s effectiveness in synthesizing concise and clinically relevant Impression sections from both ground truth and framework-generated Findings, demonstrating its utility in automating radiology report generation.

Setting	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
IU X-Ray							
Test Set							
Generated from Ground-Truth Finding	0.4959 \pm 0.2494	0.4362 \pm 0.2572	0.3797 \pm 0.2678	0.3173 \pm 0.2724	0.5977 \pm 0.2979	0.2637 \pm 0.1880	1.9006 \pm 1.7117
Generated from Finding generated by our framework	0.4806 \pm 0.2567	0.4248 \pm 0.2629	0.3731 \pm 0.2688	0.3144 \pm 0.2675	0.5777 \pm 0.3092	0.2475 \pm 0.1946	1.7801 \pm 1.6598
Validation Set							
Generated from Ground-Truth Finding	0.4728 \pm 0.2364	0.3917 \pm 0.2513	0.3248 \pm 0.2638	0.2585 \pm 0.2588	0.5530 \pm 0.2885	0.2440 \pm 0.1822	1.7355 \pm 1.7544
Generated from Finding generated by our framework	0.4538 \pm 0.2409	0.3840 \pm 0.2562	0.3255 \pm 0.2646	0.2640 \pm 0.2586	0.5602 \pm 0.2840	0.2280 \pm 0.1829	1.7309 \pm 1.6639
MIMIC-CXR							
Test Set							
Generated from Ground-Truth Finding	0.3785 \pm 0.2264	0.2952 \pm 0.2265	0.2386 \pm 0.2184	0.1926 \pm 0.2052	0.4039 \pm 0.2448	0.1968 \pm 0.1558	1.0740 \pm 1.4191
Generated from Finding generated by our framework	0.3657 \pm 0.2334	0.2875 \pm 0.2348	0.2344 \pm 0.2299	0.1920 \pm 0.2163	0.4086 \pm 0.2535	0.1929 \pm 0.1615	1.1052 \pm 1.5028
Validation Set							
Generated from Ground-Truth Finding	0.4274 \pm 0.2559	0.3577 \pm 0.2612	0.3059 \pm 0.2574	0.2606 \pm 0.2440	0.4874 \pm 0.2902	0.2231 \pm 0.1854	1.3739 \pm 1.6306
Generated from Finding generated by our framework	0.4255 \pm 0.2500	0.3511 \pm 0.2549	0.2957 \pm 0.2516	0.2481 \pm 0.2392	0.4793 \pm 0.2853	0.2216 \pm 0.1819	1.3165 \pm 1.5608





Chest X-Ray Image				
Ground-Truth Findings	Normal cardiomeastinal silhouette. There is no focal consolidation. There are no XXXX of a large pleural effusion. There is no pneumothorax. There is no acute bony abnormality seen.	The heart is normal in size and contour. The lungs are clear, without evidence of infiltrate. There is no pneumothorax or effusion.	The XXXX examination consists of frontal and lateral radiographs of the chest. A total of 3 images were obtained. The cardiomeastinal contours are within normal limits allowing for low lung volumes and patient rotation. There is XXXX XXXX atelectasis. No consolidation, pleural effusion or pneumothorax. Calcified right infrahilar lymph XXXX again seen. Partially visualized lower cervical spine fusion XXXX.	Stable cardiomegaly and mediastinal contour. Increased interstitial lung markings are seen, possibly due to volume overload. There is improved aeration of the lung bases with small residual left basilar effusion. No XXXX focal consolidation or pneumothorax. Stable tunneled dialysis catheter. Visualized osseous structures appear intact.
Generated Findings	No acute cardiopulmonary abnormality. The cardiomeastinal silhouette is within normal limits. No focal consolidation, pneumothorax or large pleural effusion.	1. No evidence of active disease. The heart is normal in size and contour. The lungs are clear, without evidence of infiltrate. There is no pneumothorax or effusion.	1. Chest. No evidence of active disease. The cardiomeastinal silhouette and pulmonary vascularity are within normal limits. There is no focal consolidation, pleural effusion, or pneumothorax identified. Calcified right lower lung granulomas. Degenerative changes of the spine.	1. Stable cardiomegaly. 2. Stable right basilar interstitial opacities, compatible with chronic lung disease. The cardiomeastinal silhouette is stable in size and contour. There is a left basilar interstitial opacities, compatible with chronic lung disease. No focal consolidation, pneumothorax or large pleural effusion. Osseous structures are intact.
Ground-Truth Impression	There is no radiographic evidence of acute cardiopulmonary disease.	No acute cardiopulmonary disease.	Lung lines without evidence of acute cardiopulmonary process.	Stable cardiomegaly. Improved aeration of lung bases with persistent left basilar effusion. Prominent interstitium, possibly due to mild volume overload.
Generated Impression from Ground-Truth Findings	There is no radiographic evidence of acute cardiopulmonary disease.	No acute cardiopulmonary disease.	No evidence of acute cardiopulmonary process.	Stable cardiomegaly and mediastinal contour with improved interstitial lung markings, possibly due to volume overload.
Generated Impression from Generated Findings	There is no radiographic evidence of acute cardiopulmonary disease.	No acute cardiopulmonary disease.	No acute cardiopulmonary disease.	Stable cardiomegaly with small residual left basilar effusion. Increased interstitial lung markings, possibly due to volume overload.

Figure 42. Visualization of generated Impression sections from Findings by our proposed framework on the IU X-Ray dataset. The figure showcases examples of Impression sections generated by the framework based on both ground truth and framework-generated Findings. The visualization highlights the coherence, clinical relevance, and accuracy of the generated Impression sections, demonstrating the framework’s ability to synthesize concise summaries from detailed Findings. This example underscores the effectiveness of the framework in automating the generation of clinically useful radiology reports.





Chest X-Ray Image				
Ground-Truth Findings	Moderate to large bilateral pleural effusions are again seen, likely right greater than left. There is suspected superimposed pulmonary edema may have slightly improved since prior although detailed evaluation is limited given layering pleural effusions. Vascature appears less engorged. Cardiac silhouette cannot be assessed.	Keyword to Text: Moderate to large bilateral pleural effusions are again seen, likely right greater than left. There is suspected superimposed pulmonary edema may have slightly improved since prior although detailed evaluation is limited given layering pleural effusions. Vascature appears less engorged. Cardiac silhouette cannot be assessed.	Finding to Impression: Ill-defined patchy opacities are seen in the right lung base with an associated small right pleural effusion, which is also confirmed in the lateral view. A dense left-sided retrocardiac opacity abutting the left hemidiaphragm is unchanged since at least ____ compatible with a Bochdalek hernia. A small left pleural effusion is also likely present. There is biapical pleuro-parenchymal scarring, more conspicuous in the left apex. No other focal opacities are identified. Mild cardiomegaly is unchanged from prior. There is no pneumothorax.	Finding to Impression: Since the prior radiograph, no significant change in the widespread parenchymal opacities and moderate cardiomegaly. No change in the left the cavoatrial junction, and right pacemaker lead in the right ventricle. No new focal consolidation or larger pleural effusions.
Generated Findings	[Keyword to Text: There are moderate bilateral pleural effusions, right greater than left. The pulmonary vasculature is engorged and there is mild pulmonary edema.]	Finding to Impression: Subtle patchy opacity along the left heart border on the frontal view, not substantiated on the lateral view, may be due to atelectasis/ scarring or epicardial fat pad, less likely consolidation. No focal consolidation seen elsewhere. There is no pleural effusion or pneumothorax. Cardiac and mediastinal silhouettes are stable. Hilar contours are stable. No overt pulmonary edema is seen. Chronic changes at the right acromioclavicular joint are not well assessed.	[Keyword to Text: There is a new focal opacity in the left lower lung, partially obscuring the left hemidiaphragm. Mild scarring at the right lung apex is unchanged since _____. The cardiomeastrial silhouette and hilar contours are unchanged. There is no pleural effusion or pneumothorax. A small hiatal hernia is noted.]	[Keyword to Text: A Port-A-Cath terminates at the cavoatrial junction. Dual lead pacemaker is again noted. No new focal consolidation is seen.]
Ground-Truth Impression	Mild to large bilateral, right greater than left pleural effusions. Degree of pulmonary edema may have slightly improved since prior exam although detailed evaluation is limited.	Subtle patchy opacity along the left heart border on the frontal view, not substantiated on the lateral view, may be due to atelectasis/ scarring or epicardial fat pad, less likely consolidation.	Right lower lobe pneumonia. Small bilateral pleural effusions.	No significant change in the widespread parenchymal opacities and moderate cardiomegaly. No larger pleural effusions.
Generated Impression from Ground-Truth Findings	[Moderate to large bilateral pleural effusions, likely right greater than left. Suspected superimposed pulmonary edema may have slightly improved since prior although detailed evaluation is limited given layering pleural effusions.]	[Subtle patchy opacity along the left heart border on the frontal view, not substantiated on the lateral view, may be due to atelectasis/ scarring or epicardial fat pad, less likely consolidation.]	[Right basilar opacity compatible with pneumonia. Small bilateral pleural effusions. Bochdalek hernia.]	[No significant interval change in widespread parenchymal opacities and moderate cardiomegaly.]
Generated Impression from Generated Findings	[Moderate to large bilateral pleural effusions with possible slight improvement in superimposed pulmonary edema.]	[Subtle patchy opacity along the left heart border on the frontal view, not substantiated on the lateral view, may be due to atelectasis/scarring or epicardial fat pad, less likely consolidation.]	[1. Patchy ill-defined opacities in the right lung base with associated small right pleural effusion, concerning for infection. 2. Stable dense left retrocardiac opacity compatible with Bochdalek hernia.]	[No significant interval change.]

Figure 43. Visualization of generated Impression sections from Findings by our proposed framework on the MIMIC-CXR dataset. The figure presents examples of Impression sections generated by the framework based on both ground truth and framework-generated Findings. The visualization emphasizes the coherence, clinical relevance, and accuracy of the generated Impression sections, illustrating the framework's capability to produce concise and actionable summaries from detailed Findings. This example further validates the framework's robustness and applicability across diverse datasets in radiology report generation.

6. Conclusion

This study presents a novel framework for radiology report generation that integrates automatic keyword adaptation and frequency-based multi-label classification to improve both performance and transparency. By replacing traditional black-box visual features with interpretable keyword lists, our approach enhances explainability while reducing errors inherent in conventional methods. Extensive experiments on the IU X-Ray and MIMIC-CXR datasets demonstrate the superiority of our framework over state-of-the-art methods across all key evaluation metrics.

Prior studies in chest X-ray image analysis often focus on narrowing the scope of target tasks, such as lung region segmentation to isolate infection-prone areas ([187], [188], [189]). Similarly, our framework employs a generalizable strategy by utilizing extracted keywords as the starting point for radiology report generation. These keywords, refined through the RadLex dictionary and prioritized using a frequency-based multi-label classification strategy, ensure clinical relevance while balancing computational efficiency. This approach aligns with the principle of "Garbage in, Garbage out" ([190]) underscoring the importance of high-quality, context-appropriate inputs for reliable outputs.

Our findings also highlight the potential and limitations of commercial large language models (LLMs), such as ChatGPT, in medical vision-language processing. While these models offer efficient pipelines for generating reports, their performance heavily depends on large datasets, which are common in natural image contexts but scarce in medical domains like chest X-ray reporting. Additionally, high-resolution imaging modalities such as pathology imaging ([191]) can provide sufficient data through slicing techniques, but chest X-ray datasets paired with

radiology reports remain relatively small in scale, limiting the generalizability of LLMs in this area.

Further challenges arise from the cost, computational demands, and lack of explainability of commercial LLMs. Many LLMs rely on external servers, raising concerns about data privacy and integration with clinical workflows. Moreover, the opaque nature of these models makes it difficult for radiologists to validate the logic or evidence behind generated reports, often relegating them to post-generation editing tasks ([192]). This not only increases workload but also discourages adoption of AI-assisted workflows in favor of manual report drafting.

In contrast, our proposed framework addresses these challenges by introducing a transparent intermediate step: generating interpretable keyword lists. These lists allow radiologists to validate and refine extracted features before finalizing reports which is shown in Figure 44, fostering a collaborative workflow that enhances usability and aligns with clinical needs. Additionally, the framework promotes clarity and adaptability within the domain of radiology informatics.

Although our framework achieved high performance, it inevitably introduced a longer generation time compared with the conventional single-stage encoder–decoder structure. This increase in processing time stems from the multi-step design, which, while improving the quality of the generated reports, adds additional computation. We consider that, in the context of automatic radiology report generation, prioritizing accuracy over minimal generation time is more valuable for clinical applicability.

Furthermore, due to the limitations of the available datasets, our framework has not yet undergone external validation in real-world clinical settings. As a result, we were unable to fully assess its robustness and generalizability when applied to diverse patient populations

and imaging protocols. Conducting such external validation will be an important focus of our future work.

Looking forward, we plan to validate our framework in real-world clinical settings through external validations and stress tests, assessing its robustness under diverse conditions. A critical focus will be on ensuring semantic consistency between the generated keywords and the corresponding radiology report sentences. Preliminary findings reveal occasional semantic mismatches, where generated text misrepresents or contradicts the intended meaning of the keywords. Inspired by prior research ([193]), we will explore techniques to align keyword semantics with generated sentences to address this issue. By continuing to refine and adapt our framework, we aim to advance automated radiology report generation and foster trust and adoption of AI-driven medical solutions.

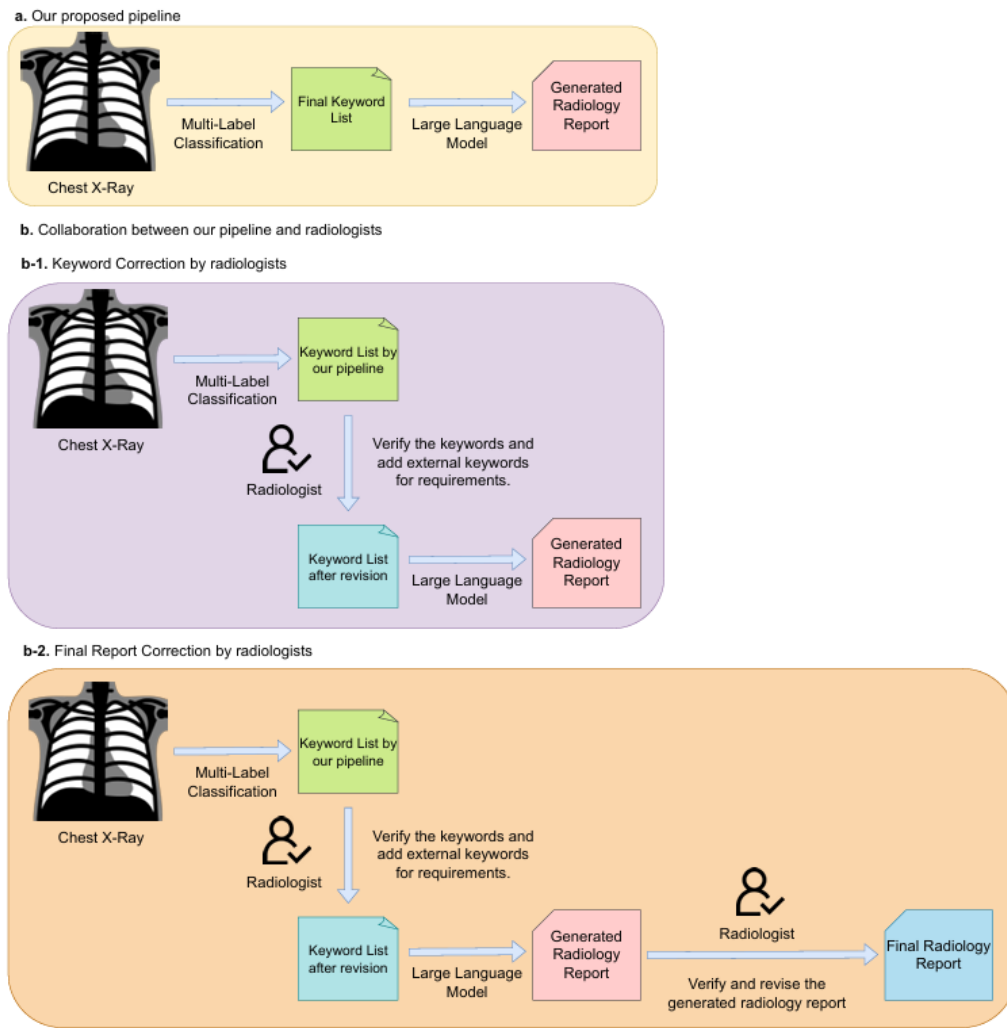


Figure 44. Proposed future collaboration workflows between the pipeline and radiologists for refining keywords and finalizing radiology reports. The diagram illustrates three approaches: the proposed pipeline (a) and two collaborative workflows (b that contain b-1 and b-2). In the proposed pipeline (a), medical imaging data is processed for keyword extraction, undergoing automatic verification and refinement before being input into a pretrained large language model (LLM) to generate clinically relevant reports. In the (b-1) workflow, the process is enhanced by radiologists double-checking the refined keywords before they are input into the LLM for report generation. In the (b-2) workflow, the refined keywords undergo the same process as (b-1), but the generated reports are further reviewed and revised by radiologists to produce the final radiology report, ensuring the highest quality and clinical accuracy.

7. Reference

1. Jing, B., P. Xie, and E. Xing, *On the automatic generation of medical imaging reports*. arXiv preprint arXiv:1711.08195, 2017.
2. Xue, Y., et al. *Multimodal recurrent model with attention for automated radiology report generation*. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. 2018. Springer.
3. Harzig, P., et al., *Addressing data bias problems for chest x-ray image report generation*. arXiv preprint arXiv:1908.02123, 2019.
4. Xie, X., et al. *Attention-based abnormal-aware fusion network for radiology report generation*. in *Database Systems for Advanced Applications: DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22–25, 2019, Proceedings 24*. 2019. Springer.
5. Yuan, J., et al. *Automatic radiology report generation based on multi-view image fusion and medical concept enrichment*. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. 2019. Springer.
6. Li, C.Y., et al. *Knowledge-driven encode, retrieve, paraphrase for medical image report generation*. in *Proceedings of the AAAI conference on artificial intelligence*. 2019.
7. Jing, B., Z. Wang, and E. Xing, *Show, describe and conclude: On exploiting the structure information of chest x-ray reports*. arXiv preprint arXiv:2004.12274, 2020.
8. Chen, Z., et al., *Generating radiology reports via memory-driven transformer*. arXiv preprint arXiv:2010.16056, 2020.
9. Zhang, Y., et al. *When radiology report generation meets knowledge graph*. in *Proceedings of the AAAI conference on artificial intelligence*. 2020.
10. Wang, Z., et al. *A self-boosting framework for automated radiographic report generation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
11. Alfarghaly, O., et al., *Automated radiology report generation using conditioned transformers*. *Informatics in Medicine Unlocked*, 2021. **24**: p. 100557.
12. Liu, F., et al., *Contrastive attention for automatic chest x-ray report generation*. arXiv preprint arXiv:2106.06965, 2021.
13. Liu, F., et al. *Exploring and distilling posterior and prior knowledge for radiology report generation*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
14. Yang, S., et al., *Knowledge matters: Chest radiology report generation with general and specific knowledge*. *Medical image analysis*, 2022. **80**: p. 102510.
15. Nooralahzadeh, F., et al., *Progressive transformer-based generation of radiology reports*. arXiv preprint arXiv:2102.09777, 2021.
16. Yang, S., et al., *Radiology report generation with a learned knowledge base and multi-modal alignment*. *Medical Image Analysis*, 2023. **86**: p. 102798.
17. Hou, B., et al. *Ratchet: Medical transformer for chest x-ray diagnosis and reporting*. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. 2021. Springer.

18. Zhou, Y., et al. *Visual-textual attentive semantic consistency for medical report generation*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
19. Yan, A., et al., *Weakly supervised contrastive learning for chest x-ray report generation*. arXiv preprint arXiv:2109.12242, 2021.
20. Wang, Z., et al. *A medical semantic-assisted transformer for radiographic report generation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022. Springer.
21. Li, J., et al. *A self-guided framework for radiology report generation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022. Springer.
22. You, D., et al. *Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation*. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III* 24. 2021. Springer.
23. Wang, L., et al. *An inclusive task-aware framework for radiology report generation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022. Springer.
24. Sirshar, M., et al., *Attention based automated radiology report generation using CNN and LSTM*. Plos one, 2022. **17**(1): p. e0262209.
25. Yan, S., et al., *Attributed abnormality graph embedding for clinically accurate x-ray report generation*. IEEE Transactions on Medical Imaging, 2023. **42**(8): p. 2211-2222.
26. Wang, Z., et al., *Automated radiographic report generation purely on transformer: A multicriteria supervised approach*. IEEE Transactions on Medical Imaging, 2022. **41**(10): p. 2803-2813.
27. Yu, H. and Q. Zhang. *Clinically coherent radiology report generation with imbalanced chest x-rays*. in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2022. IEEE.
28. Chen, Z., et al., *Cross-modal memory networks for radiology report generation*. arXiv preprint arXiv:2204.13258, 2022.
29. Wang, J., A. Bhalerao, and Y. He. *Cross-modal prototype driven network for radiology report generation*. in *European Conference on Computer Vision*. 2022. Springer.
30. Nishino, T., et al. *Factual accuracy is not enough: Planning consistent description order for radiology report generation*. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022.
31. Nicolson, A., J. Dowling, and B. Koopman, *Improving chest X-ray report generation by leveraging warm starting*. Artificial intelligence in medicine, 2023. **144**: p. 102633.
32. Delbrouck, J.-B., et al., *Improving the factual correctness of radiology report generation with semantic rewards*. arXiv preprint arXiv:2210.12186, 2022.
33. You, J., et al. *Jpg-jointly learn to align: Automated disease prediction and radiology report generation*. in *Proceedings of the 29th international conference on computational linguistics*. 2022.
34. Wu, X., et al., *Multimodal contrastive learning for radiology report generation*. Journal of Ambient Intelligence and Humanized Computing, 2023. **14**(8): p. 11185-11194.

35. Dalla Serra, F., et al. *Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations*. in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2022.
36. Yan, B., et al., *Prior guided transformer for accurate radiology reports generation*. *IEEE Journal of Biomedical and Health Informatics*, 2022. **26**(11): p. 5631-5640.
37. Wang, S., et al., *Prior knowledge enhances radiology report generation*. *AMIA Summits on Translational Science Proceedings*, 2022. **2022**: p. 486.
38. Qin, H. and Y. Song. *Reinforced cross-modal alignment for radiology report generation*. in *Findings of the Association for Computational Linguistics: ACL 2022*. 2022.
39. Wang, Y., et al. *Self adaptive global-local feature enhancement for radiology report generation*. in *2023 IEEE International Conference on Image Processing (ICIP)*. 2023. IEEE.
40. Kong, M., et al. *Transq: Transformer-based semantic query for medical report generation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022. Springer.
41. Li, M., et al. *Dynamic graph enhanced contrastive learning for chest x-ray report generation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
42. Tanida, T., et al. *Interactive and explainable region-guided radiology report generation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
43. Yang, Y., et al., *Joint embedding of deep visual and semantic features for medical image report generation*. *IEEE Transactions on Multimedia*, 2021. **25**: p. 167-178.
44. Kale, K., et al. *KGVL-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation*. in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023.
45. Huang, Z., X. Zhang, and S. Zhang. *Kiut: Knowledge-injected u-transformer for radiology report generation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
46. Wang, Z., et al. *Metransformer: Radiology report generation by transformer with multiple learnable expert tokens*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
47. Hou, W., et al., *ORGAN: observation-guided radiology report generation via tree reasoning*. arXiv preprint arXiv:2306.06466, 2023.
48. Wang, Z., et al., *R2gengpt: Radiology report generation with frozen llms*. *Meta-Radiology*, 2023. **1**(3): p. 100033.
49. Kale, K. and K. Jadhav, *Replace and report: NLP assisted radiology report generation*. arXiv preprint arXiv:2306.17180, 2023.
50. Yan, B., et al., *Style-aware radiology report generation with radgraph and few-shot prompting*. arXiv preprint arXiv:2310.17811, 2023.
51. Li, Y., et al. *Unify, align and refine: Multi-level semantic alignment for radiology report generation*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
52. Mohsan, M.M., et al., *Vision transformer and language model based radiology report generation*. *IEEE Access*, 2022. **11**: p. 1814-1824.
53. Chen, W., et al., *Visual-linguistic causal intervention for radiology report generation*. arXiv preprint arXiv:2303.09117, 2023. **1**(8).

54. Zhang, K., et al., *Semi-supervised medical report generation via graph-guided hybrid feature consistency*. IEEE Transactions on Multimedia, 2023. **26**: p. 904-915.
55. Liu, C., et al. *Bootstrapping Large Language Models for Radiology Report Generation*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024.
56. Zhou, Z., et al., *Large Model driven Radiology Report Generation with Clinical Quality Reinforcement Learning*. arXiv preprint arXiv:2403.06728, 2024.
57. Yi, X., et al., *TSGET: Two-Stage Global Enhanced Transformer for Automatic Radiology Report Generation*. IEEE Journal of Biomedical and Health Informatics, 2024.
58. Parres, D., A. Albiol, and R. Paredes, *Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation*. Bioengineering, 2024. **11**(4): p. 351.
59. Zhang, K., et al., *Attribute Prototype-guided Iterative Scene Graph for Explainable Radiology Report Generation*. IEEE Transactions on Medical Imaging, 2024.
60. Yi, X., et al., *LHR-RFL: Linear Hybrid-Reward-Based Reinforced Focal Learning for Automatic Radiology Report Generation*. IEEE Transactions on Medical Imaging, 2024. **44**: p. 1494-1504.
61. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks*. Communications of the ACM, 2012. **60**: p. 84 - 90.
62. Simonyan, K., *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556, 2014.
63. Szegedy, C., et al., *Going deeper with convolutions*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: p. 1-9.
64. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
65. Ioffe, S. and C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. ArXiv, 2015. **abs/1502.03167**.
66. Szegedy, C., et al., *Rethinking the Inception Architecture for Computer Vision*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: p. 2818-2826.
67. Xie, S., et al. *Aggregated residual transformations for deep neural networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
68. Zagoruyko, S. and N. Komodakis, *Wide Residual Networks*. ArXiv, 2016. **abs/1605.07146**.
69. Huang, G., Z. Liu, and K.Q. Weinberger, *Densely Connected Convolutional Networks*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: p. 2261-2269.
70. Szegedy, C., et al., *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. ArXiv, 2016. **abs/1602.07261**.
71. Hu, J., et al., *Squeeze-and-Excitation Networks*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: p. 7132-7141.
72. Zhang, X., et al., *ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: p. 6848-6856.
73. Sandler, M., et al., *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: p. 4510-4520.
74. Ma, N., et al., *ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design*. ArXiv, 2018. **abs/1807.11164**.

75. Howard, A.G., et al., *Searching for MobileNetV3*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: p. 1314-1324.
76. Gao, S.-H., et al., *Res2net: A new multi-scale backbone architecture*. IEEE transactions on pattern analysis and machine intelligence, 2019. **43**(2): p. 652-662.
77. Tan, M. and Q.V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. ArXiv, 2019. **abs/1905.11946**.
78. Sun, K., et al., *Deep High-Resolution Representation Learning for Human Pose Estimation*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: p. 5686-5696.
79. Wang, C.-Y., et al., *CSPNet: A New Backbone that can Enhance Learning Capability of CNN*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019: p. 1571-1580.
80. Dosovitskiy, A., et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ArXiv, 2020. **abs/2010.11929**.
81. Radosavovic, I., et al., *Designing Network Design Spaces*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: p. 10425-10433.
82. Touvron, H., et al. *Training data-efficient image transformers & distillation through attention*. in *International Conference on Machine Learning*. 2020.
83. Liu, Z., et al., *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: p. 9992-10002.
84. Ding, X., et al., *RepVGG: Making VGG-style ConvNets Great Again*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: p. 13728-13737.
85. Yuan, L., et al., *Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: p. 538-547.
86. Peng, Z., et al., *Conformer: Local Features Coupling Global Representations for Visual Recognition*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: p. 357-366.
87. Chu, X., et al. *Twins: Revisiting the Design of Spatial Attention in Vision Transformers*. in *Neural Information Processing Systems*. 2021.
88. Yu, W., et al., *MetaFormer is Actually What You Need for Vision*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: p. 10809-10819.
89. Li, Y., et al., *MViTv2: Improved Multiscale Vision Transformers for Classification and Detection*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: p. 4794-4804.
90. Liu, Z., et al., *Swin Transformer V2: Scaling Up Capacity and Resolution*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: p. 11999-12009.
91. Bao, H., L. Dong, and F. Wei, *BEiT: BERT Pre-Training of Image Transformers*. ArXiv, 2021. **abs/2106.08254**.
92. Radford, A., et al. *Learning Transferable Visual Models From Natural Language Supervision*. in *International Conference on Machine Learning*. 2021.
93. Tan, M. and Q.V. Le. *EfficientNetV2: Smaller Models and Faster Training*. in *International Conference on Machine Learning*. 2021.
94. El-Nouby, A., et al. *XCiT: Cross-Covariance Image Transformers*. in *Neural Information Processing Systems*. 2021.

95. Chen, X., S. Xie, and K. He, *An Empirical Study of Training Self-Supervised Vision Transformers*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: p. 9620-9629.
96. He, K., et al., *Masked Autoencoders Are Scalable Vision Learners*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: p. 15979-15988.
97. Wu, K., et al., *TinyViT: Fast Pretraining Distillation for Small Vision Transformers*. ArXiv, 2022. **abs/2207.10666**.
98. Liu, Z., et al., *A ConvNet for the 2020s*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: p. 11966-11976.
99. Guo, M.-H., et al., *Visual attention network*. Computational Visual Media, 2022. **9**: p. 733 - 752.
100. Trockman, A. and J.Z. Kolter, *Patches Are All You Need?* Trans. Mach. Learn. Res., 2022. **2023**.
101. Maaz, M., et al. *EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications*. in *ECCV Workshops*. 2022.
102. Anasosalu Vasu, P.K., et al., *MobileOne: An Improved One millisecond Mobile Backbone*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: p. 7907-7917.
103. Li, Y., et al., *EfficientFormer: Vision Transformers at MobileNet Speed*. ArXiv, 2022. **abs/2206.01191**.
104. Peng, Z., et al., *BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers*. ArXiv, 2022. **abs/2208.06366**.
105. Li, J., et al. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. in *International Conference on Machine Learning*. 2022.
106. Woo, S., et al., *ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: p. 16133-16142.
107. Li, J., et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. in *International Conference on Machine Learning*. 2023.
108. Sakthivel-Wainford, K., *Chest X-ray Interpretation for Radiographers, Nurses and Allied Health Professionals*. 2019: M&K Update Limited.
109. Cox, C.W., C.S. Rose, and D.A. Lynch, *State of the art: Imaging of occupational lung disease*. Radiology, 2014. **270 3**: p. 681-96.
110. Jacobi, A., et al., *Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review*. Clinical Imaging, 2020. **64**: p. 35 - 42.
111. United, N. *sources and effects of ionizing radiation*. 2000.
112. Akhter, Y., R. Singh, and M. Vatsa, *AI-based radiodiagnosis using chest X-rays: A review*. Frontiers in Big Data, 2023. **6**.
113. Graf, M.M., K.K. Bressem, and L.C. Adams, *[Transformation of free-text radiology reports into structured data]*. Radiologie, 2025.
114. Jiang, H., et al., *Transforming free-text radiology reports into structured reports using ChatGPT: A study on thyroid ultrasonography*. European journal of radiology, 2024. **175**: p. 111458.
115. Ganeshan, D., et al., *Structured Reporting in Radiology*. Academic radiology, 2018. **25 1**: p. 66-73.
116. Weiss, D.L. and C. Langlotz, *Structured reporting: patient care enhancement or productivity nightmare?* Radiology, 2008. **249 3**: p. 739-47.

117. Cowan, I.A., S.L.S. MacDonald, and R.A. Floyd, *Measuring and managing radiologist workload: Measuring radiologist reporting times using data from a Radiology Information System*. Journal of Medical Imaging and Radiation Oncology, 2013. **57**.
118. Cadth, *Canadian Medical Imaging Inventory 2022–2023: The Medical Imaging Team*. Canadian Journal of Health Technologies, 2024.
119. Mayor, S., *Waiting times for x ray results in England are increasing, figures show*. BMJ : British Medical Journal, 2015. **350**.
120. Woznitza, N., et al., *Impact of radiographer immediate reporting of chest x-rays from general practice on the lung cancer pathway (radioX)*. Lung Cancer, 2019.
121. Lakhani, P., et al., *Machine Learning in Radiology: Applications Beyond Image Interpretation*. Journal of the American College of Radiology : JACR, 2017. **15 2**: p. 350-359.
122. Vaswani, A., *Attention is all you need*. Advances in Neural Information Processing Systems, 2017.
123. Sutskever, I., O. Vinyals, and Q.V. Le, *Sequence to Sequence Learning with Neural Networks*. ArXiv, 2014. **abs/1409.3215**.
124. Demner-Fushman, D., et al., *Preparing a collection of radiology examinations for distribution and retrieval*. Journal of the American Medical Informatics Association, 2016. **23(2)**: p. 304-310.
125. Johnson, A.E., et al., *MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports*. Scientific data, 2019. **6(1)**: p. 317.
126. Radford, A. and K. Narasimhan. *Improving Language Understanding by Generative Pre-Training*. 2018.
127. Kenton, J.D.M.-W.C. and L.K. Toutanova. *Bert: Pre-training of deep bidirectional transformers for language understanding*. in *Proceedings of naacL-HLT*. 2019. Minneapolis, Minnesota.
128. Monajatipoor, M., et al. *Berthop: An effective vision-and-language model for chest x-ray disease diagnosis*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2022. Springer.
129. Huang, S.-C., et al. *Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
130. Wang, Z., et al., *Medclip: Contrastive learning from unpaired medical images and text*. arXiv preprint arXiv:2210.10163, 2022.
131. Cong, F., et al. *Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension*. in *Proceedings of the 30th ACM International Conference on Multimedia*. 2022.
132. Li, Y., H. Wang, and Y. Luo. *A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports*. in *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*. 2020. IEEE.
133. Müller, P., et al. *Joint learning of localized representations from medical images and reports*. in *European Conference on Computer Vision*. 2022. Springer.
134. Huang, S.-C., et al., *GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: p. 3922-3931.
135. Boecking, B., et al., *Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing*. ArXiv, 2022. **abs/2204.09817**.

136. Zhang, K., et al., *Multi-Task Paired Masking With Alignment Modeling for Medical Vision-Language Pre-Training*. IEEE Transactions on Multimedia, 2023. **26**: p. 4706-4721.
137. Lei, Y., et al., *CLIP-Lung: Textual Knowledge-Guided Lung Nodule Malignancy Prediction*. ArXiv, 2023. **abs/2304.08013**.
138. Tiu, E., et al., *Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning*. Nature Biomedical Engineering, 2022. **6**: p. 1399 - 1406.
139. Lai, H., et al., *CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification*. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: p. 11137-11146.
140. Eslami, S., C. Meinel, and G. De Melo. *PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?* in *Findings*. 2023.
141. Zhang, X., et al., *PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering*. ArXiv, 2023. **abs/2305.10415**.
142. Sonsbeek, T.v., et al. *Open-Ended Medical Visual Question Answering Through Prefix Tuning of Language Models*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2023.
143. Gan, Z., et al., *Vision-Language Pre-training: Basics, Recent Advances, and Future Trends*. ArXiv, 2022. **abs/2210.09263**.
144. Hartung, M.P., et al., *How to create a great radiology report*. Radiographics, 2020. **40**(6): p. 1658-1670.
145. Boland, G., A. Guimaraes, and P. Mueller, *Radiology report turnaround: expectations and solutions*. European Radiology, 2008. **18**: p. 1326-1328.
146. Monshi, M.M.A., J. Poon, and V. Chung, *Deep learning in generating radiology reports: A survey*. Artificial Intelligence in Medicine, 2020. **106**: p. 101878.
147. Jain, S., et al., *Radgraph: Extracting clinical entities and relations from radiology reports*. arXiv preprint arXiv:2106.14463, 2021.
148. Soleimani, M., et al., *Practical evaluation of ChatGPT performance for radiology report generation*. Academic Radiology, 2024.
149. Iturralde, M.P., *Dictionary and handbook of nuclear medicine and clinical imaging*. 2018: CRC Press.
150. Reliyanti, E. and O. Damayanti, *Composing Radiographic Dictionary for Radiology Students and Radiographers*. International Journal of Ethno-Sciences and Education Research, 2023. **3**(4): p. 121-126.
151. Langlotz, C., *RadLex: a new method for indexing online educational materials*. Radiographics : a review publication of the Radiological Society of North America, Inc, 2006. **26** **6**: p. 1595-7.
152. Wu, P.-H., et al., *Keyword extraction and structuralization of medical reports*. Health Information Science and Systems, 2020. **8**: p. 1-25.
153. Kim, Y., et al., *Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records*. Scientific reports, 2020. **10**(1): p. 20265.
154. Grootendorst, M., *KeyBERT: Minimal keyword extraction with BERT*. 2020.
155. Woo, S., et al., *CBAM: Convolutional Block Attention Module*. ArXiv, 2018. **abs/1807.06521**.
156. Contributors, M., *OpenMMLab's Pre-training Toolbox and Benchmark*. 2023.
157. Read, J., et al., *Classifier chains for multi-label classification*. Machine learning, 2011. **85**: p. 333-359.

158. Yang, P., et al., *SGM: sequence generation model for multi-label classification*. arXiv preprint arXiv:1806.04822, 2018.
159. Wehrmann, J., R. Cerri, and R. Barros. *Hierarchical multi-label classification networks*. in *International conference on machine learning*. 2018. PMLR.
160. Liu, S., et al., *Query2label: A simple transformer way to multi-label classification*. arXiv preprint arXiv:2107.10834, 2021.
161. You, R., et al. *Cross-modality attention with semantic graph embedding for multi-label classification*. in *Proceedings of the AAAI conference on artificial intelligence*. 2020.
162. Chen, S.-F., et al. *Order-free rnn with visual attention for multi-label classification*. in *Proceedings of the AAAI conference on artificial intelligence*. 2018.
163. Yuan, J., et al., *Graph attention transformer network for multi-label image classification*. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. **19**(4): p. 1-16.
164. Ridnik, T., et al. *Asymmetric loss for multi-label classification*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
165. Wu, T., et al. *Distribution-balanced loss for multi-label classification in long-tailed datasets*. in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. 2020. Springer.
166. Durand, T., N. Mehrasa, and G. Mori. *Learning a deep convnet for multi-label classification with partial labels*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
167. Chen, B., et al., *Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification*. *IEEE journal of biomedical and health informatics*, 2020. **24**(8): p. 2292-2302.
168. Ma, C., H. Wang, and S.C. Hoi. *Multi-label thoracic disease image classification with cross-attention networks*. in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. 2019. Springer.
169. Ge, Z., et al., *Chest x-rays classification: A multi-label and fine-grained problem*. arXiv preprint arXiv:1807.07247, 2018.
170. Wolf, T., *Huggingface's transformers: State-of-the-art natural language processing*. arXiv preprint arXiv:1910.03771, 2019.
171. Peters, M.E., et al., *Deep Contextualized Word Representations*. ArXiv, 2018. **abs/1802.05365**.
172. Chu, Z., et al., *History, Development, and Principles of Large Language Models-An Introductory Survey*. ArXiv, 2024. **abs/2402.06853**.
173. Gao, K., et al., *Examining User-Friendly and Open-Sourced Large GPT Models: A Survey on Language, Multimodal, and Scientific GPT Models*. ArXiv, 2023. **abs/2308.14149**.
174. Lipscomb, C.E., *Medical subject headings (MeSH)*. *Bulletin of the Medical Library Association*, 2000. **88**(3): p. 265.
175. Raffel, C., et al., *Exploring the limits of transfer learning with a unified text-to-text transformer*. *Journal of machine learning research*, 2020. **21**(140): p. 1-67.
176. Lu, Q., D. Dou, and T. Nguyen. *ClinicalT5: A generative language model for clinical text*. in *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022.
177. Langlotz, C.P., *RadLex: a new method for indexing online educational materials*. 2006, Radiological Society of North America. p. 1595-1597.

178. Liu, Z., et al. *A convnet for the 2020s*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
179. Papineni, K., et al. *Bleu: a method for automatic evaluation of machine translation*. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.
180. Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries*. in *Text summarization branches out*. 2004.
181. Vedantam, R., C. Lawrence Zitnick, and D. Parikh. *Cider: Consensus-based image description evaluation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
182. Banerjee, S. and A. Lavie. *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005.
183. Irvin, J.A., et al. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. in *AAAI Conference on Artificial Intelligence*. 2019.
184. Touvron, H., et al., *LLaMA: Open and Efficient Foundation Language Models*. ArXiv, 2023. **abs/2302.13971**.
185. Bai, J., et al., *Qwen Technical Report*. ArXiv, 2023. **abs/2309.16609**.
186. Lu, H., et al., *DeepSeek-VL: Towards Real-World Vision-Language Understanding*. ArXiv, 2024. **abs/2403.05525**.
187. He, Z., A.N.N. Wong, and J.S. Yoo, *Co-ERA-Net: Co-Supervision and Enhanced Region Attention for Accurate Segmentation in COVID-19 Chest Infection Images*. Bioengineering, 2023. **10**.
188. Gao, K., et al., *Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images*. Medical Image Analysis, 2020. **67**: p. 101836 - 101836.
189. Paluru, N., et al., *Anam-Net: Anamorphic Depth Embedding-Based Lightweight CNN for Segmentation of Anomalies in COVID-19 Chest CT Images*. Ieee Transactions on Neural Networks and Learning Systems, 2021. **32**: p. 932 - 946.
190. Kilkenny, M.F. and K. Robinson, *Data quality: "Garbage in – garbage out"*. Health Information Management Journal, 2018. **47**: p. 103 - 105.
191. Wang, X., et al., *A pathology foundation model for cancer diagnosis and prognosis prediction*. Nature, 2024.
192. Tanno, R., et al., *Collaboration between clinicians and vision-language models in radiology report generation*. Nature medicine, 2024.
193. Chen, J.C.-Y., et al., *MAGICoRe: Multi-Agent, Iterative, Coarse-to-Fine Refinement for Reasoning*. ArXiv, 2024. **abs/2409.12147**.