Computers in Biology and Medicine 196 (2025) 110625

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Radiology report generation using automatic keyword adaptation, frequency-based multi-label classification and text-to-text large language models

Zebang He[a], Alex Ngai Nick Wong[b], Jung Sun Yoo[a,*]

[a] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong Special Administrative Region of China
[b] DOBI Medical International Inc., Hangzhou, China

ARTICLE INFO

Keywords:
Radiology report generation
Automatic keyword adaptation
Frequency-based multi-label classification
Large language model

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 referring physicians. However, the manual generation of radiology 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.

Methods: 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 conclusion, we developed a novel deep-learning based radiology report generation method for preparing high-quality and explainable radiology report for chest X-ray images using the multi-label classification and a text-to-text large language model. Our method could address the lack of explainability in the current workflow and provide a clear and flexible automated pipeline to reduce the workload of radiologists and support the further applications related to Human–AI interactive communications.

* Corresponding author.
E-mail addresses: zebang.he@connect.polyu.hk (Z. He), axwong93@gmail.com (A.N.N. Wong), jungsun.yoo@polyu.edu.hk (J.S. Yoo).

https://doi.org/10.1016/j.compbiomed.2025.110625
Received 4 March 2025; Received in revised form 7 May 2025; Accepted 16 June 2025
Available online 3 July 2025

1. Introduction

Radiology reports are a critical component of medical imaging, serving as the primary communication medium between radiologists and referring clinicians. These reports offer essential insights for diagnosis, treatment planning, and ongoing patient management. The accuracy, clarity, and timeliness of radiology reports directly influence clinical decision-making and patient outcomes. However, the manual drafting of radiology reports remains a time-intensive and laborious task, particularly as imaging volumes continue to rise. On average, composing a chest X-ray radiology report takes approximately 1 min and 38 s to 2 min per case [1]. This burden is further exacerbated by the global shortage of radiologists, creating significant bottlenecks in healthcare systems.

These challenges are especially pronounced in countries with high clinical demand and limited imaging resources, such as Canada [2] and the UK [3]. In such settings, patients may wait up to 32 days from the time of imaging to receive a diagnosis, largely due to reporting delays and strained radiology departments. These circumstances underscore a growing demand for automated solutions that can expedite report generation and alleviate workforce pressures.

Artificial intelligence (AI)-assisted automatic radiology report generation has emerged as a promising approach to address these challenges. By automatically translating radiological findings into structured and clinically relevant narratives, such systems have the potential to significantly reduce reporting time, streamline clinical workflows, and enhance diagnostic efficiency. Compared to manual report drafting, which takes several minutes [1], automated systems can generate reports in less than one second per case. This rapid turnaround time enables near-instantaneous availability of reports, which has been shown to reduce diagnostic delays by nearly half when immediate reporting is implemented [4].

Despite these benefits, existing deep learning-based report generation methods—especially those built on encoder–decoder architectures—still face key limitations. Chief among them is the lack of transparency and interpretability. The high-level features extracted by visual encoders are often abstract and not clinically meaningful, leading to errors, irrelevant details, or missing critical information in the generated text. This "black-box" nature of current models limits clinicians' ability to validate or trust the generated reports, presenting a significant barrier to their integration in clinical practice.

To address these challenges, we propose a novel framework for radiology report generation that emphasizes explainability, accuracy, and adaptability. Fig. 1 illustrates the core motivation behind our approach, highlighting the transition from unexplainable features to an interpretable, controlled keyword list that serves as the foundation for report generation. Central to our framework is the systematic multi-label classification process that generates an interpretable keyword list. Initially, general language models are used to extract relevant keywords from medical imaging data. These keywords are then verified using a radiology-specific dictionary, such as RadLex, and ranked by frequency to cluster them into meaningful categories for multi-label classification. By changing the report generation process from the unexplainable features to the controllable keyword list, our approach not only enhanced the interpretability by also mitigate the impact of uncontrollable features, which is a key aspect of the "Garbage in – garbage out" principle, to mislead the report generation. This principle emphasizes that the quality of input data directly affects the quality of the generated output. By filtering out unrelated or erroneous radiological information, our framework ensures that only relevant, accurate data are fed into the system, thereby improving the quality of the automatically generated reports.

The "Garbage in – garbage out" concept [5], widely recognized in health data management, underscores the importance of accurate, valid, and comprehensive data collection. Poor input data inevitably leads to unreliable outputs, as seen in clinical documentation and healthcare analytics. By ensuring that our input data—keywords extracted from radiological images and verified using domain-specific dictionaries—is precise and relevant, we enhance both the quality control of report generation and the transparency of the AI workflow. This transparency enables medical professionals, particularly radiologists, to easily visualize and refine the keyword list, facilitating quality assurance and the potential for manual adjustments to improve report accuracy and relevance.

In addition to improving data quality, our framework incorporates a frequency-based multi-label classification strategy to address keyword imbalance, where rare but clinically significant terms may be overlooked. By clustering keywords based on their frequency, our model reduces prediction errors while ensuring accurate identification of critical information. These keywords are then utilized by a text-to text large language model (TT-LLM), pretrained on medical datasets, to generate human-like radiology reports. Unlike traditional text decoders that rely on black-box features in computer vision, our text-to text LLM transforms interpretable keywords into coherent, natural language reports that maintain clinical relevance and readability. Our framework not only improves the quality and transparency of radiology report generation but also offers practical benefits for clinical workflows. Radiologists can easily review and modify the keyword list, providing a straightforward quality control mechanism and enabling personalized enhancements to report content. Furthermore, the generated reports are both comprehensive and interpretable, empowering clinicians to make informed decisions with confidence.

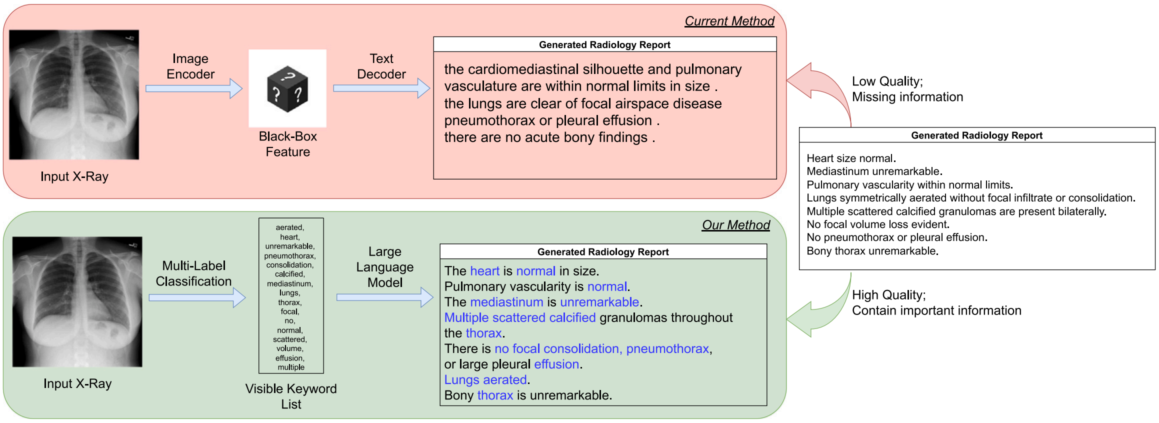
We evaluate our approach using two public datasets, IU-XRay and MIMIC-CXR, demonstrating its superior performance compared to state-of-the-art methods. The results show that our framework produces high-quality, human-like radiology reports that capture essential clinical information while maintaining transparency and adaptability. Through ablation studies, we validate the contributions of each component, including automatic keyword adaptation, frequency-based multi-label classification, and text-to text LLM-based report generation.

The main contributions are as follows:

- **Framework Innovation:** We introduce a deep learning-based framework that replaces black-box features in computer vision with a transparent keyword list, integrating automatic keyword adaptation, frequency-based multi-label classification, and text-to text LLM-based report generation.
- **Automatic Keyword Adaptation:** We propose a dynamic mechanism to generate and verify keywords using general language models and radiology-specific dictionaries, ensuring robust keyword extraction across diverse clinical settings.
- **Frequency-Based Multi-Label Classification:** We address keyword imbalance by clustering terms based on their frequency, reducing prediction errors and accurately identifying rare but important keywords.
- **Text-to text LLM-Based Report Generation:** We employ a pretrained text-to text LLM to transform verified keywords into coherent, human-like radiology reports, ensuring clarity, accuracy, and clinical utility.
- **Comprehensive Validation:** We validate our approach on public datasets, demonstrating superior performance compared to existing methods and highlighting the effectiveness of each framework component.

2. Related work

In this section, we first explore recent advancements in vision-language applications within medical contexts, establishing the foundation for our work and emphasizing the growing synergy between visual and textual data in healthcare. Additionally, we review current methodologies in radiology report generation for chest X-ray images, focusing on the mechanisms that inform and inspire our proposed framework.


Current Method
Input X-Ray
Image Encoder
Black-Box Feature
Text Decoder
Low Quality;
Missing information
Our Method
Input X-Ray
Multi-Label Classification
aerated, heart, unremarkable, pneumothorax, consolidation, calcified, mediastinum, lungs, thorax, focal, no, normal, scattered, volume, effusion, multiple
Visible Keyword List
Large Language Model
High Quality;
Contain important information

Generated Radiology Report

the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size .
the lungs are clear of focal airspace disease pneumothorax or pleural effusion .
there are no acute bony findings .

Generated Radiology Report

Heart size normal.
Mediastinum unremarkable.
Pulmonary vascularity within normal limits.
Lungs symmetrically aerated without focal infiltrate or consolidation.
Multiple scattered calcified granulomas are present bilaterally.
No focal volume loss evident.
No pneumothorax or pleural effusion.
Bony thorax unremarkable.

Generated Radiology Report

The heart is normal in size.
Pulmonary vascularity is normal.
The mediastinum is unremarkable.
Multiple scattered calcified granulomas throughout the thorax.
There is no focal consolidation, pneumothorax, or large pleural effusion.
Lungs aerated.
Bony thorax is unremarkable.

Fig. 1. Illustration of the motivation behind the proposed method. Conventional approaches employing an encoder–decoder structure rely on black-box features to generate radiology reports, often resulting in missed critical information. In contrast, our method leverages a visible keyword list, derived through multi-label classification, to replace unexplainable black-box features. This ensures that the generated radiology reports are not only comprehensive in covering essential information but also maintain high quality and enhanced explainability.

2.1. Vision-language applications in medical scene

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 contexts. With advances in deep learning and growing computational power, language models like BERT [6] and Transformer [7] have been introduced for natural language processing (NLP) tasks, revolutionizing automated language processing. 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, these models reduce workload and improve performance across diverse NLP tasks.

More recently, the application of transformer structures in computer vision has highlighted structural similarities in models used for both visual and language processing tasks. This synergy has driven the rise of vision-language integration as a promising research area, where text can aid in image understanding and images can enhance language processing by providing the high-level features for interpretation. Vision-language applications in natural images include image captioning, visual recognition, and text-to-image generation.

In medical imaging, vision-language applications hold even greater potential, as image interpretation directly impacts diagnostic efficiency. Integrating medical images with associated text is especially valuable, allowing for the automated generation of descriptive analysis and assisting in identifying key diagnostic features, which is critical in light of the current radiologist shortage.

Building on advancements in vision-language applications for natural images, similar approaches have been increasingly adopted in the medical imaging domain to tackle diverse tasks. These include image–text classification support [8–10], question-answering using images and associated questions [11,12], and medical object detection with text guidance [12,13]. In these tasks, images often serve as the primary data source, with text providing auxiliary contextual information. Paired image–text datasets are therefore essential for capturing the intricate relationships between visual and textual elements.

A significant advancement in vision-language modeling is the Contrastive Language-Image Pre-Training (CLIP) framework, proposed by OpenAI [14]. CLIP enables training models using large-scale image–text pairs, reducing the need to train models from scratch and providing a robust zero-shot predictor. Leveraging CLIP as a pre-trained model has led to numerous applications in the medical domain. These include fine-tuning pre-trained checkpoints for wide-range of tasks [9,15,16] Furthermore, CLIP has been employed as a zero-shot predictor for Classification [17–19], and medical visual question answering [20–22].

The CLIP framework introduces a paradigm shift in medical vision-language applications by enabling downstream tasks to benefit from pre-trained models through fine-tuning and optimization. Its adaptability and efficiency suggest a growing trend in leveraging CLIP-driven models in medical vision-language research.

Despite its potential, general vision-language models, including CLIP, face limitations when applied to new and dynamic medical contexts. Models trained on specific datasets may struggle to generalize across varying clinical environments, particularly when encountering unfamiliar diseases or diagnostic scenarios absent from training data. While CLIP partially mitigates this issue through its large-scale training, it still encounters challenges when predictions fall outside the predefined scope of its training datasets. For instance, tasks with undefined prediction ranges may lead to suboptimal performance, which is a common scenarios in radiology report generation which is not limited in the writing style. This underscores the need for further innovations to enhance the adaptability of vision-language models in the medical domain, particularly in scenarios with high variability and uncertainty.

To address this, we propose an automatic keyword adaptation mechanism that leverages sample reports to dynamically generate relevant keywords for new scenes, thereby relieve the problem of the application of vision-language models to unfamiliar disease contexts. This adaptive approach allows our model to overcome limitations in generalization, enhancing its robustness and effectiveness across a range of clinical applications.

2.2. Radiology report generation

Radiology reports describe and summarize the observed features in medical images, including X-rays, computed tomography scans, and magnetic resonance images [23]. They provide essential insights into a patient's condition and are vital 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 [24].

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. Monshi et al. [25] This integration of deep learning has attracted attention due to its potential to enhance both the efficiency and accuracy of radiology reporting.

In the early stages of radiology report generation research, foundational datasets and benchmarks such as IU X-ray [26] and MIMIC-CXR [27] were introduced, offering chest X-ray images paired with free-text radiology reports. The IU X-ray dataset consists of 7470 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 Jing et al. [28] introduced a deep learning framework for radiology report generation based on an encoder–decoder architecture. In this model, a Convolutional Neural Network (CNN) functions as the feature extractor, capturing detailed visual information from medical images, while a Long Short-Term Memory (LSTM) network serves as the text decoder, generating radiology reports that mimic human-written descriptions. This encoder–decoder framework 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 [29–31]. Attention mechanisms have also been introduced to better align the encoded features with the content of the reports [32,33]. For the development of decoder, more sophisticated designs [34,35] are being used to generate text that captures the detailed and specific language used in radiology reports. Additionally, the connection between the encoder and decoder has been strengthened by using Graph Convolution Networks (GCNs), which help the system better understand the relationships between the report content and the image features.

A pioneering effort in this area, RadGraph, focused on extracting clinical entities and their relationships from radiology reports to construct a knowledge graph [36]. Leveraging an information extraction schema, RadGraph achieved a micro F1 score of 0.82 for MIMIC-CXR and 0.73 for the CheXpert test sets in relation extraction. Building on RadGraph's foundation, recent studies have explored knowledge-based mechanisms and graph structures to better model relationships within radiology reports [37,38]. These approaches typically involve identifying key entities and relationships, then constructing a knowledge graph to capture the underlying structure and dependencies. This process mirrors human reasoning, offering an interpretable feature set and improving the quality of generated reports.

Despite these advancements, the traditional encoder–decoder pipeline has significant limitations. The quality of generated reports heavily depends on the black-box features extracted by the encoder, which are constrained by the training data. This dependency limits the model's adaptability to new clinical settings, reducing its utility across diverse environments. Additionally, the text decoder in the standard pipeline often starts training from scratch, relying 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 address these shortcomings, recent studies have incorporated large language models (LLMs) pretrained on extensive language datasets or commercial LLMs to enhance report quality. For instance, Wang et al. [39] employed Llama2, a frozen LLM, for radiology report generation. This approach uses tokenizers and visual mappers to translate visual features into language-understandable tokens, resulting in reports with high readability and a human-like style. Similarly, Soleimani et al. [40] investigated the feasibility of using ChatGPT for radiology report generation. Their method involved using the online LLMs Claude.ai to extract keywords from the input data, which were then fed into a predefined template via ChatGPT to produce the report. However, these approaches still rely on unexplainable features extracted by models or tools as input, making them susceptible to limitations in flexibility and robustness across varied clinical contexts.

In our proposed approach, we aim to replace the traditional encoder–decoder structure with a hybrid system combining multi-label classification and an LLM. Instead of generating complex, opaque features, the multi-label classification step identifies explicit, interpretable keywords describing the image. These keywords are then synthesized into a coherent, human-like radiology report by the LLM. This approach not only enhances interpretability but also enables more adaptable, high-quality text generation across diverse medical scenarios.

2.3. Keyword extraction in medical contexts

Keyword extraction is a fundamental task in natural language processing (NLP) that involves identifying and isolating significant terms to distill complex texts into their essential elements. This process plays a pivotal role in enabling language models to perform downstream tasks effectively. In the medical domain, keyword extraction assumes even greater importance, given the specialized terminology and intricate structures inherent in medical texts. These complexities pose unique challenges for traditional machine learning and NLP methods.

Over the years, a variety of rule-based and model-driven approaches have been developed to address keyword extraction in healthcare. For instance, [90] proposed a framework that utilizes a medical dictionary to extract and organize keywords by matching and retrieving terms from clinical reports. This method captures critical clinical information by leveraging domain-specific lexicons. Similarly, [91] evaluated keyword extraction techniques applied to pathology reports in electronic health records, introducing refined methodologies to enhance accuracy and relevance. These studies underscore the feasibility and utility of keyword extraction in medical NLP while highlighting the need for precise and domain-specific adaptations.

Radiology reports, however, present unique challenges for keyword extraction. Unlike pathology reports, they often lack well-established ground truth for keyword identification. Moreover, radiology reports exhibit significant variability in writing styles and terminologies across datasets, leading to domain shift—a phenomenon where models trained on one dataset struggle to generalize effectively to another. Addressing these issues necessitates innovative solutions that can adapt to diverse radiology datasets while maintaining high accuracy.

To overcome these challenges, our proposed framework introduces an automatic keyword adaptation mechanism. This approach combines a general-purpose language model with a specialized radiology dictionary, enabling the extraction of relevant keywords even in the absence of predefined ground truth annotations. By dynamically adapting to varying styles and terminologies, the framework effectively mitigates the impact of domain shifts between datasets. Specifically, it sets the keywords dynamically before applying multi-label classification, ensuring consistency and relevance across diverse radiology contexts.

This mechanism not only improves the accuracy of keyword extraction but also enhances the alignment between extracted features and the textual content of radiology reports. By strengthening the connection between the visual and textual modalities, our framework ensures the generation of high-quality, interpretable radiology reports. This adaptability and robustness make our framework a powerful solution for vision-language integration in medical imaging applications.

2.4. 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. In natural image processing, multi-label classification has been extensively studied, with several approaches

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.

Work	Year	Model (Encoder)	Model (Decoder)	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Jing et al. [28]	2017	CNN	LSTM	0.517	0.386	0.306	0.247	0.447	0.217	0.327
Xue et al. [41]	2018	CNN	LSTM	0.464	0.358	0.270	0.195	0.366	0.274	/
Harzig et al. [42]	2019	CNN(ResNet-152)	LSTM	0.373	0.246	0.175	0.126	0.315	0.163	0.359
Xie et al. [32]	2019	CNN	LSTM	0.443	0.337	0.236	0.181	0.347	/	0.374
Yuan et al. [43]	2019	CNN(ResNet-152)	LSTM	0.529	0.372	0.315	0.255	0.453	0.343	/
Li et al. [44]	2019	CNN	Transformer	0.482	0.325	0.226	0.162	0.339	/	0.280
Jing et al. [45]	2019	CNN	LSTM	0.464	0.301	0.210	0.154	0.362	/	0.275
Chen et al. [29]	2020	Transformer	Transformer	0.470	0.304	0.219	0.165	0.371	0.187	/
Zhang et al. [38]	2020	CNN (DenseNet-121)	LSTM	0.441	0.291	0.203	0.147	0.367	/	0.304
Wang et al. [46]	2021	CNN	LSTM	0.487	0.346	0.270	0.208	0.359	/	0.452
Alfarghaly et al. [47]	2021	CheXnet (DenseNet121-CNN)	GPT2	0.387	0.245	0.166	0.111	0.289	0.164	0.257
Liu et al. [33]	2021	CNN(ResNet-50)	LSTM	0.492	0.314	0.222	0.169	0.381	0.193	/
Liu et al. [48]	2021	Transformer	Transformer	0.483	0.315	0.224	0.168	0.376	0.190	0.351
Yang et al. [49]	2021	CNN	Transformer	0.496	0.327	0.238	0.178	0.381	/	0.382
Nooralahzadeh et al. [34]	2021	CNN	Transformer	0.486	0.317	0.232	0.173	0.390	0.192	/
Yang et al. [50]	2021	CNN	Transformer	0.497	0.319	0.230	0.174	0.399	/	0.407
Zhou et al. [51]	2021	CNN	Transformer	0.536	0.391	0.314	0.252	0.448	0.228	0.339
Li et al. [52]	2022	CNN+Transformer	Transformer	0.467	0.334	0.261	0.215	0.415	0.201	/
You et al. [31]	2022	Transformer	Transformer	0.484	0.313	0.225	0.173	0.379	0.204	/
Wang et al. [53]	2022	CNN	Transformer	0.505	0.340	0.247	0.188	0.382	0.208	/
Sirshar et al. [54]	2022	CNN	LSTM	0.580	0.342	0.263	0.155	/	/	/
Yan et al. [55]	2022	CNN	Transformer	/	/	0.256	/	0.341	/	0.380
Wang et al. [56]	2022	Transformer	Transformer	0.496	0.319	0.241	0.175	0.377	/	0.449
Yu and Zhang [57]	2022	CNN(ResNet-152)	Transformer	0.457	0.305	0.216	0.171	0.391	/	0.426
Chen et al. [58]	2022	CNN	Transformer	0.475	0.309	0.222	0.170	0.375	0.191	/
Wang et al. [59]	2022	CNN(ResNet-101)	Transformer	0.525	0.357	0.262	0.199	0.411	0.220	0.359
Nicolson et al. [60]	2022	Transformer	Transformer	0.473	0.303	0.224	0.175	0.375	0.199	0.693
Delbrouck et al. [61]	2022	CNN	BERT	/	/	/	0.121	0.306	/	/
You et al. [62]	2022	CNN (ResNet)	Transformer	0.479	0.319	0.222	0.174	0.377	0.193	/
Wu et al. [63]	2022	CNN	LSTM	0.458	0.324	0.238	0.180	0.369	0.206	0.287
Yan et al. [64]	2022	CNN	Transformer	0.482	0.313	0.232	0.181	0.381	0.203	0.735
Wang et al. [65]	2022	Graph Convolution Network, CNN	Transformer	0.450	0.301	0.213	0.158	0.384	/	0.340
Qin and Song [66]	2022	CNN	Transformer	0.494	0.321	0.235	0.181	0.384	0.201	/
Tanwani et al. [67]	2022	CNN (ResNeXt-101)	BERT, Transformer	0.580	0.440	0.320	0.270	/	/	/
Wang et al. [68]	2022	CNN (ResNet-101)	Transformer	0.505	0.345	0.243	0.176	0.396	0.205	/
Kong et al. [69]	2022	Transformer	Transformer	0.484	0.333	0.238	0.175	0.415	0.207	/
Li et al. [70]	2023	Transformer	Transformer	/	/	/	0.163	0.383	0.193	0.586
Yang et al. [71]	2023	CNN	LSTM	0.478	0.344	0.248	0.180	0.398	/	0.439
Kale et al. [35]	2023	CNN(ResNet-152)	BART	0.423	0.256	0.194	0.165	0.444	0.150	/
Huang et al. [72]	2023	CNN(ResNet-101)	Transformer	0.525	0.360	0.251	0.185	0.409	0.242	/
Wang et al. [73]	2023	Transformer	Transformer	0.483	0.322	0.228	0.172	0.380	0.192	0.435
Hou et al. [74]	2023	CNN	Transformer	0.510	0.346	0.255	0.195	0.399	0.200	/
Wang et al. [39]	2023	Swin-Transformer	LLAMA2	0.488	0.316	0.228	0.173	0.377	0.211	0.438
Kale et al. [75]	2023	CNN	Transformer	0.402	0.322	0.285	0.170	0.567	0.455	0.473
Li et al. [76]	2023	VAE	Transformer	0.530	0.365	0.263	0.200	0.405	0.218	0.501
Mohsan et al. [77]	2023	Transformer	Transformer	0.532	0.344	0.233	0.158	0.387	0.218	0.500
Chen et al. [78]	2023	Transformer	Transformer	0.505	0.334	0.245	0.190	0.394	0.210	0.592
Zhang et al. [79]	2024	CNN	Transformer	0.482	0.310	0.221	0.165	0.377	0.195	/
Liu et al. [80]	2024	MiniGPT-4	MiniGPT-4	0.499	0.323	0.238	0.184	0.390	0.208	/
Zhou et al. [81]	2024	GPT4, CLIP	BLIP-2	/	/	/	0.208	0.387	0.216	/
Yi et al. [82]	2024	CNN (ResNet 101)	Transformer	0.500	0.349	0.256	0.194	0.402	0.218	/
Parres et al. [83]	2024	Swin Transformer	BERT	/	/	/	0.149	0.341	/	/
Yi et al. [84]	2024	CNN (ResNet-101)	Transformer	0.539	0.380	0.278	0.210	0.416	0.223	/

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 [92].

With the advent of deep learning, particularly Convolutional Neural Networks (CNNs), multi-label classification saw significant advancements, as deep learning models offered robust feature extraction capabilities. Numerous strategies have since been developed to optimize multi-label classification, including advanced network structures [93,94,94,95], attention mechanisms [96–98], custom loss functions [99,100], and specialized training methods for deep learning models [101]. These innovations have enhanced model performance on complex, imbalanced datasets, a frequent challenge in multi-label classification.

However, while natural image datasets such as MS-COCO, PASCAL VOC, and ImageNet provide large volumes of labeled data for multi-label classification, medical imaging datasets are often smaller and feature highly imbalanced label distributions. This imbalance can lead models to prioritize more frequent labels—often representing normal findings—at the expense of decreasing in detecting less common, clinically important abnormalities. To address this, various methods have been proposed to improve sensitivity for infrequent labels. Techniques like label co-occurrence [102], attention mechanisms [103], and custom loss functions [104] have been adapted from natural image classification to support multi-label tasks in medical imaging.

While these approaches have shown high performance in multi-label classification for medical images, directly applying multi-label classification as a keyword prediction tool in radiology images presents additional challenges. Unlike natural image classification, the keyword requirements in radiology are dynamic, varying across different medical contexts. Traditional multi-label methods assume a stable set of labels, but in radiology, label sets may shift based on the specific clinical environment or available data.

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.

Work	Year	Model (Encoder)	Model (Decoder)	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Chen et al. [29]	2020	Transformer	Transformer	0.353	0.218	0.145	0.103	0.277	0.142	/
Liu et al. [33]	2021	CNN(ResNet-50)	LSTM	0.350	0.219	0.152	0.109	0.283	0.151	/
Liu et al. [48]	2021	Transformer	Transformer	0.360	0.224	0.149	0.106	0.284	0.149	0.237
Yang et al. [49]	2021	CNN	Transformer	0.363	0.228	0.156	0.115	0.284	/	0.203
Nooralahzadeh et al. [34]	2021	CNN	Transformer	0.378	0.232	0.154	0.107	0.272	0.145	/
Yang et al. [50]	2021	CNN	Transformer	0.386	0.237	0.157	0.111	0.274	/	0.111
Hou et al. [85]	2021	CNN	Transformer	0.232	/	/	/	0.240	0.101	0.493
Zhou et al. [51]	2021	CNN	Transformer	0.372	0.241	0.168	0.123	0.335	0.190	1.121
Yan et al. [86]	2021	Transformer	BERT	0.373	/	/	0.107	0.274	0.144	/
Wang et al. [30]	2022	Transformer	Transformer	0.413	0.266	0.186	0.136	0.298	0.170	0.429
You et al. [31]	2022	Transformer	Transformer	0.378	0.235	0.156	0.112	0.283	0.158	/
Wang et al. [53]	2022	CNN	Transformer	0.395	0.253	0.170	0.121	0.284	0.147	/
Yan et al. [55]	2022	CNN	Transformer	/	/	0.145	/	0.225	/	0.160
Wang et al. [56]	2022	Transformer	Transformer	0.351	0.223	0.157	0.118	0.287	/	0.281
Yu and Zhang [57]	2022	CNN(ResNet-152)	Transformer	0.347	0.235	0.149	0.106	0.280	/	0.552
Chen et al. [58]	2022	CNN	Transformer	0.353	0.218	0.148	0.106	0.278	0.142	
Wang et al. [59]	2022	CNN(ResNet-101)	Transformer	0.344	0.215	0.146	0.105	0.279	0.138	/
Nishino et al. [87]	2022	CNN	LSTM	/	/	/	0.168	0.122	/	/
Nicolson et al. [60]	2022	Transformer	Transformer	0.392	0.247	0.171	0.126	0.286	0.154	0.389
Delbrouck et al. [61]	2022	CNN	BERT	/	/	/	0.116	0.259	/	/
Wu et al. [63]	2022	CNN	LSTM	0.34	0.212	0.145	0.103	0.270	0.139	0.109
Serra et al. [37]	2022	CNN(ResNet-101)	Transformer	0.363	0.245	0.178	0.136	0.313	0.161	/
Yan et al. [64]	2022	CNN	Transformer	0.356	0.222	0.151	0.111	0.280	0.140	0.154
Qin and Song [66]	2022	CNN	Transformer	0.381	0.232	0.155	0.109	0.287	0.151	/
Wang et al. [68]	2022	CNN (ResNet-101)	Transformer	0.363	0.235	0.164	0.118	0.301	0.136	/
Kong et al. [69]	2022	Transformer	Transformer	0.423	0.261	0.171	0.116	0.286	0.168	/
Li et al. [70]	2023	Transformer	Transformer	/	/	/	0.109	0.284	0.150	0.281
Tanida et al. [88]	2023	CNN	Transformer	0.373	0.249	0.175	0.126	0.264	0.168	0.495
Yang et al. [71]	2023	CNN	LSTM	0.362	0.251	0.188	0.143	0.326	/	0.273
Huang et al. [72]	2023	CNN(ResNet-101)	Transformer	0.393	0.243	0.159	0.113	0.285	0.160	/
Wang et al. [73]	2023	Transformer	Transformer	0.386	0.250	0.169	0.124	0.291	0.152	0.362
Hou et al. [74]	2023	CNN	Transformer	0.407	0.256	0.172	0.123	0.293	0.162	/
Wang et al. [39]	2023	Swin-Transformer	LLAMA2	0.411	0.267	0.186	0.134	0.297	0.160	0.269
Kale et al. [75]	2023	CNN	Transformer	0.253	0.188	0.169	0.163	0.348	0.268	0.331
Li et al. [76]	2023	VAE	Transformer	0.363	0.229	0.158	0.107	0.289	0.157	0.246
Chen et al. [78]	2023	Transformer	Transformer	0.400	0.245	0.165	0.119	0.28	0.150	0.190
Zhang et al. [79]	2024	CNN	Transformer	0.362	0.229	0.157	0.113	0.284	0.153	/
Liu et al. [80]	2024	MiniGPT-4	MiniGPT-4	0.402	0.262	0.180	0.128	0.291	0.175	/
Zhou et al. [81]	2024	GPT4, CLIP	BLIP-2	/	/	/	0.122	0.296	0.165	/
Yi et al. [82]	2024	CNN (ResNet 101)	Transformer	0.398	0.248	0.169	0.121	0.281	0.149	/
Parres et al. [83]	2024	Swin Transformer	BERT	/	/	/	0.116	0.265	/	/
Zhang et al. [89]	2024	Transformer	Transformer	0.391	0.258	0.182	0.129	0.282	0.175	0.526
Yi et al. [84]	2024	CNN (ResNet-101)	Transformer	0.400	0.253	0.171	0.120	0.296	0.154	/

To address this limitation, we propose an automatic keyword adaptation mechanism that dynamically adjusts the keyword set for multi-label classification based on the availability of ground-truth keywords and radiology reports. This adaptive approach enables multi-label classification to be robust across varied medical contexts, enhancing its applicability and relevance in radiology. Furthermore, to solve the common problem of label imbalance in keywords, we propose to utilize the frequency-based multi-label classification to cover both the common keywords and rare but clinical keywords for generating high-quality radiology report.

3. Methodology

In this section, we present our proposed method for radiology report generation, which integrates Automatic Keyword Adaptation, Frequency-Based Multi-Label Classification, and TT-LLM for radiology report generation based on keywords. Section 3.1 provides an overview of our approach, while Section 3.2 details the process of automatic keyword adaptation, which extracts and adapts keywords from radiology reports to set up frequency-based multi-label classification. In Section 3.3, we introduce the frequency-based multi-label classification, which utilizes frequency clusters to efficiently classify keywords and generate the corresponding keyword list. Section 3.4 describes the application of TT-LLM in generating radiology reports based on the keyword list obtained from frequency-based multi-label classification. Finally, Section 3.5 outlines the loss functions employed in the frequency-based multi-label classification and the fine-tuning of the TT-LLM.

3.1. Overview

Given a set of radiology X-ray images, the objective is to generate a detailed sequence 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 TT-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.

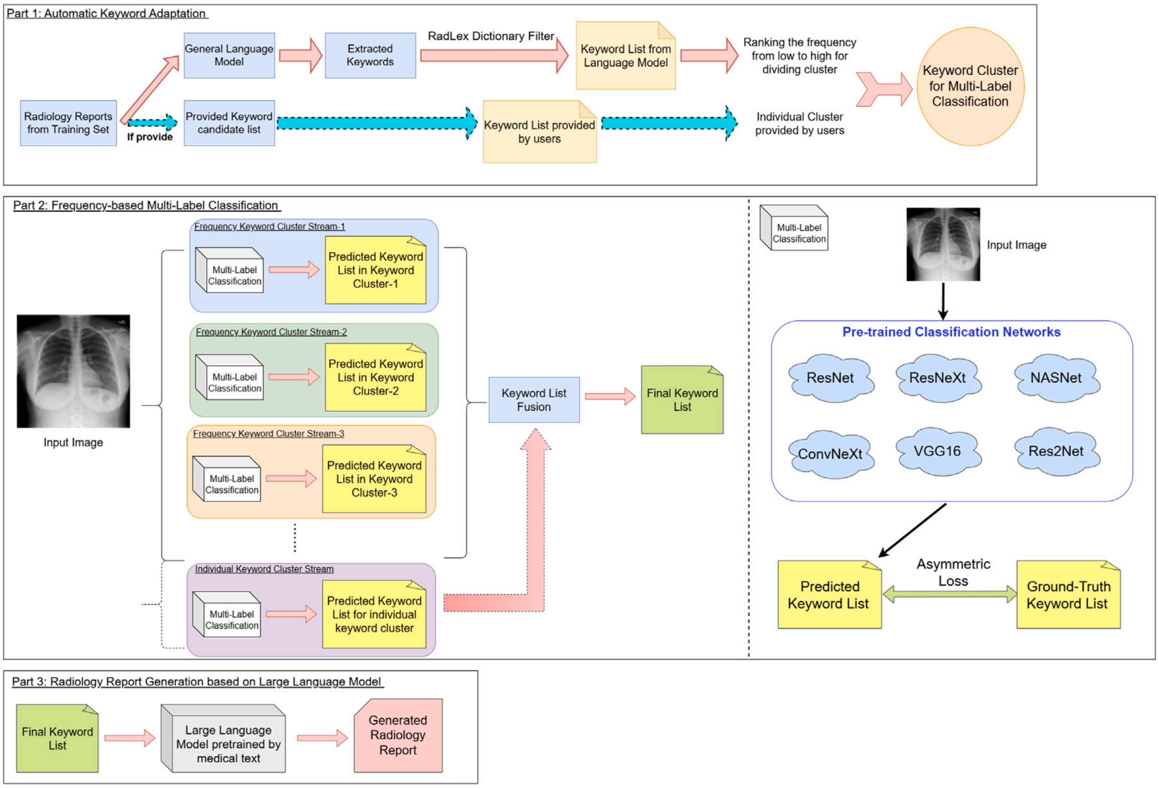

Part 1: Automatic Keyword Adaptation
Radiology Reports from Training Set
General Language Model
Extracted Keywords
RadLex Dictionary Filter
Keyword List from Language Model
Ranking the frequency from low to high for dividing cluster
Keyword Cluster for Multi-Label Classification
If provide
Provided Keyword candidate list
Keyword List provided by users
Individual Cluster provided by users
Part 2: Frequency-based Multi-Label Classification
Input Image
Frequency Keyword Cluster Stream-1
Multi-Label Classification
Predicted Keyword List in Keyword Cluster-1
Frequency Keyword Cluster Stream-2
Predicted Keyword List in Keyword Cluster-2
Frequency Keyword Cluster Stream-3
Predicted Keyword List in Keyword Cluster-3
Individual Keyword Cluster Stream
Predicted Keyword List for individual keyword cluster
Keyword List Fusion
Final Keyword List
Multi-Label Classification
Input Image
Pre-trained Classification Networks
ResNet
ResNeXt
NASNet
ConvNeXt
VGG16
Res2Net
Predicted Keyword List
Asymmetric Loss
Ground-Truth Keyword List
Part 3: Radiology Report Generation based on Large Language Model
Final Keyword List
Large Language Model pretrained by medical text
Generated Radiology Report

Fig. 2. 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.

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 TT-LLM 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 Fig. 2, 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.

3.2. Automatic keyword adaptation

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 [105,106] 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 in our pipeline [107]. 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

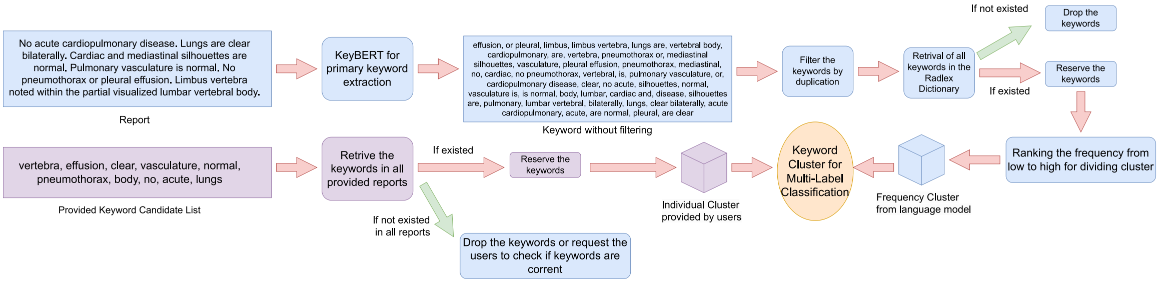

No acute cardiopulmonary disease. Lungs are clear bilaterally. Cardiac and mediastinal silhouettes are normal. Pulmonary vasculature is normal. No pneumothorax or pleural effusion. Limbus vertebra noted within the partial visualized lumbar vertebral body.
Report
KeyBERT for primary keyword extraction
effusion, or pleural, limbus, limbus vertebra, lungs are, vertebral body, cardiopulmonary, are, vertebra, pneumothorax or, mediastinal silhouettes, vasculature, pleural effusion, pneumothorax, mediastinal, no, cardiac, no pneumothorax, vertebral, is, pulmonary vasculature, or, cardiopulmonary disease, clear, no acute, silhouettes, normal, vasculature is, is normal, body, lumbar, cardiac and, disease, silhouettes are, pulmonary, lumbar vertebral, bilaterally, lungs, clear bilaterally, acute cardiopulmonary, acute, are normal, pleural, are clear
Keyword without filtering
Filter the keywords by duplication
Retrival of all keywords in the Radlex Dictionary
If not existed
Drop the keywords
If existed
Reserve the keywords
Ranking the frequency from low to high for dividing cluster
Frequency Cluster from language model
Keyword Cluster for Multi-Label Classification
vertebra, effusion, clear, vasculature, normal, pneumothorax, body, no, acute, lungs
Provided Keyword Candidate List
Retrive the keywords in all provided reports
If existed
Reserve the keywords
Individual Cluster provided by users
If not existed in all reports
Drop the keywords or request the users to check if keywords are corrent

Fig. 3. Block diagram of the proposed automatic keyword adaptation process. Beginning with the provided radiology reports, candidate keywords are extracted using the KeyBERT tool [105]. Subsequently, the extracted keywords are filtered using the RadLex Dictionary [107], 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.

dictionary RadLex, 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 Fig. 3. 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.

3.3. *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.

3.3.1. *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. 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.

3.3.2. *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. 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. 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. The visualization of the Keyword List Fusion process is shown in Fig. 4. 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. Specifically, while a base threshold of 0.5 is used as a general classification criterion, we further refine this threshold adaptively to better capture infrequent but important keywords. This is achieved by scaling the base threshold by the ratio of each keyword's frequency to the total keyword frequency. As a result, lower-frequency keywords are given proportionally more opportunity to be selected, ensuring that semantically important yet rare terms are not overlooked during the fusion process.

3.4. *Radiology report generation based on keywords with text-to-text large language model (TT-LLM)*

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 [108], utilizing the fine-tuning methodology demonstrated by Clinical-T5 [109], 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

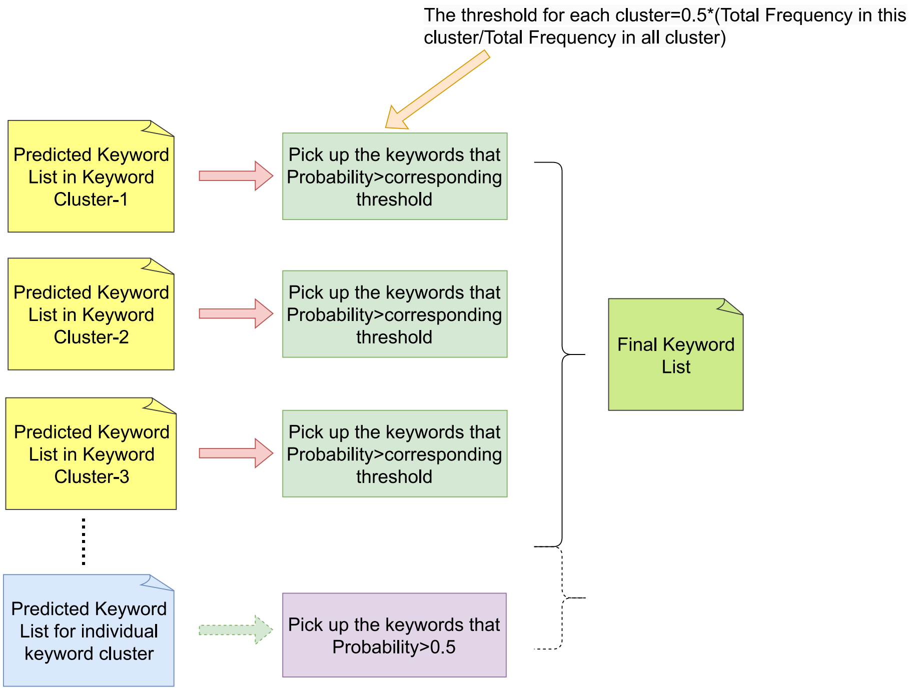

The threshold for each cluster=0.5*(Total Frequency in this cluster/Total Frequency in all cluster)
Predicted Keyword List in Keyword Cluster-1
Pick up the keywords that Probability>corresponding threshold
Predicted Keyword List in Keyword Cluster-2
Pick up the keywords that Probability>corresponding threshold
Predicted Keyword List in Keyword Cluster-3
Pick up the keywords that Probability>corresponding threshold
Final Keyword List
Predicted Keyword List for individual keyword cluster
Pick up the keywords that Probability>0.5

Fig. 4. 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.

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. 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.

3.5. Loss function

Our method comprises three main components: automatic keyword adaptation, 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 adaptation 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 [36], 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:

$$ASL(L_+) = (1-p)^{\gamma+} log(p) \tag{1}$$

$$ASL(L_-) = (p_m)^{\gamma-} log(1-p_m) \tag{2}$$

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 Pm is defined as:

$$P_m = max(p-m, 0) \tag{3}$$

Here, the probability margin m_0 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, while P(X) defined as prediction and G(X) defined as ground-truth:

$$CE(L) = -\sum_{x} G(x) log P(x) \tag{4}$$

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. Experiment

In this section, we first describe the two public datasets and the applied radiology dictionary RadLex in our experiments, and also the metrics and experimental settings in detail. Then, we present the keyword distribution analysis for two datasets generated by our automatic keyword adaptation. After that, we present both quantitative analysis and qualitative analysis of the proposed framework. Finally, to validate the effectiveness of each part in the proposed framework, we present the ablation study based on these two datasets.

4.1. Datasets and dictionary details

We conduct the experiments on two public datasets, i.e., IU X-ray [26] and MIMIC-CXR [110]. Moreover, we also introduce the radiology dictionary RadLex [107] used for Automatic Keyword Adaptation. We also show the basic dataset and dictionary information in Table 3.

IU X-ray. The IU X-ray dataset is a widely-used benchmark and radiology report dataset proposed by the Indiana University. It contains 7566 chest X-ray images associated with 3852 radiology reports in original version. We firstly follow the official split from [29] into training set, validation set and testing set, and then we filter the images that do not contains the radiology report. To simply the process of training, we assume that each image in the cases shared the same radiology report. Finally, we collect and construct 6659 pairs in training set, 295 pairs in validation set and 590 pairs in testing set.

MIMIC-CXR. The MIMIC-CXR dataset for radiology reports is currently the largest publicly available dataset of chest radiographs with free-text radiology reports. The latest version of the datasets contain 377,110 images corresponding to 227,835 radiographic studies. We utilize the version from July 23, 2024 and follow the official split provided by the PhysioNet and also filter the images without associated radiology reports. Finally, we collect and construct 270,790 pairs in training set, 2130 pairs in validation set and 3858 pairs in testing set.

RadLex. RadLex, developed by the Radiological Society of North America (RSNA), is a comprehensive ontology of radiology terms designed for use in radiology reporting and related research. For our experiments, we utilize RadLex version 4.2, which contains a vast collection of 46,838 terms. While this extensive vocabulary is valuable for covering diverse radiological concepts, its size presents challenges for direct application in radiology report generation. To address this, we leverage an automatic keyword adaptation mechanism to dynamically match extracted keywords with relevant entries in the RadLex dictionary. This approach effectively reduces the size of the keyword set, making it more manageable and tailored to the specific requirements of our report generation tasks.

4.2. Implementation details

The proposed framework employs a unified subnetwork architecture for frequency-based multi-label classification, with ConvNeXt [111] selected as the backbone network due to its advanced feature extraction capabilities. The process begins 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 remain unseen during training, the keyword frequency analysis is limited to the training sets, as illustrated in Table 3. To manage the distribution of keywords, a logarithmic split strategy (log x) is applied based on the maximum frequency observed in each dataset. For IU X-ray, frequency clusters are divided into three ranges: [10,100], [100,1000] and [1000,10,000]. Meanwhile, for MIMIC-CXR, five clusters are defined as [10,100], [100,1000],[1000,10,000],[10,000,100,000], and [100,000+]. This approach ensures a balanced representation of keywords across frequency ranges, thereby improving classification performance.

For radiology report generation, the framework employs a modified version of the Text-to-Text Transformer (T5) model [108], following the fine-tuning methodology described in [109]. Pre-trained checkpoints are sourced from HuggingFace's repository [112] to minimize initialization and training costs. The model is fine-tuned using the extracted keywords and corresponding radiology reports, which enables it to generate domain-specific, contextually accurate reports. This process adapts the model's language generation capabilities to align with the stylistic and clinical requirements of radiology documentation.

The training process is optimized for both tasks. For multi-label classification, the network is trained for 120 epochs per frequency cluster with a batch size of 4. The Adam optimizer is used with an initial learning rate of 0.0001. For the fine-tuning of the language model, the training is performed over 10 epochs with a batch size of 2, using the Adam optimizer and an initial learning rate of 0.00005. All experiments are 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.

4.3. Evaluation metrics and details

The primary objective of this study is radiology report generation, and accordingly, the evaluation metrics focus on assessing the quality of the generated reports. For the two datasets utilized in this work, IU X-ray and MIMIC-CXR, we benchmark the performance of our proposed framework against a wide range of state-of-the-art (SOTA) radiology report generation models, as listed in Tables 1 and 2 in the "Related Work" section. To ensure a fair and comprehensive comparison, we re-implement the models proposed by [29,60] using their open-source code, aligning with our dataset splits, and include their results in the evaluation. For other SOTA models where re-implementation is not feasible, the reported performance values are taken directly from their original publications. It is worth noting that if a model does not provide results for one of the two datasets, we only report the metrics presented in its original paper and avoid relying on results from secondary sources, even if such comparisons are available.

The evaluation metrics used for radiology report generation align with those commonly adopted in SOTA works. These include Bilingual Evaluation Understudy (BLEU) scores (BLEU-1, BLEU-2, BLEU-3, BLEU-4) [113], Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [114], Consensus-based Image Description Evaluation (CIDEr) [115], and Metric for Evaluation of Translation with Explicit ORdering (METEOR) [116]. In addition to language-based metrics, we incorporate Clinical Evaluation (CE) metrics to assess the clinical accuracy and usefulness of the generated reports—an increasingly important standard in recent works. Specifically, we adopt the CheXpert clinical evaluation metrics [117], which include F1 Score, Precision, and Recall calculated over a set of 14 clinical labels. Furthermore, we report the RadGraph evaluation scores [36], including both the Entity F1 Score and the Relation F1 Score, which evaluate how well the generated report captures the structured clinical entities and their relationships.

While these metrics collectively provide a robust assessment of the quality of generated reports, it should be noted that not all papers report all metrics. Consequently, some cells in the evaluation tables may remain blank, indicating the absence of corresponding metric results in the original publications.

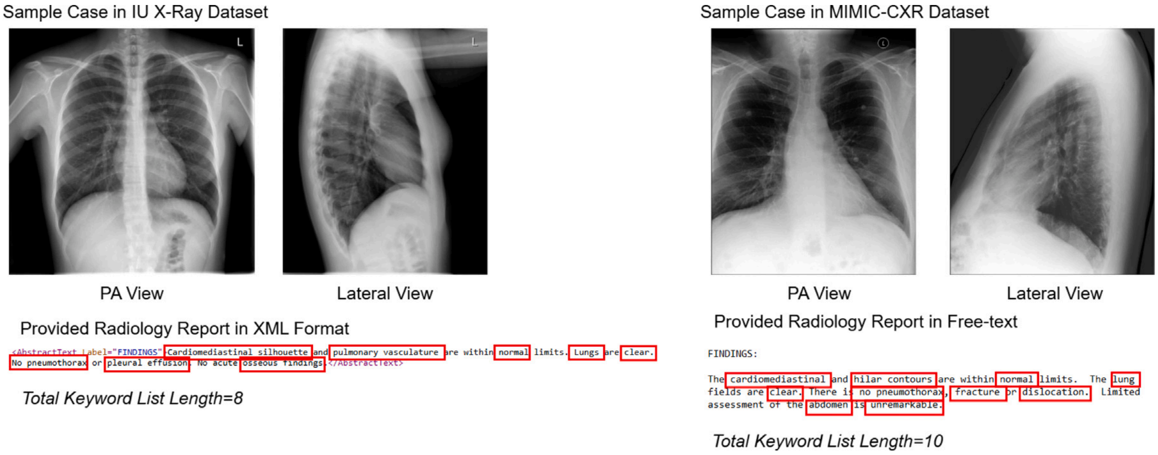
Table 3
Dataset descriptions for the IU X-ray and MIMIC-CXR datasets, including details of keyword cluster information derived from the automatic keyword adaptation process. Additionally, the table provides information on the radiology dictionary used in the experiments and presents the proportion of keywords in the two experimental datasets relative to the full version of the dictionary.

Setting	Description
Applied Radiology Dictionary	
Dictionary Name	RadLex [107]
Dictionary Version	4.2
Dictionary Source	https://radlex.org/
Total Number of Keywords	46,838
Dataset 1: IU X-ray	
Basic Information	
Open Source	https://openi.nlm.nih.gov/faq
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	
Open Source	https://physionet.org/content/mimic-cxr/2.1.0/
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	196 051
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]	96 119
Cluster [100000+] 196 051	

Additionally, following established practices in radiology report generation research, all performance comparisons are conducted exclusively on the test sets of each dataset. This ensures consistency and comparability across different models and datasets, providing a reliable benchmark for evaluating the effectiveness of our framework.

4.4. Analysis of automatic keyword adaptation

It is important to note that ground-truth keywords for each radiology report were not provided by radiologists in either the IU X-ray or MIMIC-CXR datasets. Given the scale of these datasets—exceeding 200,000 cases—manual annotation by radiologists is not feasible due to resource constraints. Moreover, the number of keywords per report can vary significantly depending on the stylistic conventions of the radiologist. To better understand this, Fig. 5 presents examples of manually extracted keywords, verified using the RadLex radiology dictionary. Additionally, Fig. 6 highlights representative failure cases where predicted keywords are compared with manually identified ones. These examples reveal that clinically relevant terms, such as “cardiomediastinal”, are often overlooked by automatic extraction methods.

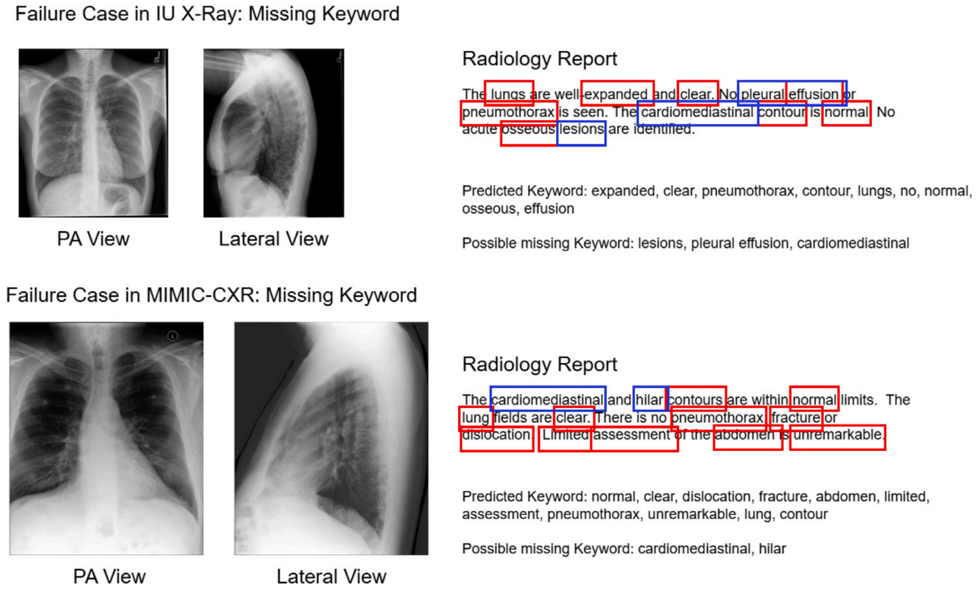

Sample Case in IU X-Ray Dataset
PA View
Lateral View
Provided Radiology Report in XML Format
Total Keyword List Length=8
Sample Case in MIMIC-CXR Dataset
PA View
Lateral View
Provided Radiology Report in Free-text
Total Keyword List Length=10

```
<AbstractText Label="FINDINGS">Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Lungs are clear.
No pneumothorax or pleural effusion. No acute osseous findings.</AbstractText>
```

```
FINDINGS:

The cardiomediastinal 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.
```

Fig. 5. 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. Keywords were manually annotated and verified in the RadLex radiology dictionary. 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
Failure Case in MIMIC-CXR: Missing Keyword
PA View
Lateral View
Radiology Report

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

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

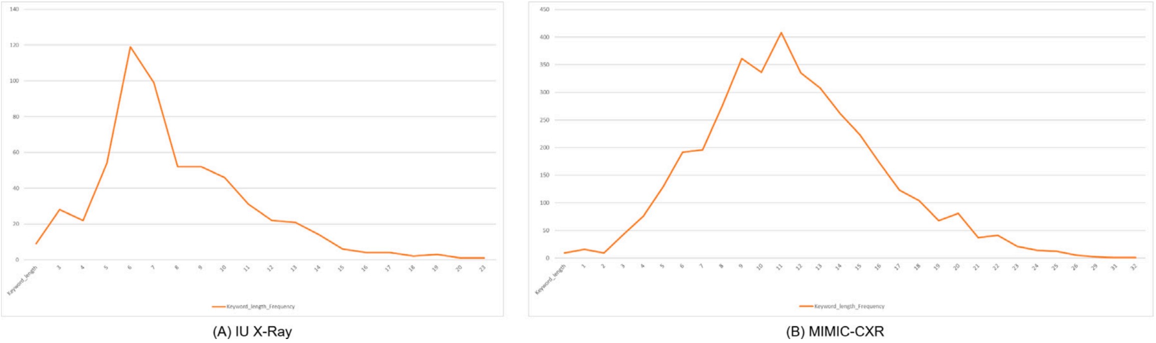
Possible missing Keyword: lesions, pleural effusion, cardiomediastinal

The cardiomediastinal 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: cardiomediastinal, hilar

Fig. 6. 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.



(A) IU X-Ray

(B) MIMIC-CXR

Fig. 7. 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.

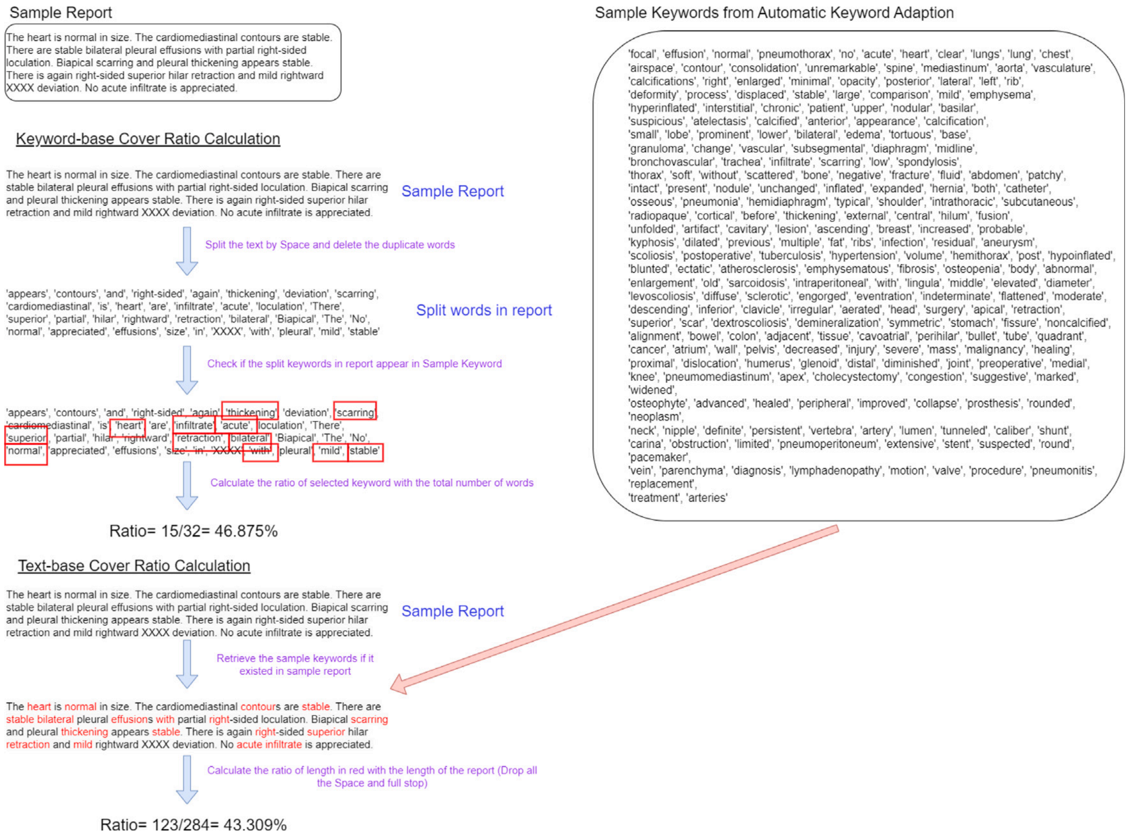

Sample Report
The heart is normal in size. The cardiomediastinal contours are stable. There are stable bilateral pleural effusions with partial right-sided loculation. Biapical scarring and pleural thickening appears stable. There is again right-sided superior hilar retraction and mild rightward XXXX deviation. No acute infiltrate is appreciated.
Keyword-base Cover Ratio Calculation
The heart is normal in size. The cardiomediastinal contours are stable. There are stable bilateral pleural effusions with partial right-sided loculation. Biapical scarring and pleural thickening appears stable. There is again right-sided superior hilar retraction and mild rightward XXXX deviation. No acute infiltrate is appreciated.
Sample Report
Split the text by Space and delete the duplicate words
'appears', 'contours', 'and', 'right-sided', 'again', 'thickening', 'deviation', 'scarring', 'cardiomediastinal', 'is', 'heart', 'are', 'infiltrate', 'acute', 'loculation', 'There', 'superior', 'partial', 'hilar', 'rightward', 'retraction', 'bilateral', 'Biapical', 'The', 'No', 'normal', 'appreciated', 'effusions', 'size', 'in', 'XXXX', 'with', 'pleural', 'mild', 'stable'
Split words in report
Check if the split keywords in report appear in Sample Keyword
'appears', 'contours', 'and', 'right-sided', 'again', 'thickening', 'deviation', 'scarring', 'cardiomediastinal', 'is', 'heart', 'are', 'infiltrate', 'acute', 'loculation', 'There', 'superior', 'partial', 'hilar', 'rightward', 'retraction', 'bilateral', 'Biapical', 'The', 'No', 'normal', 'appreciated', 'effusions', 'size', 'in', 'XXXX', 'with', 'pleural', 'mild', 'stable'
Calculate the ratio of selected keyword with the total number of words
Ratio= 15/32= 46.875%
Text-base Cover Ratio Calculation
The heart is normal in size. The cardiomediastinal contours are stable. There are stable bilateral pleural effusions with partial right-sided loculation. Biapical scarring and pleural thickening appears stable. There is again right-sided superior hilar retraction and mild rightward XXXX deviation. No acute infiltrate is appreciated.
Sample Report
Retrieve the sample keywords if it existed in sample report
The heart is normal in size. The cardiomediastinal contours are stable. There are stable bilateral pleural effusions with partial right-sided loculation. Biapical scarring and pleural thickening appears stable. There is again right-sided superior hilar retraction and mild rightward XXXX deviation. No acute infiltrate is appreciated.
Calculate the ratio of length in red with the length of the report (Drop all the Space and full stop)
Ratio= 123/284= 43.309%
Sample Keywords from Automatic Keyword Adaption
'focal', 'effusion', 'normal', 'pneumothorax', 'no', 'acute', 'heart', 'clear', 'lungs', 'lung', 'chest',
'airspace', 'contour', 'consolidation', 'unremarkable', 'spine', 'mediastinum', 'aorta', 'vasculature',
'calcifications', 'right', 'enlarged', 'minimal', 'opacity', 'posterior', 'lateral', 'left', 'rib',
'deformity', 'process', 'displaced', 'stable', 'large', 'comparison', 'mild', 'emphysema',
'hyperinflated', 'interstitial', 'chronic', 'patient', 'upper', 'nodular', 'basilar',
'suspicious', 'atelectasis', 'calcified', 'anterior', 'appearance', 'calcification',
'small', 'lobe', 'prominent', 'lower', 'bilateral', 'edema', 'tortuous', 'base',
'granuloma', 'change', 'vascular', 'subsegmental', 'diaphragm', 'midline',
'bronchovascular', 'trachea', 'infiltrate', 'scarring', 'low', 'spondylosis',
'thorax', 'soft', 'without', 'scattered', 'bone', 'negative', 'fracture', 'fluid', 'abdomen', 'patchy',
'intact', 'present', 'nodule', 'unchanged', 'inflated', 'expanded', 'hernia', 'both', 'catheter',
'osseous', 'pneumonia', 'hemidiaphragm', 'typical', 'shoulder', 'intrathoracic', 'subcutaneous',
'radiopaque', 'cortical', 'before', 'thickening', 'external', 'central', 'hilum', 'fusion',
'unfolded', 'artifact', 'cavitary', 'lesion', 'ascending', 'breast', 'increased', 'probable',
'kyphosis', 'dilated', 'previous', 'multiple', 'fat', 'ribs', 'infection', 'residual', 'aneurysm',
'scoliosis', 'postoperative', 'tuberculosis', 'hypertension', 'volume', 'hemithorax', 'post', 'hypoinflated',
'blunted', 'ectatic', 'atherosclerosis', 'emphysematous', 'fibrosis', 'osteopenia', 'body', 'abnormal',
'enlargement', 'old', 'sarcoidosis', 'intraperitoneal', 'with', 'lingula', 'middle', 'elevated', 'diameter',
'levoscoliosis', 'diffuse', 'sclerotic', 'engorged', 'eventration', 'indeterminate', 'flattened', 'moderate',
'descending', 'inferior', 'clavicle', 'irregular', 'aerated', 'head', 'surgery', 'apical', 'retraction',
'superior', 'scar', 'dextroscoliosis', 'demineralization', 'symmetric', 'stomach', 'fissure', 'noncalcified',
'alignment', 'bowel', 'colon', 'adjacent', 'tissue', 'cavoatrial', 'perihilar', 'bullet', 'tube', 'quadrant',
'cancer', 'atrium', 'wall', 'pelvis', 'decreased', 'injury', 'severe', 'mass', 'malignancy', 'healing',
'proximal', 'dislocation', 'humerus', 'glenoid', 'distal', 'diminished', 'joint', 'preoperative', 'medial',
'knee', 'pneumomediastinum', 'apex', 'cholecystectomy', 'congestion', 'suggestive', 'marked',
'widened',
'osteophyte', 'advanced', 'healed', 'peripheral', 'improved', 'collapse', 'prosthesis', 'rounded',
'neoplasm',
'neck', 'nipple', 'definite', 'persistent', 'vertebra', 'artery', 'lumen', 'tunneled', 'caliber', 'shunt',
'carina', 'obstruction', 'limited', 'pneumoperitoneum', 'extensive', 'stent', 'suspected', 'round',
'pacemaker',
'vein', 'parenchyma', 'diagnosis', 'lymphadenopathy', 'motion', 'valve', 'procedure', 'pneumonitis',
'replacement',
'treatment', 'arteries'

Fig. 8. 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.

This underscores the potential value of allowing radiologists to manually supplement the predicted keyword list before frequency-based multi-label classification, thereby improving report accuracy.

Prior to conducting experiments on the IU X-ray and MIMIC-CXR datasets, we analyzed the keyword statistics across training, validation, and test sets. The results, presented in Table 4, show that the MIMIC-CXR dataset, due to its larger sample size, has a higher average number of keywords per case compared to the IU X-ray dataset. Fig. 7 further illustrates that medium-length keyword lists occur most frequently in both datasets, and their distributions generally approximate a normal distribution. Furthermore, we observed that applying filters to remove extremely low-frequency keywords and those not listed in the RadLex dictionary significantly reduces the total number of keywords. These filtering steps demonstrate the effectiveness of our automatic keyword adaptation mechanism in minimizing classification workload while improving the accuracy of subsequent keyword prediction.

To better support downstream generation, we also examined the frequency distribution of extracted keywords in the training sets of the IU X-ray and MIMIC-CXR datasets. The keyword frequency in each dataset was examined, and the distributions were visualized in Figs. 9 and 10, with additional statistical details, such as cluster splits and highest frequencies, summarized in Table 3. To manage the distribution of keywords, we employed a logarithmic split (logx) strategy based on the maximum observed frequency in each dataset. While the logarithmic method was used in this study to reduce the number of clusters, it is not a fixed requirement; the frequency-based clusters can be flexibly adjusted based on specific needs. To simplify the network and mitigate extreme class imbalance, keywords with a frequency of less than 10 were excluded.

The frequency distributions revealed a general trend: as the dataset scale increases, the imbalance in keyword frequency becomes more pronounced. For instance, in the smaller IU X-ray dataset, the most frequent keyword, “no”, appears over 5000 times, leading to three clusters spanning frequencies from 10 to 5000. In contrast, the larger MIMIC-CXR dataset exhibits a more significant imbalance, with the keyword “pneumothorax” appearing over 190,000 times. To address this, we divided the MIMIC-CXR dataset into five clusters, covering frequencies 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 dictionary. Table 3 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 Fig. 8.

The Keyword-Based Strategy involves extracting unique keywords from sample radiology reports by splitting text into individual words and removing duplicates. These unique keywords are then matched against those generated by the adaptation mechanism, with the coverage ratio calculated as the percentage of matching keywords relative to the total unique keywords in the report. In the Text-Based Strategy, the generated keywords are directly searched within the radiology reports, and matching words or phrases are highlighted. The coverage ratio is computed as the proportion of the total length of matched words or phrases to the total text length of the report.

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.

Table 4
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	/

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%

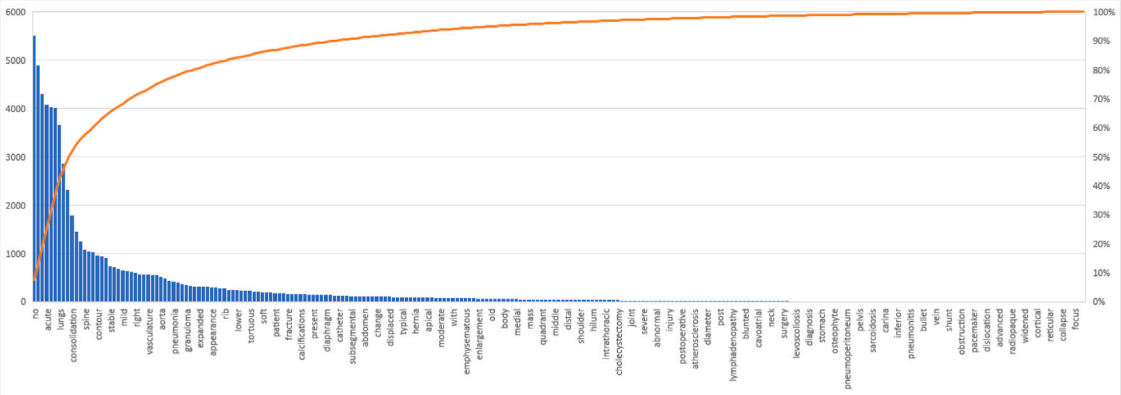


Fig. 9. Pareto chart illustrating the keyword distribution in the IU X-ray dataset. Blue bars represent the frequency of each keyword in the training set, while the orange line indicates the cumulative frequency ratio from the most frequent keyword to the current keyword compared with the total frequency of keyword.

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 Tables 7–9 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: an 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), a 78.0% increase in CIDEr (1.996 vs. 1.121), 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

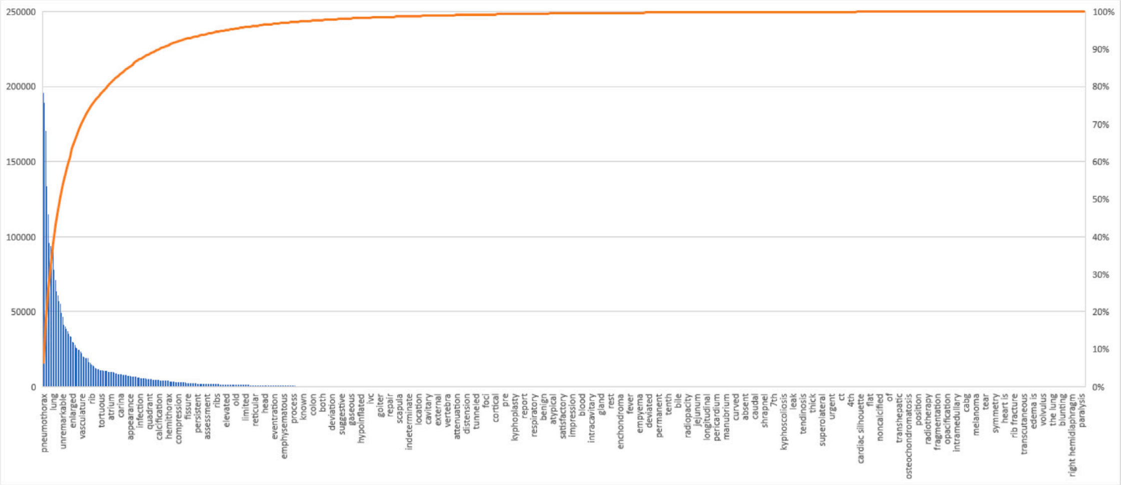


Fig. 10. Pareto chart illustrating the keyword distribution in the MIMIC-CXR dataset. Blue bars represent the frequency of each keyword in the training set, while the orange line indicates the cumulative frequency ratio from the most frequent keyword to the current keyword compared with the total frequency of keyword.

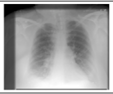
X-Ray Image	Reference Report	Predicted Keywords in report	Our proposed method	State-of-the-art Method1: Cvt2Distgen (Official Checkpoint)	State-of-the-art Method2: R2Gen (Official Checkpoint)
	Frontal and lateral views of the chest show an unchanged cardiomediastinal silhouette. Right-sided aortic XXXX. Normal mediastinal contour, pulmonary XXXX and vasculature, central airways and aeration of the lungs. There is right basal XXXX patchy opacity and bibasal 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 cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. the lungs are clear of focal airspace disease pneumothorax or pleural effusion. there are no acute bony findings.	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .
	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 cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. the lungs are clear of focal airspace disease pneumothorax or pleural effusion. there are no acute bony findings.	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .
	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 cardiomediastinal silhouette is normal in size and contour . no focal consolidation pneumothorax or large pleural effusion . normal xxxx .

Fig. 11. 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 TT-LLM.

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.

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 select high-performing SOTA methods and generate their radiology reports using their publicly available source code. The generated reports are 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 Fig. 11 for the IU X-ray dataset and Fig. 12 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 identifies and incorporates keywords such as “pneumothorax”, “vasculature”, “effusion”, and “granuloma”, while in Sample 2 of the MIMIC-CXR dataset, it highlights terms like “consolidation”, “atelectasis”, “engorged”, “enlargement”, and “aorta”. These keywords are 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 TT-LLM. 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 TT-LLM offers a robust and interpretable solution

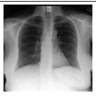
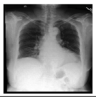
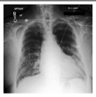
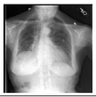
X-Ray Image	Reference Report	Predicted Keywords in report	Our proposed method	State-of-the-art Method1: Cvt2Distgen (Official Checkpoint)	State-of-the-art Method2: R2Gen (Official Checkpoint)
	The cardiac, mediastinal and hilar contours appear stable. The heart is normal in size. There is no pleural 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 cardiomediastinal 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 cardiomediastinal 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 cardiomediastinal 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, stable. 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 cardiomediastinal 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 lobar 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 lobar consolidation is seen . there is no pleural effusion or pneumothorax . the cardiac and mediastinal silhouettes are stable .

Fig. 12. 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.

for radiology report generation, producing reports that are both structurally coherent and clinically relevant.

4.7. Ablation study

The integration of frequency-based multi-label classification and automatic keyword adaptation enables our framework to achieve state-of-the-art performance in radiology report generation. To better understand the contributions of each component, we conduct an ablation study to evaluate the roles of automatic keyword adaptation, frequency-based multi-label classification, and radiology report generation within the framework.

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 plays a critical role in the overall performance of radiology report generation. However, evaluating multi-label classification performance is 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 use the keywords extracted by the automatic keyword adaptation mechanism as pseudo ground truth. This allows us to monitor classification performance and compare the impact of different network architectures.

In addition to the ConvNeXt backbone used in our experiments, we evaluate the performance of several alternative network architectures, including ResNeXt [118], ResNet [119], VGG16 [120], EfficientNet [121], NASNet [122], and Res2Net [123]. These networks are tested in the multi-label classification stage and subsequently in radiology report generation, using the same TT-LLM to ensure consistency. As there is no directly comparable work on keyword extraction and prediction from radiology reports, we focus 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 indicates 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 emerges 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.

4.7.2. Influence of keyword numbers and the combination of high- and low-frequency keywords on radiology report generation performance

Before evaluating the effectiveness of frequency-based keyword clustering, we first investigate how the number of keywords per case influences radiology report generation performance within our framework. To this end, we link several commonly used language evaluation metrics—BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr—with the length of the keyword list in each case. The corresponding trends are visualized in Fig. 13 for the IU X-ray dataset and Fig. 14 for the MIMIC-CXR dataset. These plots reveal a consistent trend across both datasets: as the number of keywords increases, the performance of the report generation model tends to decrease. This observation suggests that generating accurate and coherent reports becomes more challenging as the keyword list grows longer—likely due to the increased semantic complexity and the greater demand for contextual alignment among keywords.

Building upon this insight, we proceed to evaluate the efficiency of our frequency-based clustering strategy, particularly the integration of high- and low-frequency keywords within the radiology report generation pipeline. 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 examine 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 validate the generated reports using

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 the Natural Language Generation (NLG) Metrics. 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.

Work	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Paper Report Performance							
Jing et al. [28]	0.517	0.386	0.306	0.247	0.447	0.217	0.327
Xue et al. [41]	0.464	0.358	0.27	0.195	0.366	0.274	/
Harzig et al. [42]	0.373	0.246	0.175	0.126	0.315	0.163	0.359
Xie et al. [32]	0.443	0.337	0.236	0.181	0.347	/	0.374
Yuan et al. [43]	0.529	0.372	0.315	0.255	0.453	0.343	/
Li et al. [44]	0.482	0.325	0.226	0.162	0.339	/	0.28
Jing et al. [45]	0.464	0.301	0.21	0.154	0.362	/	0.275
Chen et al. [29]	0.47	0.304	0.219	0.165	0.371	0.187	/
Zhang et al. [38]	0.441	0.291	0.203	0.147	0.367	/	0.304
Wang et al. [46]	0.487	0.346	0.27	0.208	0.359	/	0.452
Alfarghaly et al. [47]	0.387	0.245	0.166	0.111	0.289	0.164	0.257
Liu et al. [33]	0.492	0.314	0.222	0.169	0.381	0.193	/
Liu et al. [48]	0.483	0.315	0.224	0.168	0.376	0.19	0.351
Yang et al. [49]	0.496	0.327	0.238	0.178	0.381	/	0.382
Nooralahzadeh et al. [34]	0.486	0.317	0.232	0.173	0.39	0.192	/
Yang et al. [50]	0.497	0.319	0.23	0.174	0.399	/	0.407
Zhou et al. [51]	0.536	0.391	0.314	0.252	0.448	0.228	0.339
Li et al. [52]	0.467	0.334	0.261	0.215	0.415	0.201	/
You et al. [31]	0.484	0.313	0.225	0.173	0.379	0.204	/
Wang et al. [53]	0.505	0.34	0.247	0.188	0.382	0.208	/
Sirshar et al. [54]	0.58	0.342	0.263	0.155	/	/	/
Yan et al. [55]	/	/	0.256	/	0.341	/	0.38
Wang et al. [56]	0.496	0.319	0.241	0.175	0.377	/	0.449
Yu and Zhang [57]	0.457	0.305	0.216	0.171	0.391	/	0.426
Chen et al. [58]	0.475	0.309	0.222	0.17	0.375	0.191	/
Wang et al. [59]	0.525	0.357	0.262	0.199	0.411	0.22	0.359
Nicolson et al. [60]	0.4732	0.3039	0.2242	0.1754	0.3758	0.1997	0.6935
Delbrouck et al. [61]	/	/	/	0.121	0.306	/	/
You et al. [62]	0.479	0.319	0.222	0.174	0.377	0.193	/
Wu et al. [63]	0.458	0.324	0.238	0.18	0.369	0.206	0.287
Yan et al. [64]	0.482	0.313	0.232	0.181	0.381	0.203	0.735
Wang et al. [65]	0.45	0.301	0.213	0.158	0.384	/	0.34
Qin and Song [66]	0.494	0.321	0.235	0.181	0.384	0.201	/
Tanwani et al. [67]	0.58	0.44	0.32	0.27	/	/	/
Wang et al. [68]	0.505	0.345	0.243	0.176	0.396	0.205	/
Kong et al. [69]	0.484	0.333	0.238	0.175	0.415	0.207	/
Li et al. [70]	/	/	/	0.163	0.383	0.193	0.586
Yang et al. [71]	0.478	0.344	0.248	0.18	0.398	/	0.439
Kale et al. [35]	0.423	0.256	0.194	0.165	0.444	0.15	/
Huang et al. [72]	0.525	0.36	0.251	0.185	0.409	0.242	/
Wang et al. [73]	0.483	0.322	0.228	0.172	0.38	0.192	0.435
Hou et al. [74]	0.51	0.346	0.255	0.195	0.399	0.20	/
Wang et al. [39]	0.488	0.316	0.228	0.173	0.377	0.211	0.438
Kale et al. [75]	0.402	0.322	0.285	0.17	0.567	0.455	0.473
Li et al. [76]	0.53	0.365	0.263	0.2	0.405	0.218	0.501
Mohsan et al. [77]	0.532	0.344	0.233	0.158	0.387	0.218	0.5
Chen et al. [78]	0.505	0.334	0.245	0.19	0.394	0.21	0.592
Zhang et al. [79]	0.482	0.31	0.221	0.165	0.377	0.195	/
Liu et al. [80]	0.499	0.323	0.238	0.184	0.39	0.208	/
Zhou et al. [81]	/	/	/	0.208	0.387	0.216	/
Yi et al. [82]	0.5	0.349	0.256	0.194	0.402	0.218	/
Parres et al. [83]	/	/	/	0.149	0.341	/	/
Yi et al. [84]	0.539	0.380	0.278	0.210	0.416	0.223	/
Re-Train Performance							
R2Gen ([29])	0.4514	0.2988	0.2163	0.1631	0.3377	0.201	0.5988
Cvt2Distgen2 ([60])	0.4182	0.2758	0.2037	0.1594	0.3315	0.1923	0.6784
Our Performance	0.7190 ± 0.2101	0.6250 ± 0.2713	0.5645 ± 0.3057	0.5215 ± 0.3376	0.6392 ± 0.2554	0.3861 ± 0.2218	3.2749 ± 1.0106

single clusters. Then, we extend 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 ensures 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.

Table 7

Performance comparison between the proposed method and other state-of-the-art radiology report generation methods on the MIMIC-CXR dataset for the Natural Language Generation (NLG) Metrics. 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.

Work	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
Paper Report Performance							
Chen et al. [29]	0.353	0.218	0.145	0.103	0.277	0.142	/
Liu et al. [33]	0.35	0.219	0.152	0.109	0.283	0.151	/
Liu et al. [48]	0.36	0.224	0.149	0.106	0.284	0.149	0.237
Yang et al. [49]	0.363	0.228	0.156	0.115	0.284	/	0.203
Nooralahzadeh et al. [34]	0.378	0.232	0.154	0.107	0.272	0.145	/
Yang et al. [50]	0.386	0.237	0.157	0.111	0.274	/	0.111
Hou et al. [85]	0.232	/	/	/	0.24	0.101	0.493
Zhou et al. [51]	0.372	0.241	0.168	0.123	0.335	0.19	1.121
Yan et al. [86]	0.373	/	/	0.107	0.274	0.144	/
Wang et al. [30]	0.413	0.266	0.186	0.136	0.298	0.17	0.429
You et al. [31]	0.378	0.235	0.156	0.112	0.283	0.158	/
Wang et al. [53]	0.395	0.253	0.17	0.121	0.284	0.147	/
Yan et al. [55]	/	/	0.145	/	0.225	/	0.16
Wang et al. [56]	0.351	0.223	0.157	0.118	0.287	/	0.281
Yu and Zhang [57]	0.347	0.235	0.149	0.106	0.28	/	0.552
Chen et al. [58]	0.353	0.218	0.148	0.106	0.278	0.142	/
Wang et al. [59]	0.344	0.215	0.146	0.105	0.279	0.138	/
Nishino et al. [87]	/	/	/	0.168	0.122	/	/
Nicolson et al. [60]	0.3928	0.2478	0.1713	0.1267	0.2863	0.1545	0.3892
Delbrouck et al. [61]	/	/	/	0.116	0.259	/	/
Wu et al. [63]	0.34	0.212	0.145	0.103	0.27	0.139	0.109
Serra et al. [37]	0.363	0.245	0.178	0.136	0.313	0.161	/
Yan et al. [64]	0.356	0.222	0.151	0.111	0.28	0.14	0.154
Qin and Song [66]	0.381	0.232	0.155	0.109	0.287	0.151	/
Wang et al. [68]	0.363	0.235	0.164	0.118	0.301	0.136	/
Kong et al. [69]	0.423	0.261	0.171	0.116	0.286	0.168	/
Li et al. [70]	/	/	/	0.109	0.284	0.15	0.281
Tanida et al. [88]	0.373	0.249	0.175	0.126	0.264	0.168	0.495
Yang et al. [71]	0.362	0.251	0.188	0.143	0.326	/	0.273
Huang et al. [72]	0.393	0.243	0.159	0.113	0.285	0.16	/
Wang et al. [73]	0.386	0.25	0.169	0.124	0.291	0.152	0.362
Hou et al. [74]	0.407	0.256	0.172	0.123	0.293	0.162	/
Wang et al. [39]	0.411	0.267	0.186	0.134	0.297	0.16	0.269
Kale et al. [75]	0.253	0.188	0.169	0.163	0.348	0.268	0.331
Li et al. [76]	0.363	0.229	0.158	0.107	0.289	0.157	0.246
Chen et al. [78]	0.4	0.245	0.165	0.119	0.28	0.15	0.19
Zhang et al. [79]	0.362	0.229	0.157	0.113	0.284	0.153	/
Liu et al. [80]	0.402	0.262	0.18	0.128	0.291	0.175	/
Zhou et al. [81]	/	/	/	0.122	0.296	0.165	/
Yi et al. [82]	0.398	0.248	0.169	0.121	0.281	0.149	/
Parres et al. [83]	/	/	/	0.116	0.265	/	/
Zhang et al. [89]	0.391	0.258	0.182	0.129	0.282	0.175	0.526
Yi et al. [84]	0.400	0.253	0.171	0.120	0.296	0.154	/
Re-Train Performance							
R2Gen ([29])	0.3058	0.1834	0.1221	0.0868	0.2386	0.1299	0.1466
Cvt2Distgen2 ([60])	0.2952	0.1839	0.1263	0.0927	0.2473	0.1308	0.1814
Our Performance	0.5599 ± 0.1607	0.4379 ± 0.1736	0.3557 ± 0.1824	0.2953 ± 0.1958	0.4699 ± 0.1687	0.2842 ± 0.1018	1.9964 ± 1.4518

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 Fig. 15 (IU X-ray) and Fig. 16 (MIMIC-CXR).

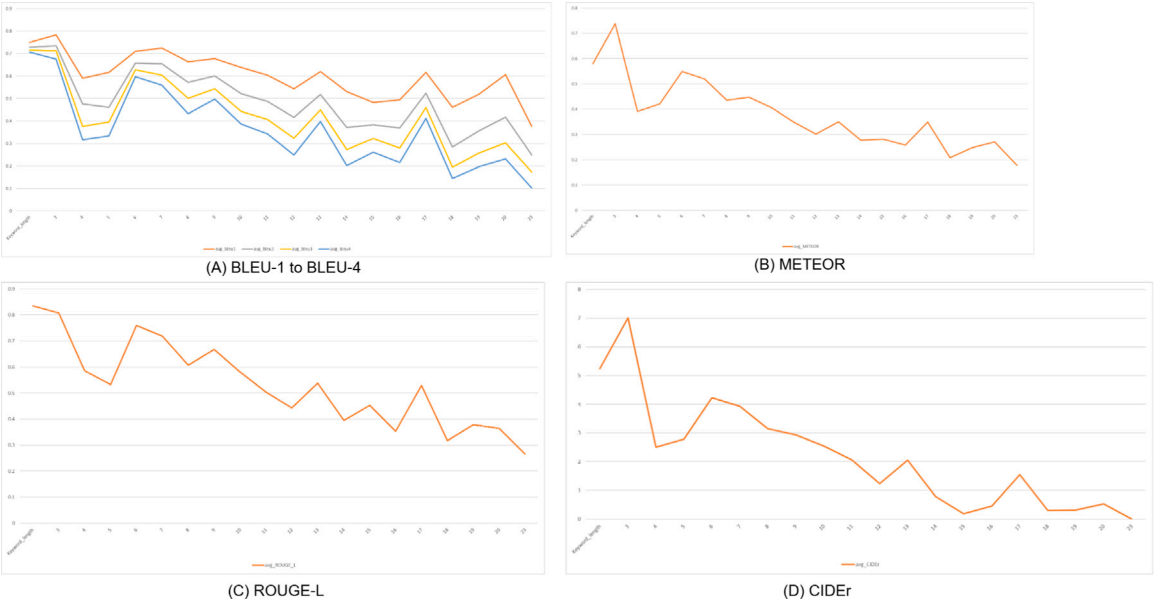
The results reveal distinct trends between the datasets. For the IU X-ray dataset with three clusters, the general observation is that high-frequency keywords yield better radiology report accuracy compared to low-frequency keywords, even when the latter includes more keywords. Specifically, although the clusters [10,100] and [100,1000] contain more keywords than [1000,10,000], the latter achieves higher performance, likely due to the higher quality of multi-label classification in the high-frequency cluster. Visualization of the generated reports confirms that high-frequency clusters include more relevant information, resulting in superior report quality. This trend extends to mixed clusters, where combining multiple clusters generally improves performance compared to individual clusters.

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

This analysis highlights a key optimization insight: while individual low-frequency clusters may perform poorly, combining them with high-frequency clusters can significantly enhance overall performance. This suggests 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 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 ± 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. [29]	0.333	0.273	0.276
Liu et al. [33]	0.352	0.298	0.303
Yang et al. [49]	0.458	0.348	0.371
Nooralahzadeh et al. [34]	0.240	0.428	0.308
Yang et al. [50]	0.420	0.339	0.352
Yu and Zhang [57]	0.447	0.593	0.503
Chen et al. [58]	0.334	0.275	0.278
Nicolson et al. [60]	0.367	0.418	0.391
Serra et al. [37]	0.428	0.459	0.443
Yan et al. [64]	0.353	0.310	0.297
Qin and Song [66]	0.342	0.294	0.292
Kong et al. [69]	0.482	0.563	0.519
Tanida et al. [88]	0.461	0.475	0.447
Huang et al. [72]	0.371	0.318	0.321
Wang et al. [73]	0.364	0.309	0.311
Hou et al. [74]	0.416	0.418	0.385
Wang et al. [39]	0.392	0.387	0.389
Chen et al. [78]	0.489	0.340	0.401
Zhang et al. [79]	0.38	0.342	0.335
Liu et al. [80]	0.465	0.482	0.473
Yi et al. [82]	0.319	0.509	0.393
Zhang et al. [89]	0.486	0.493	0.462
Yi et al. [84]	0.392	0.335	0.342
Re-Train Performance			
R2Gen ([29])	0.5047 ± 0.2413	0.3838 ± 0.2109	0.4361 ± 0.2094
Cvt2Distgen2([60])	0.4627 ± 0.2310	0.3423 ± 0.2566	0.3935 ± 0.2280
Our Performance	0.7448 ± 0.1215	0.6610 ± 0.1820	0.7004 ± 0.1513



(A) BLEU-1 to BLEU-4

(B) METEOR

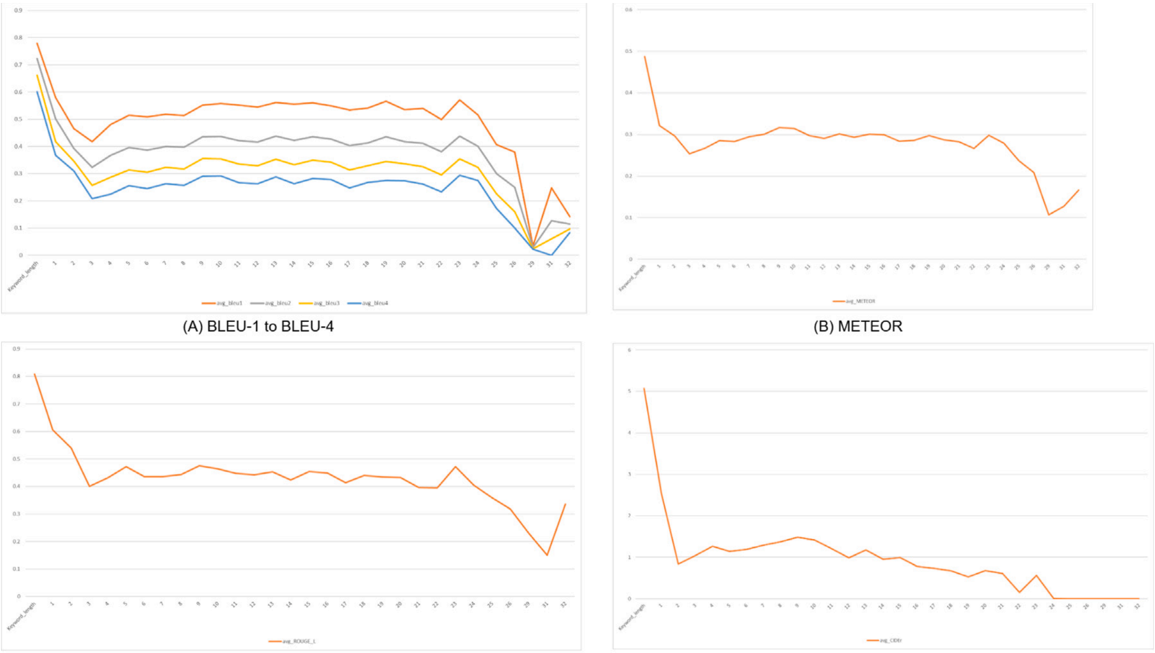
(C) ROUGE-L

(D) CIDEr

Fig. 13. 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.

4.7.3. Performance of radiology report generation with different 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.



(A) BLEU-1 to BLEU-4

(B) METEOR

(C) ROUGE-L

(D) CIDEr

Fig. 14. 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.

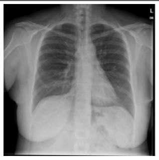
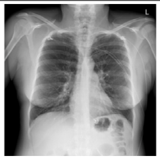
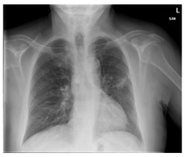
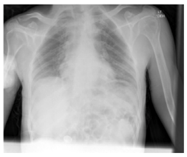
X-Ray Image				
Reference Report	The heart size is normal. The mediastinal contour is within normal limits. The lungs are free of any focal infiltrates. There are no nodules or masses. No visible pneumothorax. No visible pleural fluid. The XXXX are grossly normal. There is no visible free intraperitoneal air under the diaphragm. Surgical clips are seen the right upper quadrant.	The cardiomediastinal 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 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	*intraperitoneal, heart, contour, pneumothorax, fluid, quadrant, lungs, focal, no, upper, normal, diaphragm* 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.	*aerated, pneumothorax, consolidation, vasculature, lungs, focal, normal, effusion* The cardiomediastinal 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.	*heart, infiltrate, surface, effusion, chronic, calcifications, small, perihilar, calcified, stable, lobe, emphysema, pneumothorax, opacity, no, normal, left, extension, acute* 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.	*patient, heart, low, consolidation, enlarged, lungs, focal, no, lung, volume, effusion* 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]	*intraperitoneal,quadrant* Cardiomegaly with intraperitoneal effusions. Calcified right hilar lymph node and left midlung infiltrate/abdominal contours.	*aerated* Aerated and aerated. The cardiac silhouette is well-aerated.	*perihilar* The cardiac silhouette and cardiomediastinal silhouette are within normal limits in size. The perihilar contours are within normal limits in size.	*volume* Volume loss with no visible active cardiopulmonary disease. Volume loss with mild volume loss.
Keyword and generated report in Cluster [100,1000]	*upper,diaphragm,fluid* 1. Right upper diaphragm fluid in the right upper lobe. 2. No evidence of active disease. The cardiomediastinal silhouette and pulmonary vascularity are within normal limits. There is a right upper diaphragm fluid in the right upper lobe.	*vasculature* Vasculature within normal limits for size and configuration of the cardiac silhouette. Vasculature within normal limits for size and configuration of the cardiac silhouette.	*calcifications,opacity,left,stable,emphysema,chronic, calcified,small,lobe,infiltrate* 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 granulomas seen in the left lower lobe. There is stable emphysema. There are streaky calcifications of aortic XXXX.	*enlarged,patient,low* Cardiomegaly with enlarged cardiac silhouette. Low patient volume study.
Keyword and generated report in Cluster [1000,10000]	*focal,normal,pneumothorax,no,heart,lungs,contour* 1. No evidence of active disease. The cardiomediastinal silhouette is within normal limits for size and configuration. The lungs are free of focal infiltrates. No pleural effusions or pneumothorax.	*focal,effusion,normal,pneumothorax,lungs,consolidation* No evidence of active disease. The lungs are free of focal consolidation, pleural effusion, or pneumothorax. Cardiomediastinal silhouette is within normal limits. Pulmonary vascularity is within normal limits.	*effusion,normal,pneumothorax,no,acute,heart* No acute cardiopulmonary abnormality. Heart size within normal limits. No pneumothorax or pleural effusion.	*focal,effusion,no,heart,lungs,lung,consolidation* No evidence of active disease. The lungs and pleural spaces show no focal infiltrates. There is no pleural effusion or pneumothoraces. The cardiac silhouette and pulmonary vascularity appear within normal limits. The lungs are free of focal consolidation.
Keyword and generated report in Cluster [10,100]+[100,1000]	*upper,diaphragm,fluid,intraperitoneal,quadrant* 1. Right upper diaphragm XXXX with streaky infiltrates. The cardiomediastinal silhouette and pulmonary vascularity are within normal limits. There is a right upper diaphragm XXXX with streaky infiltrates and a XXXX pleural effusions. There is no pneumothoraces.	*vasculature,aerated* Cardiomegaly and pulmonary vasculature are well-aerated. The cardiomediastinal silhouette is well-aerated. Pulmonary vasculature is within normal limits in size.	*calcifications,opacity,left,stable,emphysema,chronic,calcified, small,lobe,infiltrate,perihilar* 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 vascularity, and mediastinal contours are within normal limits. There are streaky calcifications in the left lower lobe and left lower lobe.	*enlarged,patient,low,volume* Cardiomegaly with enlarged cardiomediastinal silhouette. Lung volumes are mildly enlarged. Patient is slightly enlarged.
Keyword and generated report in Cluster [10,100]+[1000,10000]	*focal,normal,pneumothorax,no,heart,lungs,contour, intraperitoneal,quadrant* 1. No evidence of active disease. The cardiomediastinal 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 sternotomy XXXX.	*focal,effusion,normal,pneumothorax,lungs,consolidation, aerated* 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.	*effusion,normal,pneumothorax,no,acute,heart,perihilar* No acute cardiopulmonary abnormality. The aortic XXXX and mediastinal contours are within normal limits. Pulmonary vascularity is within normal limits. There is no pleural effusion or pneumothorax.	*focal,effusion,no,heart,lungs,lung,consolidation,volume* 1. No evidence of active disease. The cardiomediastinal silhouette and pulmonary vascularity are within normal limits in size. The lungs are free of focal consolidation. No pleural effusions or pneumothoraces.
Keyword and generated report in Cluster [100,1000]+[1000,10000]	*focal,normal,pneumothorax,no,heart,lungs,contour,upper, diaphragm,fluid* 1. No evidence of active disease. The cardiac silhouette and pulmonary vascularity 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.	*focal,effusion,normal,pneumothorax,lungs,consolidation, vasculature* 1. No evidence of active disease. The cardiomediastinal silhouette and pulmonary vascularity are within normal limits. The lungs are free of focal consolidation, pleural effusion, or pneumothorax. There is no evidence of active effusion.	*effusion,normal,pneumothorax,no,acute,heart,calcifications, opacity,left,stable,emphysema,chronic,calcified,small,lobe, infiltrate* 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.	*focal,effusion,no,heart,lungs,lung,consolidation,enlarged, patient,low* 1. Cardiomegaly and enlarged cardiomediastinal silhouette. Low lung volumes without focal consolidation, pleural effusion, or pneumothoraces. The lungs are hyperexpanded with flattening of the right costophrenic XXXX blunting/effusion. There is no focal consolidation.

Fig. 15. Visualization of ablation study results validating the frequency-based multi-label classification on the IU X-ray dataset. Blue text highlights the predicted keywords within their respective clusters.

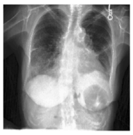
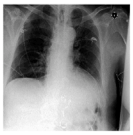
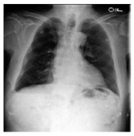
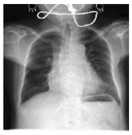
X-Ray Image				
Reference Report	Diffuse interstitial opacities, predominantly in the right lung base and probably very mild in the left lung base are present. When compared to the prior chest CT from ___, these interstitial opacities appear new, and given the clinical history, likely represent lung infection (atypical viral) 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 postradiation. Ill-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.	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.	The lungs are well expanded and clear. The cardiomediastinal and hilar contours are unremarkable. There is moderate aortic tortuosity, unchanged. A small right-sided pleural effusion is unchanged. There is no pneumothorax. Sternotomy wires are intact. Multiple fractures in early stages of healing are noted in the right rib cage.	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 parenchymal abnormalities are seen, and no pneumothorax is identified in the apical area. Again observed is a fractured second rib in the left apical area, apparently the result of previous sternotomy and intrathoracic 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 absorbed completely.
All predicted Keywords and generated report	*mild, diffuse, chest, interstitial, fibrosis, opacity, atelectasis, border, effusion, lower, minimal, unchanged, carcinomatosis, bilateral, lung, heart* 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 cardiomediastinal silhouette is unchanged. Mild obscuration of the left heart border is likely due to a combination of pleural effusion and atelectasis.	*atelectasis, left, normal, effusion, preoperative, minimal, small, lung, pneumonia* As compared to the previous radiograph, there is unchanged evidence of a right pleural effusion 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.	*rib, healing, no, moderate, effusion, intact, pneumothorax, right, multiple, lungs, small, expanded, clear, unchanged, unremarkable* The lungs are well expanded and clear. There is a small right pleural effusion. Moderate cardiomegaly is unchanged. Cardiomediastinal and hilar contours are unremarkable. Multiple healing right rib fractures are noted. Sternotomy wires are intact. No pneumothorax.	*chest, ascending, valve, contour, vasculature, metallic, atrium, rib, comparison, moderate, edema, effusion, fractured, enlarged, pneumothorax, enlargement, intrathoracic, patient, apical* 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.
Keyword and generated report in Cluster [10,100]	*carcinomatosis* The cardiac, mediastinal and hilar contours are unchanged since the prior examination. There is no evidence of pleural effusion or pneumothorax	*pneumo* 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.	*pneumo* 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.	*thoracic,pneumo* 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 pleurodesis and possible small effusions.
Keyword and generated report in Cluster [100,1000]	*carcinoma* In comparison with the study of ___, there is no change in the appearance of the known right hilar and mediastinal carcinoma carcinoma	*pre,all,preoperative* preoperative changes. Preoperative size of the preoperative preoperative is all preoperative.	*healing,all* All healing healing is appreciated.	*metallic,ascending,fractured,intrathoraci,all* All metallic metallic denfracfracfractures are fractured. Ascending metallic ascending aorta is unchanged. There is no evidence of intrathoracic or mediastinal mass.
Keyword and generated report in Cluster [1000,10000]	*minimal,diffuse,border,fusion,fibrosis* 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.	*minimal,fusion* Minimal cervical spinal fusion hardware.	*multiple,thorax,fusion* Multiple thorax is present, with multiple surgical clips in the left hemi thorax. Cervical spinal fusion hardware is partially visualized.	*apical,enlargement,valve,thorax,fusion* There is marked enlargement of the cardiac silhouette with valve replacement and median sternotomy wires. The mediastinal and hilar contours are stable, including mild apical pulmonary edema and right apical pleural thickening. Post-surgical 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.
Keyword and generated report in Cluster [10000,100000]	*chest,lung,heart,atelectasis,lateral,unchanged,opacity, mild,bilateral,low,interstitial,lower,change* Compared to chest radiograph from earlier the same day. Lung volumes remain low. Heart size is unchanged. Mild bilateral interstitial lung lung interstitial prominence is unchanged. Unchanged heart size. Mild retrocardiac opacity is unchanged, likely atelectasis. Bilateral pleural effusions are unchanged. Mild lower lung atelectasis. Overall, no significant change.	*lung,atelectasis,left,pneumonia,small* As compared to the previous radiograph, the lung volumes remain low. Small left and small right pleural effusions are present. Bibasilar areas of atelectasis are present in the retrocardiac lung areas. Streaky atelectasis is present at both lung bases, but no evidence of pneumonia or aspiration.	*clear,lung,unremarkable,unchanged,right,intact,small,rib, moderate,expanded,change* Unchanged from prior. The right lung is unremarkable. Unchanged small right rib fractures. Sternal intact. Moderate cardiomegaly unchanged. No change in the left lung, unremarkable mediastinal and hilar contours.	*chest,edema,enlarged,vasculature,patient,rib,moderate, large,contour,atrium,comparison,fracture* Comparison is made for the patient's chest comparison. Cardiac contour is moderate to severely enlarged. The cardiac contour is moderately enlarged. There is also a left-sided pacer 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 high positioned fracture of the eighth rib which is fracture fragment in the chest cavity.
Keyword and generated report in Cluster [100000+]	*effusion,no* As compared to the previous radiograph, there is no evidence of newly appeared parenchymal effusion. No other parenchymal abnormalities. No larger pleural effusions.	*effusion,no,normal* The cardiac, mediastinal and hilar contours are normal. No effusion. No sinister bony effusion.	*pneumothorax,effusion,no,lungs* The lungs are clear, the cardiomediastinal silhouette and hila are unremarkable. There is no pleural effusion or pneumothorax.	*pneumothorax,effusion* There is no focal consolidation, pleural effusion, pneumothorax, or pulmonary edema.
Keyword and generated report in Cluster ([10,100]+[100,1000]+[1000,10000])	*minimal,diffuse,border,fusion,fibrosis,carcinoma, carcinomatosis* Cardiac fusion, mediastinal and hilar contours are unchanged. The right border of the cardiac border is minimally more diffuse than on the previous study. This may reflect lymphanginomatosis 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 radioma fusion). Background fibrosis also seen in the right upper lobe compatible.	*minimal,fusion,pre,all,preoperative,pneumo* All preoperative fusion is minimal. All preoperative pneumo, right inf brachichial soft tissue thickening and cyst preoperative changes are unchanged. All mediastinal clips are consistent with preoperative thyroid thyroid surgery.	*multiple,thorax,fusion,healing,all,pneumo* Multiple healing thorax fractures are identified. There is all evidence of healing pneumo. Multiple surgical clips are seen in the right hemi thorax, and there is no evidence of mediastinal widening or radiopaquect within the left bronchial wall. Lumbar spinal fusion hardware is incompletely evaluated.	*apical,enlargement,valve,thorax,fusion,metallic,ascending, fractured,intrathoracic,all,thoracic,pneumo* There is stable enlargement of the ascending aortic valve, as well as an aortic valve replacement. Cervical fusion hardware is fractured and unchanged in position. All intrathoracic spinal fusion hardware is incompletely imaged. The thorax is grossly unremarkable except for a small metallic density projecting over the left mid hemi thorax, which may represent a surgical clip or pneumo-thoracic pleurodesis. Intrathoracic apical hyperdensity.
Keyword and generated report in Cluster ([10,100]+[100,1000]+[1000,10000])+[10000,100000]	*minimal,diffuse,border,fusion,fibrosis,carcinoma, carcinomatosis,chest,lung,heart,atelectasis,lateral, unchanged,opacity,mild,bilateral,low,interstitial,lower,change* Frontal and lateral views of the chest. Mild cardiomegaly is unchanged. Lung volumes are low. The mediast, bilateral fusion, breast carcinomatosis, lymphoma, lung carcinomatosis, and fibrosis remain consistent with patient's known lung carcinomatosct. Right lower lung fusion demonstrates diffuse interstitial fibrosis. In addition, there is minimal bilateral lower lung lung fibrosis. Unchanged right heart border. Unchanged opacity in the left lung may also reflect.	*minimal,fusion,pre,all,preoperative,pneumo,lung, atelectasis,left,pneumonia,small* 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 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.	*multiple,thorax,fusion,healing,all,pneumo,clear,lung, unremarkable,unchanged,right,intact,small,rib,moderate, expanded,change* Multiple rib fractures are unchanged. The right lung is well-expanded and clear. Moderate cardiomununrerereable. Small, left pleural effusion is unchanged. Fusion has multiple rib fracture	*apical,enlargement,valve,thorax,fusion,metallic,ascending, fractured,intrathoracic,all,thoracic,pneumo,chest,edema, enlarged,vasculature,patient,rib,moderate,large,contour,atrium, comparison,fracture* 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 ___. The cardiac enlargement is moderately enlarged. Mild degree of intrathoracic enlargement as well as widening of the ascending aortic contour suggests 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.
Keyword and generated report in Cluster ([10,100]+[100,1000]+[1000,10000])+[100000+]	*minimal,diffuse,border,fusion,fibrosis,carcinoma,carcinomatosis, effusion,no* Cardiac fibrosis, lymphomatosis carcinoma carcinomatosis in the right hilar and diffuse fibrosis is consistent with known lung carcinoma carcinomatosis. The right border of the cardiac border is minimally decreased. Diffuse fibrosis accompanied by diffuse fibrosis fibrosis. No pleural effusion. No evidence of pleural effusion.	*minimal,fusion,pre,all,preoperative,pneumo,effusion,no, normal* preoperative pneumo preoperative pneumo. All preoperative fusion is in preoperative position. There is minimal cardiomo preoperative pneumo, all in the preoperative right pancrecrecrecrection. No pleural effusion. Normal mediastinal and cardiac contours.	*multiple,thorax,fusion,healing,all,pneumo,pneumothorax, effusion,no,lungs* Multiple healing thorax. All of the healing tubing is seen in the right hemi thorax. The lungs are well-expanded and clear without evidence of pneumo, effusion or pneumothorax. No pleural effusion. No pneumo or pneumothorax. Multiple healing fractures within the lungs.	*apical,enlargement,valve,thorax,fusion,metallic,ascending, fractured,intrathoracic,all,thoracic,pneumo,pneumothorax,effusion* 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 sternotomy wires are unchanged in position. Fusion hardware within the intrathoracic thoracic aortic valve is unchanged. Multi-level thoracic vertebral bodies appear fractured.
Keyword and generated report in Cluster [10000,100000]+[100000+]	*effusion,no,chest,lung,heart,atelectasis,lateral,unchanged,opacity, mild,bilateral,low,interstitial,lower,change* AP and lateral views of the chest. Lung volumes are low. Heart size is unchanged. Mild bilateral interstitial prominence is unchanged. There is mild lower lung lung effusion. Retrocardiac opacity likely reflects atelectasis. Overall, there has been no significant change. Bibasilar atelectasis remains unchanged. Mild heart size is unchanged.	*effusion,no,normal,lung,atelectasis,left,pneumonia,small* The lung volumes are normal. Normal size of the cardiac silhouette. Small left pleural effusion. No evidence of pneumonia. Left retrocardiac atelectasis is small.	*pneumothorax,effusion,no,lungs,clear,lung,unremarkable, unchanged,right,intact,small,rib,moderate,expanded,change* Unremarkable cardiomedias. The lungs are well expanded and clear. Moderate cardiomegaly is unchanged. Small right plepleural effusion is unchanged. There is no pneumothorax. Sternal wires are intact. No significant change from the prior exam. Multiple rib fractures are again seen. Lungs are otherwise unremarkable. The lungs are well expanded and clear. Right lung is unremarkable.	*pneumothorax,effusion,chest,edema,enlarged,vasculature,patient, rib,moderate,large,contour,atrium,comparison,fracture* AP single view of the chest has been obtained with patient in comparison 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 comparison with the present present present present present atrium. In comparison with the current study of the patient's known fracture though central pulmonary vasculature is compatible with edem

Fig. 16. Visualization of ablation study results validating the frequency-based multi-label classification on the MIMIC-CXR dataset. Blue text highlights the predicted keywords within their respective clusters.

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 utilized a modified version of the Text-to-Text Transformer (T5) model [108], a medium-sized LLM, and adopted the pretrained configuration of Clinical-T5 [109] to initialize the model. To evaluate the impact of different LLM versions and pretrained settings, we compared the performance of the original T5 model with its advanced variant Flan-T5 [124], as well as with larger LLMs, including BART [125] and Pegasus [126], 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 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

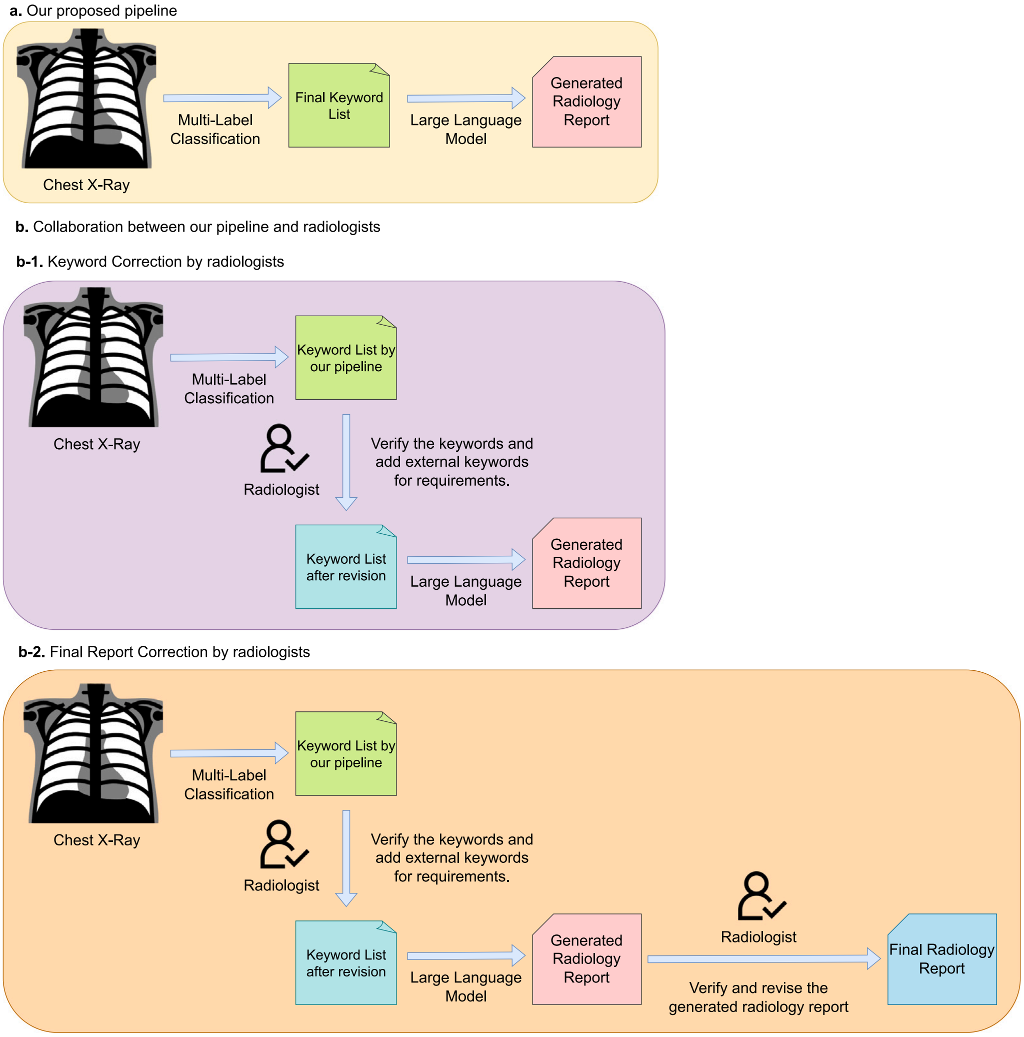

a. Our proposed pipeline
Chest X-Ray
Multi-Label Classification
Final Keyword List
Large Language Model
Generated Radiology Report
b. Collaboration between our pipeline and radiologists
b-1. Keyword Correction by radiologists
Chest X-Ray
Multi-Label Classification
Keyword List by our pipeline
Radiologist
Verify the keywords and add external keywords for requirements.
Keyword List after revision
Large Language Model
Generated Radiology Report
b-2. Final Report Correction by radiologists
Chest X-Ray
Multi-Label Classification
Keyword List by our pipeline
Radiologist
Verify the keywords and add external keywords for requirements.
Keyword List after revision
Large Language Model
Generated Radiology Report
Radiologist
Verify and revise the generated radiology report
Final Radiology Report

Fig. 17. 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.

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 ± 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. [61]	0.441	0.299
Parres et al. [83]	/	0.354
Re-Train Performance		
R2Gen ([29])	0.2545 ± 0.1395	0.1096 ± 0.1248
Cvt2Distgen2([60])	0.2497 ± 0.1494	0.1056 ± 0.1279
Our Performance	0.5980 ± 0.1651	0.3840 ± 0.2106

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 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 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 LLMs a practical and scalable choice for diverse medical applications, balancing computational efficiency with clinical effectiveness.

5. Conclusion

This paper 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

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 methods.

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

(*continued on next page*)

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 [127–129]. 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" [5], 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 [130] 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.

Table 11 (*continued*).

Network	Frequency cluster	Accuracy	Sensitivity	Specificity	F1-score	MCC
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 ± 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 ± 0.2101	0.6250 ± 0.2713	0.5645 ± 0.3057	0.5215 ± 0.3376	0.6392 ± 0.2554	0.3861 ± 0.2218	3.2749 ± 3.0106
ResNeXt	0.4802 ± 0.1282	0.4299 ± 0.1599	0.3965 ± 0.1831	0.3718 ± 0.2032	0.4484 ± 0.1561	0.2732 ± 0.1366	2.4615 ± 1.9166
ResNet	0.2453 ± 0.0655	0.2196 ± 0.0817	0.2025 ± 0.0935	0.1899 ± 0.1038	0.2291 ± 0.0698	0.1395 ± 0.0797	1.2574 ± 0.9790
VGG16	0.3170 ± 0.0846	0.2838 ± 0.1055	0.2617 ± 0.1209	0.2454 ± 0.1341	0.2960 ± 0.0902	0.1803 ± 0.1030	1.6250 ± 1.2653
EfficientNet	0.1803 ± 0.0481	0.1614 ± 0.0600	0.1488 ± 0.0687	0.1396 ± 0.0763	0.1684 ± 0.0513	0.1026 ± 0.0586	0.9241 ± 0.7196
NASNet	0.2332 ± 0.0622	0.2088 ± 0.0776	0.1925 ± 0.0889	0.1805 ± 0.0986	0.2177 ± 0.0663	0.1326 ± 0.0758	1.1952 ± 0.9306
Res2Net	0.2144 ± 0.0572	0.1920 ± 0.0714	0.1770 ± 0.0817	0.1660 ± 0.0907	0.2002 ± 0.0610	0.1220 ± 0.0697	1.0990 ± 0.8557
MIMIC-CXR Test set							
ConvNeXt	0.5599 ± 0.1607	0.4379 ± 0.1736	0.3557 ± 0.1824	0.2953 ± 0.1958	0.4699 ± 0.1687	0.2842 ± 0.1018	1.9964 ± 1.4518
ResNeXt	0.3557 ± 0.1001	0.2775 ± 0.1081	0.2247 ± 0.1133	0.1860 ± 0.1209	0.2945 ± 0.1044	0.1795 ± 0.0629	0.6221 ± 0.9035
ResNet	0.3415 ± 0.0961	0.2664 ± 0.1038	0.2157 ± 0.1088	0.1785 ± 0.1161	0.2827 ± 0.1002	0.1724 ± 0.0604	0.5972 ± 0.8673
VGG16	0.3572 ± 0.1006	0.2787 ± 0.1086	0.2256 ± 0.1138	0.1868 ± 0.1214	0.2957 ± 0.1048	0.1803 ± 0.0632	0.6247 ± 0.9073
EfficientNet	0.4096 ± 0.1153	0.3195 ± 0.1245	0.2587 ± 0.1304	0.2141 ± 0.1392	0.3391 ± 0.1202	0.2067 ± 0.0725	0.7163 ± 1.0403
NASNet	0.4336 ± 0.1221	0.3383 ± 0.1318	0.2739 ± 0.1381	0.2267 ± 0.1474	0.3590 ± 0.1272	0.2189 ± 0.0767	0.7584 ± 1.1014
Res2Net	0.4093 ± 0.1152	0.3193 ± 0.1244	0.2585 ± 0.1303	0.2140 ± 0.1391	0.3388 ± 0.1201	0.2066 ± 0.0724	0.7157 ± 1.0395

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.

Setting	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr
IU X-ray Test							
Single Cluster							
Cluster [10,100]	0.0038 ± 0.0552	0.0016 ± 0.0300	0.0010 ± 0.0189	0.0006 ± 0.0110	0.0702 ± 0.0595	0.0238 ± 0.0310	0.0141 ± 0.0853
Cluster [100,1000]	0.1555 ± 0.1520	0.0989 ± 0.1151	0.0712 ± 0.0974	0.0532 ± 0.0852	0.1859 ± 0.1262	0.1008 ± 0.0765	0.2240 ± 0.4427
Cluster [1000,10000]	0.4443 ± 0.1833	0.3149 ± 0.1824	0.2407 ± 0.1889	0.1913 ± 0.1974	0.3645 ± 0.1799	0.2162 ± 0.1207	0.6171 ± 1.0652
Mixed Cluster							
Cluster [10,100]+[100,1000]	0.2083 ± 0.1654	0.1337 ± 0.1284	0.0973 ± 0.1081	0.0734 ± 0.0936	0.2108 ± 0.1348	0.1146 ± 0.0791	0.3096 ± 0.4984
Cluster [10,100]+[1000,10000]	0.4568 ± 0.1799	0.3207 ± 0.1798	0.2402 ± 0.1833	0.1857 ± 0.1881	0.3686 ± 0.1669	0.2161 ± 0.1131	0.6073 ± 1.0344
Cluster [100,1000]+[1000,10000]	0.5013 ± 0.2028	0.3828 ± 0.2381	0.3140 ± 0.2633	0.2680 ± 0.2811	0.4331 ± 0.2354	0.2729 ± 0.1675	1.2750 ± 2.1064
Cluster All Keywords	0.7190 ± 0.2101	0.6250 ± 0.2713	0.5645 ± 0.3057	0.5215 ± 0.3376	0.6392 ± 0.2554	0.3861 ± 0.2218	3.2749 ± 3.0106

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 [131]. 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 Fig. 17, 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.

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

Table 14
Performance comparison of radiology report generation on the MIMIC-CXR dataset across different keyword clusters. Results are presented 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.

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

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 ±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 ± 0.1808	0.5520 ± 0.2157	0.4840 ± 0.2425	0.4290 ± 0.2946	0.5894 ± 0.2294	0.3452 ± 0.1619	1.5191 ± 1.5254
T5-base	Medical	222 M	0.6906 ± 0.2187	0.6163 ± 0.2683	0.5678 ± 0.3050	0.5333 ± 0.3357	0.6300 ± 0.2717	0.3777 ± 0.1973	2.8264 ± 2.4610
Flan-T5-Small	General	77.0 M	0.6707 ± 0.2112	0.6010 ± 0.2594	0.5550 ± 0.2957	0.5215 ± 0.3248	0.6618 ± 0.2618	0.3915 ± 0.2064	3.1248 ± 2.6836
Flan-T5-Small	Medical	77.0 M	0.6970 ± 0.1818	0.6149 ± 0.2353	0.5594 ± 0.2755	0.5191 ± 0.3087	0.6609 ± 0.2405	0.4055 ± 0.1880	3.0916 ± 2.7142
Flan-T5-base	General	247 M	0.6724 ± 0.2287	0.5923 ± 0.2757	0.5414 ± 0.3146	0.5064 ± 0.3423	0.6158 ± 0.2774	0.3579 ± 0.2179	2.7875 ± 2.6798
Flan-T5-base	Medical	247 M	0.7141 ± 0.1661	0.6107 ± 0.2066	0.5362 ± 0.2372	0.4766 ± 0.2807	0.6412 ± 0.2139	0.3807 ± 0.1421	2.1922 ± 1.9878
BART	General	139 M	0.6750 ± 0.1528	0.6023 ± 0.1881	0.5486 ± 0.2114	0.5042 ± 0.2518	0.5914 ± 0.2211	0.3747 ± 0.1547	1.6933 ± 0.9086
BART	Medical	139 M	0.6958 ± 0.1295	0.6089 ± 0.1695	0.5458 ± 0.1983	0.4949 ± 0.2375	0.5905 ± 0.2022	0.3834 ± 0.1351	1.8835 ± 1.0064
Pegasus	General	272 M	0.6691 ± 0.1886	0.5903 ± 0.2272	0.5320 ± 0.2553	0.4840 ± 0.2980	0.6452 ± 0.2364	0.3717 ± 0.1697	1.9380 ± 1.8382
Pegasus	Medical	272 M	0.7108 ± 0.1685	0.6238 ± 0.2094	0.5594 ± 0.2382	0.5061 ± 0.2744	0.6722 ± 0.2224	0.4009 ± 0.1502	2.3568 ± 2.0371
Ours	/	60.5 M	0.7190 ± 0.2101	0.6250 ± 0.2713	0.5645 ± 0.3057	0.5215 ± 0.3376	0.6392 ± 0.2554	0.3861 ± 0.2218	3.2749 ± 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 ±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 ± 0.1831	0.3176 ± 0.1810	0.2495 ± 0.1816	0.2020 ± 0.1813	0.3788 ± 0.1710	0.2153 ± 0.1136	0.6195 ± 1.3788
T5-base	Medical	222 M	0.5392 ± 0.1728	0.4139 ± 0.1839	0.3315 ± 0.1918	0.2733 ± 0.1971	0.4457 ± 0.1777	0.2745 ± 0.1126	1.1082 ± 1.7353
Flan-T5-Small	General	77.0 M	0.4521 ± 0.1720	0.3484 ± 0.1722	0.2780 ± 0.1726	0.2264 ± 0.1823	0.4090 ± 0.1613	0.2284 ± 0.0997	0.5560 ± 1.0986
Flan-T5-Small	Medical	77.0 M	0.5337 ± 0.1713	0.4096 ± 0.1814	0.3279 ± 0.1895	0.2702 ± 0.1947	0.4438 ± 0.1784	0.2726 ± 0.1110	1.0907 ± 1.6834
Flan-T5-base	General	247 M	0.4536 ± 0.1703	0.3485 ± 0.1702	0.2774 ± 0.1710	0.2253 ± 0.1806	0.4063 ± 0.1606	0.2278 ± 0.0993	0.5574 ± 1.1004
Flan-T5-base	Medical	247 M	0.5500 ± 0.1657	0.4215 ± 0.1788	0.3374 ± 0.1882	0.2776 ± 0.1940	0.4457 ± 0.1741	0.2769 ± 0.1107	1.1310 ± 1.6780
BART	General	139 M	0.4730 ± 0.1613	0.3571 ± 0.1621	0.2800 ± 0.1633	0.2246 ± 0.1738	0.3945 ± 0.1560	0.2292 ± 0.0961	0.5461 ± 1.0467
BART	Medical	139 M	0.5223 ± 0.2960	0.4245 ± 0.3278	0.3458 ± 0.3483	0.2864 ± 0.3619	0.4501 ± 0.3020	0.2753 ± 0.2751	1.3830 ± 3.4673
Pegasus	General	272 M	0.4505 ± 0.1706	0.3456 ± 0.1700	0.2750 ± 0.1703	0.2230 ± 0.1797	0.4034 ± 0.1583	0.2268 ± 0.0993	0.5414 ± 1.0983
Pegasus	Medical	272 M	0.5538 ± 0.1621	0.4329 ± 0.1727	0.3515 ± 0.1798	0.2914 ± 0.1916	0.4677 ± 0.1635	0.2819 ± 0.1004	0.9722 ± 1.4213
Ours	/	60.5 M	0.5599 ± 0.1607	0.4379 ± 0.1736	0.3557 ± 0.1824	0.2953 ± 0.1958	0.4699 ± 0.1687	0.2842 ± 0.1018	1.9964 ± 1.4518

occasional semantic mismatches, where generated text misrepresents or contradicts the intended meaning of the keywords. Inspired by works like [132], 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.

CRediT authorship contribution statement

Zebang He: Writing – review & editing, Writing – original draft, Software, Data curation, Conceptualization. **Alex Ngai Nick Wong:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Jung Sun Yoo:** Writing – review & editing, Supervision, Conceptualization.

Funding

The work described in this paper was supported by the internal grant of The Hong Kong Polytechnic University and the General Research Fund (Project No. PolyU 15101422) from the Research Grants Council of the Hong Kong Special Administrative Region of China.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve the clearness of paragraph. After using this tool /service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All of the data utilized in the work reported in this paper, which are IU-XRay and MIMIC-CXR, are publicly available data.

References

[1] I.A. Cowan, S.L.S. MacDonald, R.A. Floyd, Measuring and managing radiologist workload: Measuring radiologist reporting times using data from a radiology information system, J. Med. Imaging Radiat. Oncol. 57 (2013) URL: https://api.semanticscholar.org/CorpusID:46325246.

[2] Cadth, Canadian medical imaging inventory 2022–2023: The medical imaging team, Can. J. Heal. Technol. (2024) URL: https://api.semanticscholar.org/CorpusID:272031953.

[3] S. Mayor, Waiting times for x ray results in England are increasing, figures show, BMJ Br. Med. J. 350 (2015) URL: https://api.semanticscholar.org/CorpusID:35737094.

[4] N. Woznitza, A. Devaraj, S.M. Janes, S.W. Duffy, A. Bhowmik, S. Rowe, K. Piper, S. Maughn, D.R. Baldwin, Impact of radiographer immediate reporting of chest X-rays from general practice on the lung cancer pathway (radiox): study protocol for a randomised control trial, Trials 18 (2017) URL: https://api.semanticscholar.org/CorpusID:30470118.

[5] M.F. Kilkenny, K.M. Robinson, Data quality:"garbage in–garbage out", Heal. Inf. Manag. J. 47 (3) (2018) 103–105.

[6] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[7] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).

[8] M. Monajatipoor, M. Rouhsedaghat, L.H. Li, C.C. Jay Kuo, A. Chien, K.W. Chang, Berthop: An effective vision-and-language model for chest x-ray disease diagnosis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 725–734.

[9] S.C. Huang, L. Shen, M.P. Lungren, S. Yeung, 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, pp. 3942–3951.

[10] Z. Wang, Z. Wu, D. Agarwal, J. Sun, Medclip: Contrastive learning from unpaired medical images and text, 2022, arXiv preprint arXiv:2210.10163.

[11] F. Cong, S. Xu, L. Guo, Y. Tian, Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3569–3577.

[12] Y. Li, H. Wang, 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, IEEE, 2020, pp. 1999–2004.

[13] P. Müller, G. Kaissis, C. Zou, D. Rueckert, Joint learning of localized representations from medical images and reports, in: European Conference on Computer Vision, Springer, 2022, pp. 685–701.

[14] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[15] B. Boecking, N. Usuyama, S. Bannur, D.C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, et al., Making the most of text semantics to improve biomedical vision–language processing, in: European Conference on Computer Vision, Springer, 2022, pp. 1–21.

[16] K. Zhang, Y. Yang, J. Yu, H. Jiang, J. Fan, Q. Huang, W. Han, Multi-task paired masking with alignment modeling for medical vision-language pre-training, IEEE Trans. Multimed. (2023).

[17] Y. Lei, Z. Li, Y. Shen, J. Zhang, H. Shan, CLIP-lung: Textual knowledge-guided lung nodule malignancy prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 403–412.

[18] E. Tiu, E. Talius, P. Patel, C.P. Langlotz, A.Y. Ng, P. Rajpurkar, Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning, Nat. Biomed. Eng. 6 (12) (2022) 1399–1406.

[19] H. Lai, Q. Yao, Z. Jiang, R. Wang, Z. He, X. Tao, S.K. Zhou, Carzero: Cross-attention alignment for radiology zero-shot classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11137–11146.

[20] S. Eslami, C. Meinel, G. De Melo, Pubmedclip: How much does clip benefit visual question answering in the medical domain? in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1181–1193.

[21] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, W. Xie, Pmc-vqa: Visual instruction tuning for medical visual question answering, 2023, arXiv preprint arXiv:2305.10415.

[22] T. Van Sonsbeek, M.M. Derakhshani, I. Najdenkoska, C.G. Snoek, M. Worring, Open-ended medical visual question answering through prefix tuning of language models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 726–736.

[23] M.P. Hartung, I.C. Bickle, F. Gaillard, J.P. Kanne, How to create a great radiology report, Radiographics 40 (6) (2020) 1658–1670.

[24] G. Boland, A. Guimaraes, P. Mueller, Radiology report turnaround: expectations and solutions, Eur. Radiol. 18 (2008) 1326–1328.

[25] M.M.A. Monshi, J. Poon, V. Chung, Deep learning in generating radiology reports: A survey, Artif. Intell. Med. 106 (2020) 101878.

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

[27] A.E. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.y. Deng, R.G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (1) (2019) 317.

[28] B. Jing, P. Xie, E.P. Xing, On the Automatic Generation of Medical Imaging Reports, in: Annual Meeting of the Association for Computational Linguistics, 2017, URL: https://api.semanticscholar.org/CorpusID:5776384.

[29] Z. Chen, Y. Song, T.H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, 2020, ArXiv, arXiv:2010.16056, URL: https://api.semanticscholar.org/CorpusID:226222210.

[30] Z. Wang, M. Tang, L. Wang, X. Li, L. Zhou, A medical semantic-assisted transformer for radiographic report generation, 2022, ArXiv, arXiv:2208.10358, URL: https://api.semanticscholar.org/CorpusID:251719362.

[31] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, X. Wu, AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation, 2022, ArXiv, arXiv:2203.10095, URL: https://api.semanticscholar.org/CorpusID:238208093.

[32] X. Xie, Y. Xiong, P.S. Yu, K. Li, S. Zhang, Y. Zhu, Attention-Based Abnormal-Aware Fusion Network for Radiology Report Generation, in: International Conference on Database Systems for Advanced Applications, 2019, URL: https://api.semanticscholar.org/CorpusID:129949265.

[33] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, X. Sun, Contrastive attention for automatic chest X-ray report generation, Findings (2021) URL: https://api.semanticscholar.org/CorpusID:235422047.

[34] F. Nooralahzadeh, N.A.P. Gonzalez, T. Frauenfelder, K. Fujimoto, M. Krauthammer, Progressive transformer-based generation of radiology reports, 2021, ArXiv, arXiv:2102.09777, URL: https://api.semanticscholar.org/CorpusID:231979557.

[35] K. Kale, P. Bhattacharyya, M. Gune, A. Shetty, R. Lawyer, 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, pp. 3401–3411.

[36] S. Jain, A. Agrawal, A. Saporta, S.Q. Truong, D.N. Duong, T. Bui, P. Chambon, Y. Zhang, M.P. Lungren, A.Y. Ng, et al., Radgraph: Extracting clinical entities and relations from radiology reports, 2021, arXiv preprint arXiv:2106.14463.

[37] F.D. Serra, W. Clackett, H. MacKinnon, C. Wang, F. Deligianni, J. Dalton, A.Q. O'Neil, Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations, AACL (2022) URL: https://api.semanticscholar.org/CorpusID:253762055.

[38] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A.L. Yuille, D. Xu, When radiology report generation meets knowledge graph, in: AAAI Conference on Artificial Intelligence, 2020, URL: https://api.semanticscholar.org/CorpusID:211171529.

[39] Z. Wang, L. Liu, L. Wang, L. Zhou, R2gengpt: Radiology report generation with frozen llms, Meta Radiol. 1 (3) (2023) 100033.

[40] M. Soleimani, N. Seyyedi, S.M. Ayyoubzadeh, S.R.N. Kalhori, H. Keshavarz, Practical evaluation of ChatGPT performance for radiology report generation, Academic Radiol. (2024).

[41] Y. Xue, T. Xu, L.R. Long, Z. Xue, S.K. Antani, G.R. Thoma, X. Huang, Multimodal recurrent model with attention for automated radiology report generation, Int. Conf. Med. Image Comput. Comput. Assist. Interv. (2018) URL: https://api.semanticscholar.org/CorpusID:52275869.

[42] P. Harzig, Y.Y. Chen, F. Chen, R. Lienhart, Addressing Data Bias Problems for Chest X-ray Image Report Generation, in: British Machine Vision Conference, 2019, URL: https://api.semanticscholar.org/CorpusID:199453142.

[43] J. Yuan, H. Liao, R. Luo, J. Luo, Automatic radiology report generation based on multi-view image fusion and medical concept enrichment, 2019, ArXiv, arXiv:1907.09085, URL: https://api.semanticscholar.org/CorpusID:198148007.

[44] Y. Li, X. Liang, Z. Hu, E.P. Xing, Knowledge-driven encode, retrieve, paraphrase for medical image report generation, 2019, ArXiv, arXiv:1903.10122, URL: https://api.semanticscholar.org/CorpusID:59276384.

[45] B. Jing, Z. Wang, E.P. Xing, Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports, in: Annual Meeting of the Association for Computational Linguistics, 2019, URL: https://api.semanticscholar.org/CorpusID:196199713.

[46] Z. Wang, L. Zhou, L. Wang, X. Li, A self-boosting framework for automated radiographic report generation, CVPR, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2433–2442, URL: https://api.semanticscholar.org/CorpusID:235703340.

[47] O. Alfarghaly, R. Khaled, A.M. Elkorany, M. Helal, A. Fahmy, Automated radiology report generation using conditioned transformers, Inform. Med. Unlocked 24 (2021) 100557, URL: https://api.semanticscholar.org/CorpusID:233855443.

[48] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, CVPR, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13748–13757, URL: https://api.semanticscholar.org/CorpusID:235421693.

[49] S. Yang, X. Wu, S. Ge, S. Zhou, L. Xiao, Knowledge matters: Chest radiology report generation with general and specific knowledge, Med. Image Anal. 80 (2021) 102510, URL: https://api.semanticscholar.org/CorpusID:249557147.

[50] S. Yang, X. Wu, S. Ge, X. Wu, S. Zhou, L. Xiao, Radiology report generation with a learned knowledge base and multi-modal alignment, Med. Image Anal. 86 (2021) 102798, URL: https://api.semanticscholar.org/CorpusID:245634857.

[51] Y. Zhou, L. Huang, T. Zhou, H. Fu, L. Shao, Visual-Textual Attentive Semantic Consistency for Medical Report Generation, ICCV, in: 2021 IEEE/CVF International Conference on Computer Vision, 2021, pp. 3965–3974, URL: https://api.semanticscholar.org/CorpusID:263876333.

[52] J. Li, S. Li, Y. Hu, H. Tao, A self-guided framework for radiology report generation, 2022, ArXiv, arXiv:2206.09378, URL: https://api.semanticscholar.org/CorpusID:249890082.

[53] L. Wang, M. Ning, D. Lu, D. Wei, Y. Zheng, J. lian Chen, An inclusive task-aware framework for radiology report generation, Int. Conf. Med. Image Comput. Comput. Assist. Interv. (2022) URL: https://api.semanticscholar.org/CorpusID:252369300.

[54] M. Sirshar, M.F.K. Paracha, M.U. Akram, N.S. Alghamdi, S.Z.Y. Zaidi, T. Fatima, Attention based automated radiology report generation using CNN and LSTM, PLoS One 17 (2022) URL: https://api.semanticscholar.org/CorpusID:245801513.

[55] S. Yan, W.K. Cheung, K.W.H. Chiu, T.M. Tong, C.K. Cheung, S. See, Attributed abnormality graph embedding for clinically accurate X-Ray report generation, IEEE Trans. Med. Imaging 42 (2022) 2211–2222, URL: https://api.semanticscholar.org/CorpusID:250264390.

[56] Z. Wang, H. Han, L. Wang, X. Li, L. Zhou, Automated radiographic report generation purely on transformer: A multicriteria supervised approach, IEEE Trans. Med. Imaging 41 (2022) 2803–2813, URL: https://api.semanticscholar.org/CorpusID:248514098.

[57] H. Yu, Q. Zhang, Clinically Coherent Radiology Report Generation with Imbalanced Chest X-rays, BIBM, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, 2022, pp. 1781–1786, URL: https://api.semanticscholar.org/CorpusID:255417861.

[58] Z. Chen, Y. Shen, Y. Song, X. Wan, Cross-modal memory networks for radiology report generation, 2022, ArXiv, arXiv:2204.13258, URL: https://api.semanticscholar.org/CorpusID:236460168.

[59] J. Wang, A. Bhalerao, Y. He, Cross-modal prototype driven network for radiology report generation, 2022, ArXiv, arXiv:2207.04818, URL: https://api.semanticscholar.org/CorpusID:250426167.

[60] A. Nicolson, J.A. Dowling, B. Koopman, Improving chest X-Ray report generation by leveraging warm-starting, Artif. Intell. Med. 144 (2022) 102633, URL: https://api.semanticscholar.org/CorpusID:246240082.

[61] J.B. Delbrouck, P. Chambon, C. Blüthgen, E.B. Tsai, O. Almusa, C.P. Langlotz, Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards, in: Conference on Empirical Methods in Natural Language Processing, 2022, URL: https://api.semanticscholar.org/CorpusID:253098780.

[62] J. You, D. Li, M. Okumura, K. Suzuki, JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation, in: International Conference on Computational Linguistics, 2022, URL: https://api.semanticscholar.org/CorpusID:252818921.

[63] X. Wu, J. Li, J. Wang, Q. Qian, Multimodal contrastive learning for radiology report generation, J. Ambient. Intell. Humaniz. Comput. 14 (2022) 11185–11194, URL: https://api.semanticscholar.org/CorpusID:252228943.

[64] B. Yan, M. Pei, M. Zhao, C. Shan, Z. Tian, Prior guided transformer for accurate radiology reports generation, IEEE J. Biomed. Heal. Inform. 26 (2022) 5631–5640, URL: https://api.semanticscholar.org/CorpusID:251401224.

[65] S. Wang, L. Tang, M. Lin, G.L. Shih, Y. Ding, Y. Peng, Prior knowledge enhances radiology report generation, in: AMIA ... Annual Symposium proceedings, in: AMIA Symposium, vol. 2022, 2022, pp. 486–495, URL: https://api.semanticscholar.org/CorpusID:245853976.

[66] H. Qin, Y. Song, Reinforced cross-modal alignment for radiology report generation, Findings (2022) URL: https://api.semanticscholar.org/CorpusID:248780118.

[67] A.K. Tanwani, J. Barral, D. Freedman, Repsnet: Combining vision with language for automated medical reports, Int. Conf. Med. Image Comput. Comput. Assist. Interv. (2022) 714–724.

[68] Y. Wang, K. Wang, X. Liu, T. Gao, J. Zhang, G. Wang, Self adaptive global-local feature enhancement for radiology report generation, ICIP, in: 2023 IEEE International Conference on Image Processing, IEEE, 2023, pp. 2275–2279.

[69] M. Kong, Z. Huang, K. Kuang, Q. Zhu, F. Wu, Transq: Transformer-based semantic query for medical report generation, Int. Conf. Med. Image Comput. Comput. Assist. Interv. (2022) 610–620.

[70] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, X. Chang, 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, pp. 3334–3343.

[71] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, Q. Huang, Joint embedding of deep visual and semantic features for medical image report generation, IEEE Trans. Multimed. 25 (2021) 167–178.

[72] Z. Huang, X. Zhang, 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, pp. 19809–19818.

[73] Z. Wang, L. Liu, L. Wang, L. Zhou, Metransformer: Radiology report generation by transformer with multiple learnable expert tokens, CVPR, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11558–11567.

[74] W. Hou, K. Xu, Y. Cheng, W. Li, J. Liu, ORGAN: observation-guided radiology report generation via tree reasoning, 2023, arXiv preprint arXiv:2306.06466.

[75] K. Kale, K. Jadhav, et al., Replace and report: NLP assisted radiology report generation, 2023, arXiv preprint arXiv:2306.17180.

[76] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, Y. Zou, Unify, align and refine: Multi-level semantic alignment for radiology report generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2863–2874.

[77] M.M. Mohsan, M.U. Akram, G. Rasool, N.S. Alghamdi, M.A.A. Baqai, M. Abbas, Vision transformer and language model based radiology report generation, IEEE Access 11 (2022) 1814–1824.

[78] W. Chen, Y. Liu, C. Wang, G. Li, J. Zhu, L. Lin, Visual-linguistic causal intervention for radiology report generation, 1, (8) 2023, arXiv preprint arXiv:2303.09117.

[79] K. Zhang, H. Jiang, J. Zhang, Q. Huang, J. Fan, J. Yu, W. Han, Semi-supervised medical report generation via graph-guided hybrid feature consistency, IEEE Trans. Multimed. 26 (2023) 904–915.

[80] C. Liu, Y. Tian, W. Chen, Y. Song, Y. Zhang, Bootstrapping Large Language Models for Radiology Report Generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (17) 2024, pp. 18635–18643.

[81] Z. Zhou, M. Shi, M. Wei, O. Alabi, Z. Yue, T. Vercauteren, Large model driven radiology report generation with clinical quality reinforcement learning, 2024, arXiv preprint arXiv:2403.06728.

[82] X. Yi, Y. Fu, R. Liu, H. Zhang, R. Hua, TSGET: Two-stage global enhanced transformer for automatic radiology report generation, IEEE J. Biomed. Heal. Inform. (2024).

[83] D. Parres, A. Albiol, R. Paredes, Improving radiology report generation quality and diversity through reinforcement learning and text augmentation, Bioengineering 11 (4) (2024) 351.

[84] X. Yi, Y. Fu, J. Yu, R. Liu, H. Zhang, R. Hua, LHR-RFL: Linear hybrid-reward-based reinforced focal learning for automatic radiology report generation, IEEE Trans. Med. Imaging 44 (2024) 1494–1504, URL: https://api.semanticscholar.org/CorpusID:274366843.

[85] B. Hou, G. Kaissis, R.M. Summers, B. Kainz, RATCHET: Medical transformer for chest X-ray diagnosis and reporting, 2021, ArXiv, arXiv:2107.02104, URL: https://api.semanticscholar.org/CorpusID:235732014.

[86] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, C.N. Hsu, Weakly supervised contrastive learning for chest x-ray report generation, 2021, arXiv preprint arXiv:2109.12242.

[87] T. Nishino, Y. Miura, T. Taniguchi, T. Ohkuma, Y. Suzuki, S. Kido, N. Tomiyama, Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation, in: Conference on Empirical Methods in Natural Language Processing, 2022, URL: https://api.semanticscholar.org/CorpusID:256460887.

[88] T. Tanida, P. Müller, G. Kaissis, D. Rueckert, Interactive and explainable region-guided radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7433–7442.

[89] K. Zhang, Y. Yang, J. Yu, J. Fan, H. Jiang, Q. Huang, W. Han, Attribute prototype-guided iterative scene graph for explainable radiology report generation, IEEE Trans. Med. Imaging (2024).

[90] P.H. Wu, A. Yu, C.W. Tsai, J.L. Koh, C.C. Kuo, A.L. Chen, Keyword extraction and structuralization of medical reports, Heal. Inf. Sci. Syst. 8 (2020) 1–25.

[91] Y. Kim, J.H. Lee, S. Choi, J.M. Lee, J.H. Kim, J. Seok, H.J. Joo, Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records, Sci. Rep. 10 (1) (2020) 20265.

[92] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (2011) 333–359.

[93] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, H. Wang, SGM: sequence generation model for multi-label classification, 2018, arXiv preprint arXiv:1806.04822.

[94] J. Wehrmann, R. Cerri, R. Barros, Hierarchical multi-label classification networks, in: International Conference on Machine Learning, PMLR, 2018, pp. 5075–5084.

[95] S. Liu, L. Zhang, X. Yang, H. Su, J. Zhu, Query2label: A simple transformer way to multi-label classification, 2021, arXiv preprint arXiv:2107.10834.

[96] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, S. Wen, Cross-modality attention with semantic graph embedding for multi-label classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (07) 2020, pp. 12709–12716.

[97] S.F. Chen, Y.C. Chen, C.K. Yeh, Y.C. Wang, Order-free rnn with visual attention for multi-label classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (1) 2018.

[98] J. Yuan, S. Chen, Y. Zhang, Z. Shi, X. Geng, J. Fan, Y. Rui, Graph attention transformer network for multi-label image classification, ACM Trans. Multimed. Comput. Commun. Appl. 19 (4) (2023) 1–16.

[99] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 82–91.

[100] T. Wu, Q. Huang, Z. Liu, Y. Wang, D. Lin, 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, Springer, 2020, pp. 162–178.

[101] T. Durand, N. Mehrasa, 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, pp. 647–657.

[102] B. Chen, J. Li, G. Lu, H. Yu, D. Zhang, Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification, IEEE J. Biomed. Heal. Inform. 24 (8) (2020) 2292–2302.

[103] C. Ma, H. Wang, 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, Springer, 2019, pp. 730–738.

[104] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, R. Chakravorty, Chest x-rays classification: A multi-label and fine-grained problem, 2018, arXiv preprint arXiv:1807.07247.

[105] P. Sharma, Y. Li, Self-supervised contextual keyword and keyphrase retrieval with self-labelling, 2019, Preprints.

[106] M. Grootendorst, KeyBERT: Minimal Keyword Extraction with BERT, Zenodo, 2020, http://dx.doi.org/10.5281/zenodo.4461265, URL: https://doi.org/10.5281/zenodo.4461265.

[107] C.P. Langlotz, RadLex: a new method for indexing online educational materials, Radiographics 26 (6) (2006) 1595–1597.

[108] C. Raffel, N.M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2019) 140:1–140:67, URL: https://api.semanticscholar.org/CorpusID:204838007.

[109] Q. Lu, D. Dou, T.H. Nguyen, ClinicalT5: A generative language model for clinical text, in: Conference on Empirical Methods in Natural Language Processing, 2022, URL: https://api.semanticscholar.org/CorpusID:256631112.

[110] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C. ying Deng, R.G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (2019) URL: https://api.semanticscholar.org/CorpusID:209342303.

[111] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, CVPR, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11966–11976, URL: https://api.semanticscholar.org/CorpusID:245837420.

[112] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, HuggingFace's transformers: State-of-the-art natural language processing, 2019, ArXiv, arXiv:1910.03771, URL: https://api.semanticscholar.org/CorpusID:269498120.

[113] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Annual Meeting of the Association for Computational Linguistics, 2002, URL: https://api.semanticscholar.org/CorpusID:11080756.

[114] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Annual Meeting of the Association for Computational Linguistics, 2004, URL: https://api.semanticscholar.org/CorpusID:964287.

[115] R. Vedantam, C.L. Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, CVPR, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4566–4575, URL: https://api.semanticscholar.org/CorpusID:9026666.

[116] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: IEEvaluation@ACL, 2005, URL: https://api.semanticscholar.org/CorpusID:7164502.

[117] J.A. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R.L. Ball, K.S. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J.K. Sandberg, R. Jones, D.B. Larson, C. Langlotz, B.N. Patel, M.P. Lungren, A. Ng, CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: AAAI Conference on Artificial Intelligence, 2019, URL: https://api.semanticscholar.org/CorpusID:58981871.

[118] S. Xie, R.B. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, CVPR, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5987–5995, URL: https://api.semanticscholar.org/CorpusID:8485068.

[119] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, CVPR, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 770–778, URL: https://api.semanticscholar.org/CorpusID:206594692.

[120] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, CoRR, arXiv:1409.1556, URL: https://api.semanticscholar.org/CorpusID:14124313.

[121] M. Tan, Q.V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, 2019, ArXiv, arXiv:1905.11946, URL: https://api.semanticscholar.org/CorpusID:167217261.

[122] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning Transferable Architectures for Scalable Image Recognition, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 8697–8710, URL: https://api.semanticscholar.org/CorpusID:12227989.

[123] S. Gao, M.M. Cheng, K. Zhao, X. Zhang, M.H. Yang, P.H.S. Torr, Res2Net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2019) 652–662, URL: https://api.semanticscholar.org/CorpusID:91184391.

[124] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S.S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A.W. Yu, V. Zhao, Y. Huang, A.M. Dai, H. Yu, S. Petrov, E.H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q.V. Le, J. Wei, Scaling instruction-finetuned language models, 2022, ArXiv, arXiv:2210.11416, URL: https://api.semanticscholar.org/CorpusID:253018554.

[125] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Annual Meeting of the Association for Computational Linguistics, 2019, URL: https://api.semanticscholar.org/CorpusID:204960716.

[126] J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International Conference on Machine Learning, PMLR, 2020, pp. 11328–11339.

[127] Z. He, A.N.N. Wong, J.S. Yoo, Co-ERA-Net: Co-supervision and enhanced region attention for accurate segmentation in COVID-19 chest infection images, Bioengineering 10 (8) (2023) 928.

[128] K. Gao, J. Su, Z. Jiang, L.L. Zeng, Z. Feng, H. Shen, P. Rong, X. Xu, J. Qin, Y. Yang, et al., Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images, Med. Image Anal. 67 (2021) 101836.

[129] N. Paluru, A. Dayal, H.B. Jenssen, T. Sakinis, L.R. Cenkeramaddi, J. Prakash, P.K. Yalavarthy, Anam-net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images, IEEE Trans. Neural Netw. Learn. Syst. 32 (3) (2021) 932–946.

[130] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, et al., A pathology foundation model for cancer diagnosis and prognosis prediction, Nature 634 (8035) (2024) 970–978.

[131] R. Tanno, D.G. Barrett, A. Sellergren, S. Ghaisas, S. Dathathri, A. See, J. Welbl, C. Lau, T. Tu, S. Azizi, et al., Collaboration between clinicians and vision–language models in radiology report generation, Nature Med. (2024) 1–10.

[132] J. Chih-Yao Chen, A. Prasad, S. Saha, E. Stengel-Eskin, M. Bansal, MAgICoRe: Multi-agent, iterative, coarse-to-fine refinement for reasoning, 2024, ArXiv E-Prints, arXiv–2409.