Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 21 May 2026
Revised 27 June 2026
Accepted 22 July 2026

Natural Resources for Human Health 2026, Vol. 6 (10s): 760–772

KEYWORDS:
Radiology Report Generation, Retrieval-Augmented Generation, Gradio, Hugging Face Spaces, Clinical Decision Support

ClinCrossT5: Design and Deployment of a Retrieval-Augmented System for Automated Chest X-ray Report Generation

Nilam Sureshrao Khairnar[1], Dr. Shirish S. Sane[2]

[1]Research Scholar, Department of Computer Engineering, MET Institute of Engineering, Nashik, Maharashtra, India.
[2]Research Supervisor, MET Institute of Engineering, Nashik, Maharashtra, India; Principal and Professor, R. H. Sapat College of Engineering, Management Studies and Research, Nashik, Maharashtra, India.

ABSTRACT: Automated radiology report generation can improve reporting efficiency, but implementing research models is challenging. In this paper, we introduce ClinCrossT5, an end-to-end deployment framework for automated chest X-ray report generation, in the form of an interactive web application. Our system combines Vision Transformer (ViT-B/16) for image feature extraction, retrieval of visually similar prior cases using cosine similarity, and Clinical-T5 for language generation to produce structured Findings and Impression sections from chest radiographs and clinical indications.

The application was built using Gradio and deployed on Hugging Face Spaces. Users can upload chest X-ray images, enter clinical indications, generate structured radiology reports and export the results as Microsoft Word (.docx) documents. This deployment integrates image preprocessing, visual feature extraction, similarity retrieval, multimodal inference, and automated document generation into a browser-accessible interface.

Deployment experiments verified the successful end-to-end execution of the complete inference pipeline that included image preprocessing, feature extraction, case retrieval, report generation, visualization and document export. Latency evaluation on the production deployment using an NVIDIA T4 GPU showed a mean warm-inference latency of 6.07 s for repeated requests to a single image ($n = 9$) and 8.98 s across 20 chest X-ray images with three repeated runs per image ($n = 60$). We show representative examples generated by the deployed application on unseen chest X-ray images. The paper further presents the deployment architecture, implementation workflow and operational evaluation in terms of latency, providing a practical platform for research and decision-support evaluation of retrieval-augmented multi-modal radiology report generation systems.

1. INTRODUCTION

Chest radiography is one of the most commonly performed diagnostic imaging examinations because of its low cost, rapid acquisition and widespread availability. Radiologists analyze chest X-ray images and produce structured reports including Findings, which describe anatomical observations and imaging abnormalities, and Impression, which summarizes the most clinically significant conclusions. This task of producing accurate radiology reports, which involves careful interpretation of imaging findings with the use of clinically relevant language, is therefore complex and time-consuming.

In recent years, automated radiology report generation has been improved significantly with the help of multimodal frameworks that combine visual feature extraction and natural language generation, by using deep learning. Transformer-based models have demonstrated a strong ability of learning relations between medical images and clinical text. Retrieval-augmented generation (RAG) improves report quality by adding information of clinically relevant historical cases during inference. Retrieving visually similar prior examinations adds clinical context, improves report consistency and reduces generic or incomplete descriptions.

Despite these advances, there has been relatively little work on deployment of radiology report generation models as accessible software applications. Most of the current work is concerned with model architecture, training strategies and benchmark performance, but practical deployment involves other considerations such as efficient model initialization, retrieval database management, browser-based interaction, report formatting, cloud hosting, and robust inference. It is essential to overcome these challenges to benchmark artificial intelligence systems in realistic workflow environments.

In this paper, we propose ClinCrossT5, an end-to-end deployment framework for automated chest X-ray report generation. We introduce a web-based application developed in Gradio and deployed on Hugging Face Spaces that integrates pretrained Vision Transformer (ViT-B/16) image feature extraction, retrieval of clinically relevant prior cases based on cosine similarity, and Clinical-T5 language generation. Users can upload an image of a chest x-ray, enter the clinical indication, generate a structured radiology report with Findings and Impression, and export the report as a Microsoft Word (.docx) document.

Unlike studies largely focused on model development and benchmark accuracy, this work focuses on the deployment, integration, and operational evaluation of a retrieval-augmented multimodal reporting system. Although the underlying ClinCrossT5 model has been evaluated before using standard metrics in radiology report generation, the main contribution of this paper is the implementation and evaluation of a full deployment pipeline. It provides a practical platform for research and decision-support evaluation by integrating multimodal inference and similarity-based retrieval, report visualization, document generation, and deployment performance analysis in a single web-based environment.

This work makes three key contributions. We first introduce a web-based deployment framework for retrieval-augmented chest X-ray report generation, based on Gradio and Hugging Face Spaces. Second, we implement an end-to-end multimodal inference pipeline in a unified architecture that combines the feature extraction of the Vision Transformer, similarity-based retrieval, and Clinical-T5 report generation. Third, we evaluate the deployed system by analyzing deployment latency, assessing inference workflow, and characterizing the operational performance. These contributions provide a practical framework to translate multimodal radiology report generation research into an accessible cloud-based platform for research and decision-support evaluation.

2. RELATED WORK

Automated radiology report generation has emerged as an important research area in medical imaging and clinical natural language processing. The aim is to produce clinically accurate and coherent radiology reports directly from medical images, supporting radiologists and improving reporting efficiency. Current approaches have evolved from convolutional neural network (CNN)-based encoder-decoder models to transformer-based multimodal frameworks and, most recently, retrieval-augmented generation (RAG) methods that utilize information from clinically relevant prior cases.

Early work adapted image captioning techniques to the radiology domain using CNN-based visual feature extraction with recurrent neural network (RNN) language models. Jing et al. proposed a hierarchical LSTM framework that generates paragraph-level radiology reports by modeling sentence-level and word-level dependencies and generated more coherent reports than conventional image captioning methods [1]. Li et al. proposed Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) which combined template retrieval and reinforcement learning to improve clinical relevance and reduce repetitive text generation [2]. However, these methods were limited in modeling long-range contextual dependencies and complex clinical semantics.

The use of transformer-based architectures has greatly improved radiology report generation due to better modeling of global contextual relationships. Delbrouck et al. further tackled the issue of factual inconsistency by proposing semantic reward optimization, demonstrating that reward functions guided by clinical knowledge may improve the factual correctness of generated reports [4]. Meanwhile, vision-language pretraining has become an effective recipe for medical imaging applications. Zhou et al. showed that cross-supervised learning between chest X-ray images and free-text radiology reports improves multimodal representations for downstream clinical tasks [5].

Visual representation learning is an essential part of multimodal report generation systems. Previously, deep learning models were predominantly based on convolutional neural networks. Residual learning architectures such as ResNet were used for strong visual feature extraction for medical image analysis [6]. The transformer architecture proposed by Vaswani et al. has made the effective modeling of long-range dependency via self-attention mechanism [7]. On this foundation, Dosovitskiy et al. proposed the Vision Transformer (ViT), demonstrating that pure transformer architectures can learn effective long-range spatial relationships and perform well across image identification tasks [8]. These visual encoders have been widely used in medical imaging, where a good representation of the images is helpful for the downstream report generation. Boecking et al. emphasized the significance of leveraging text semantics for enhancing biomedical vision-language processing and image-text representation learning [9].

Pretrained language models have also helped with medical text generation. Raffel et al. proposed T5 (Text-to-Text Transfer Transformer) which unified various natural language processing tasks into a single text-to-text framework [10]. The system uses the hossboll/clinical-t5 checkpoint, a Clinical-T5-based model built on the T5 architecture and adapted for clinical text generation which offers a domain-specific language model for clinical natural language processing tasks [11]. Domain-specific language models have been demonstrated to outperform general-purpose language models in generating clinical text.

Recently, retrieval-augmented generation has become a promising approach to improve the factual consistency for report generation. Lewis et al. proposed the Retrieval-Augmented Generation framework to integrate external retrieved knowledge with neural text generation and improve the performance of knowledge-intensive generation tasks [12]. Retrieval-based models in medical imaging provide visually clinically relevant reference reports from similar past exams, helping improve contextual grounding in report generation.

Reliable evaluation remains a major challenge in automated radiology report generation. Conventional natural language generation metrics like BLEU and ROUGE measure mostly lexical similarity and may not adequately capture clinical correctness. In [13] Irvin et al. presented CheXpert, a large-scale chest radiograph dataset and benchmark for automated chest radiograph interpretation. One of the largest publicly available chest radiograph databases with associated free-text reports for development and evaluation of artificial intelligence models is MIMIC-CXR, released by Johnson et al. [14]. Smit et al. [15] demonstrated that the integration of automatic labelers and expert annotations using BERT-based approaches increases the accuracy of radiology report labeling. NegBio also offers a powerful approach for detecting negation and uncertainty in radiology reports, thus enabling more accurate extraction of clinically relevant findings from free-text reports [16]. These resources have become standard benchmarks for radiology report generation research.

Recent work has shown increasing focus on multimodal representation learning and explainability. Messina et al. reviewed advances of deep learning methods for automatic medical report generation, and stressed the increasing importance of explainability and clinically meaningful evaluation [17]. Huang et al. proposed GLORIA to improve image-text alignment through global and local multimodal representation learning [18]. Self-supervised learning has been shown to learn useful representations from large collections of unlabeled chest X-ray images and improve performance on downstream chest X-ray analysis tasks [19] (Tiu et al.)

However, there are few studies on how to translate research models into deployable software systems. Most of the work so far is centered on model design, training strategies, or benchmark performance. However, efficient model loading, retrieval database integration, interactive web interfaces, report visualization, document generation, and robust cloud-based inference are required for practical deployment. Tackling these challenges is critical for evaluating multimodal artificial intelligence systems in real-world deployment scenarios.

Motivated by these limitations, we propose a unified deployment pipeline named ClinCrossT5 combining Vision Transformer based visual feature extraction, similarity based retrieval, and Clinical-T5 report generation. Unlike previous work that is primarily concerned with model development, this work is concerned with implementation and operational evaluation of an end-to-end retrieval-augmented reporting system deployed using a browser-accessible interface provided by Gradio and Hugging Face Spaces. The resulting platform provides a practical environment for research and decision-support evaluation of retrieval-augmented multimodal radiology report generation systems.

Natural Resources for
Human Health
eISSN: 2583-1194

3. SYSTEM ARCHITECTURE

We propose ClinCrossT5, an end-to-end retrieval-augmented multimodal framework for structured radiology report generation from chest X-ray images and clinical indications. The architecture consists of four main modules: (i) visual features extraction, (ii) similarity-based case retrieval, (iii) multimodal report generation, and (iv) report presentation and export. The overall system architecture is shown in Fig. 1.

This process starts with uploading a frontal chest X-ray image via the web interface. The image is resized to 224 × 224 pixels and normalized by mean and standard deviation values of the ImageNet dataset before being passed through the pretrained Vision Transformer (ViT-B/16) encoder. The encoder outputs a 768-dimensional vector representation of the image's radiographic content.

For additional clinical context, the extracted feature vector is compared to a pre-computed feature database that is computed from the IU Chest X-ray dataset. Visually similar exams are identified using cosine similarity, and the corresponding Findings and Impression sections are extracted as reference context. The retrieved reports contain clinically relevant information that complements the visual representation of the input image.

The report generation module combines three sources of information: visual features, the clinical indication provided by the user, and the retrieved reference reports. A linear projection layer projects visual features into the textual embedding space and multi-head attention fuses visual and textual representations. The fused representation is then passed to the Clinical-T5 decoder for report generation.

Clinical-T5 produces a structured report with Findings and Impression sections. To improve report coherence and reduce redundant text, beam search decoding and repetition control mechanisms are incorporated.

The report produced is presented through the web interface, allowing the user to see the Findings and Impression separately. The system is able to export the generated report as a Microsoft Word (.docx) document for offline review and documentation purposes.

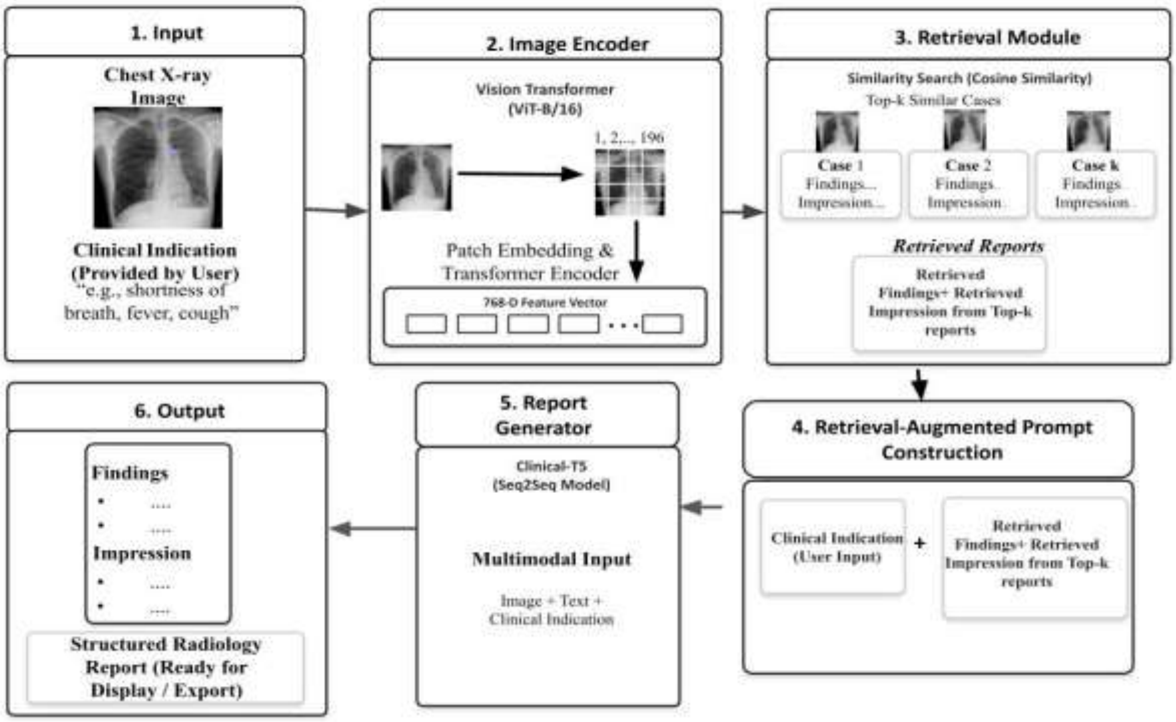


Figure 1. Overall architecture of the proposed ClinCrossT5 framework illustrating image preprocessing, ViT-B/16 feature extraction, similarity-based retrieval of reference cases, multimodal fusion with Clinical-T5, structured report generation, and document export.

4. DEPLOYMENT FRAMEWORK AND SOFTWARE IMPLEMENTATION

The ClinCrossT5 framework was implemented as a web-based application for automated chest X-ray report generation. The deployment integrates the pretrained multimodal model, similarity-based retrieval module, and report generation pipeline into a single system that supports end-to-end inference through a web browser. Figure 2 illustrates the deployment framework and the interaction between the major system components.

The interface was implemented on Gradio, a light-weight environment for image upload, input of the clinical indication, visualization of the report and export of the document. Gradio was chosen for its ease of use, compatibility with deep learning models and cloud deployment capabilities. The application runs on Hugging Face Spaces, which lets users access the system using a standard web browser, no local installation or software configuration required.

In the initialization process, the pretrained Clinical-T5 model, the tokenizer, the feature extractor of ViT-B/16 and the retrieval database are loaded into the memory. The retrieval database contains precomputed visual feature embeddings extracted from the IU Chest X-ray dataset together with their corresponding Findings and Impression reports. The inference time is reduced by pre-computing the reference embeddings since the feature extraction is not performed repeatedly at runtime.

The inference process is initiated when a user uploads a chest X-ray image together with the clinical indication associated with it. The image is preprocessed and passed through the pretrained ViT-B/16 encoder to obtain a 768-dimensional feature vector. Cosine similarity is then used to compare the extracted feature vector with the stored feature database and retrieve the most visually similar reference cases. The retrieved reports, together with the user-provided indication, are incorporated into a retrieval-augmented prompt that is passed to the Clinical-T5 model.

The generated report is displayed through the web interface as structured Findings and Impression sections. The system automatically formats the output and provides an option to export the report as a Microsoft Word (.docx) document, making it suitable for documentation and offline review.

The software follows a modular architecture in which image preprocessing, feature extraction, similarity retrieval, multimodal inference, report generation, and document creation are implemented as independent components. This design simplifies maintenance and allows individual modules to be updated or replaced without affecting the overall system. For example, alternative visual encoders, retrieval methods, or language models can be incorporated with minimal modifications.

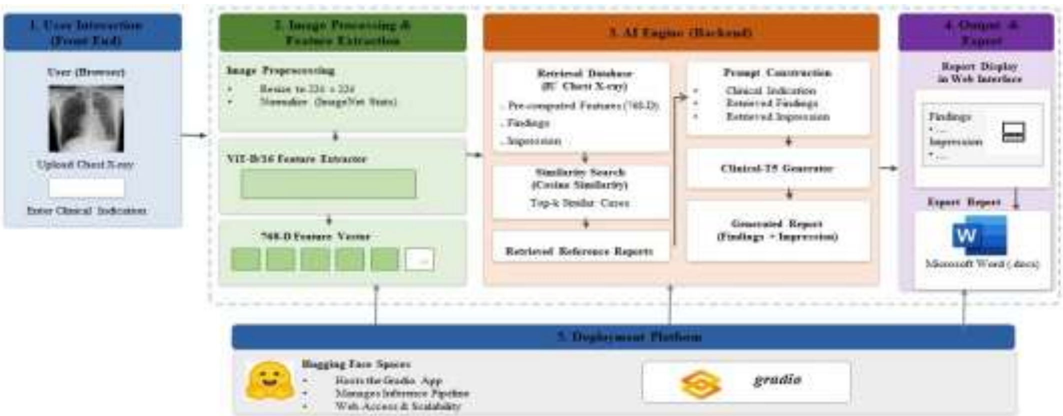


Figure 2. Deployment framework of the ClinCrossT5 application showing model initialization, image preprocessing, feature extraction, similarity-based retrieval, Clinical-T5 report generation, report visualization, and document export through the Gradio interface deployed on Hugging Face Spaces.

5. IMPLEMENTATION DETAILS AND EXPERIMENTAL SETUP

5.1 Model Components

The visual encoder is ViT-B/16, which extracts a 768 dimensional feature vector from each chest X-ray image. These feature vectors represent the radiologic content of the image and are used for similarity based retrieval and generation of the report.

Structured radiology reports are generated using the sequence-to-sequence language model Clinical-T5. The extracted visual features are projected into the textual embedding space by a learnable linear projection layer, and a multi-head attention module fuses the visual and textual representations before decoding. The model outputs two structured sections, Findings and Impression, in compliance with standard practice in radiology reporting.

5.2 Retrieval Database

The retrieval augmentation is performed using a database of precomputed image feature embeddings from the IU Chest X-ray dataset. For each entry in the database, we have a ViT-B/16 visual feature embedding along with the corresponding Findings and Impression section.

At inference time, cosine similarity is used to match the uploaded image to the stored feature embeddings and retrieve the most visually similar reference cases. The retrieved reports provide additional contextual information to enhance the visual representation of the input image before report generation.

The pre-computation of image embeddings avoids the repeated extraction of the features of the reference dataset at the time of inference. The computational efficiency is improved without degrading the retrieval quality.

5.3 Software Environment

The deployment framework was built in Python with popular deep learning and application development libraries. The primary software components include PyTorch, Hugging Face Transformers, Torchvision, Gradio, NumPy, Pandas, Pillow, and python-docx.

We implemented Clinical-T5 using Hugging Face Transformers and used PyTorch to load the pretrained models and do the inference. The web interface was created using Gradio . Downloadable Microsoft Word (.docx) reports were generated using python-docx.

5.4 Deployment Platform

The application is deployed on Hugging Face Spaces, a cloud-hosting platform for interactive machine learning applications. The software architecture is modular with separate components for preprocessing, retrieval, report generation and document creation. This design supports maintenance and allows for the update, replacement or extension of individual modules without affecting the overall application.

5.5 Inference Configuration

During inference, each uploaded chest X-ray image is resized to 224 × 224 pixels and normalized using the ImageNet mean and standard deviation values used in Vision Transformer (ViT-B/16) pretraining. The extracted image embedding is compared with the retrieval database by cosine similarity to find the most similar reference cases.

The input prompt of the Clinical-T5 model consists of the retrieved reports and the clinical indication provided by the user. We use beam search decoding with repetition control for report generation to improve the coherence and reduce the repetitive text. The generated sections of Findings and Impression are displayed on the web interface and can be downloaded as a Microsoft Word document (.docx).

The inference configuration is summarized in Table 1.

Table 1. Model and Inference Configuration

Parameter	**Configuration**
Visual encoder	Vision Transformer (ViT-B/16)
Image size	224 × 224 pixels
Visual embedding dimension	768
Language model	Clinical-T5

Natural Resources for Human Health
eISSN: 2583-1194

Retrieval method	Cosine similarity
Retrieval database	IU Chest X-ray dataset
Output sections	Findings and Impression
Deployment interface	Gradio
Cloud platform	Hugging Face Spaces
Export format	Microsoft Word (.docx)

5.6 Reproducibility

This modular framework combines preprocessing, feature extraction, retrieval, inference and report generation with pretrained models and open source libraries, facilitating reproducibility and future extension.

6. RESULTS

6.1 Deployment of the ClinCrossT5 Application

The proposed framework was deployed as an interactive web application using Gradio and hosted on Hugging Face Spaces. The interface allows users to upload chest X-ray images, enter clinical indications, generate structured radiology reports, and export the results as Microsoft Word (.docx) documents. Figure 3 shows the user interface of the deployed application.

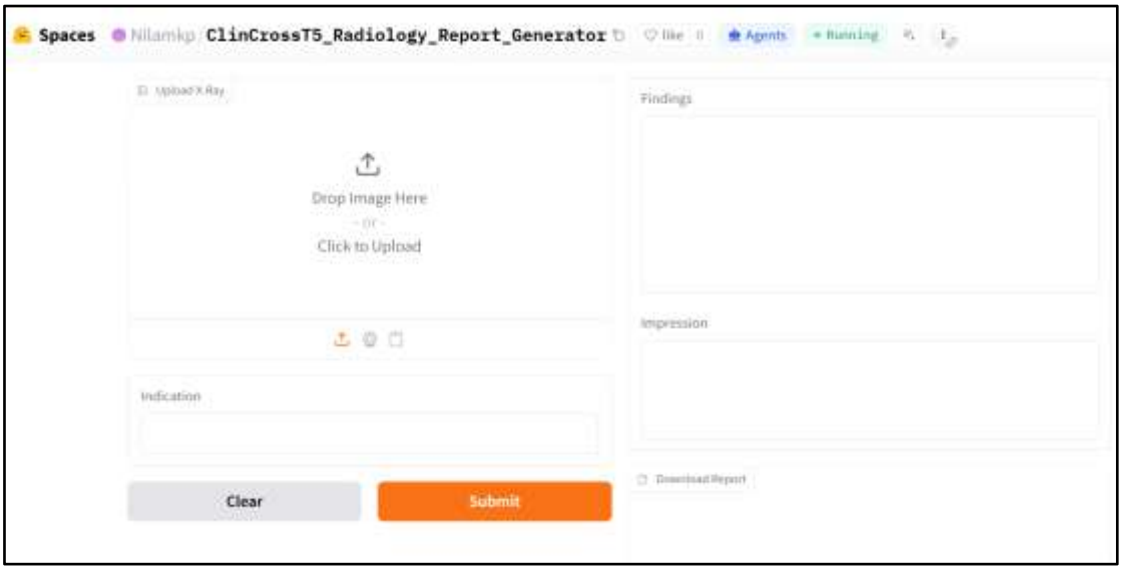


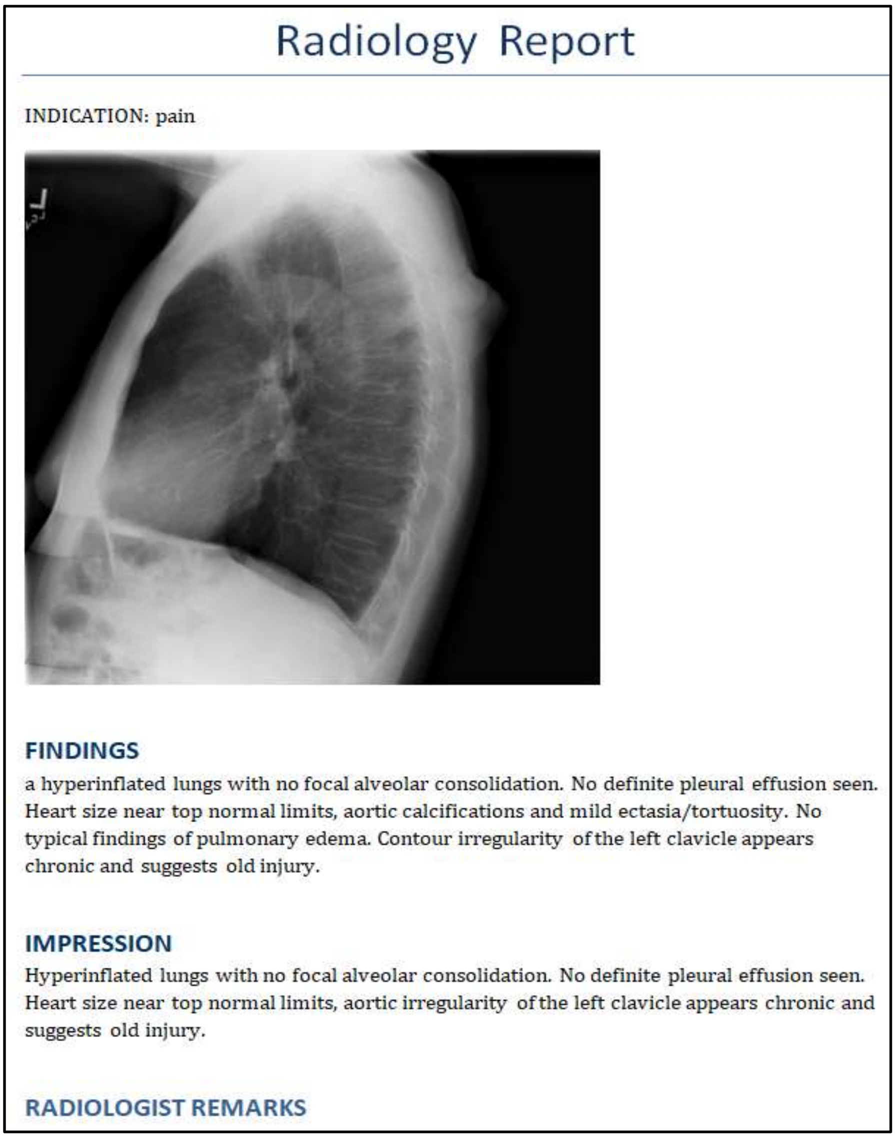
Figure 3. Browser-based Gradio interface of the deployed ClinCrossT5 application showing chest X-ray upload, clinical indication input, report generation, and document export.

6.2 End-to-End Report Generation

To evaluate the deployed system, chest X-ray images and their corresponding clinical indications were submitted through the web interface. For each input, the application successfully completed the full inference pipeline, including image preprocessing, feature extraction, similarity-based retrieval, multimodal inference, and structured report generation.

The generated reports contained two sections: Findings, describing the radiographic observations, and Impression, summarizing the main diagnostic interpretation. The generated reports followed the standard radiology reporting format and incorporated information from both the uploaded image and the retrieved reference cases. Figure 4 shows an example of a report generated by the deployed application.

Natural Resources for
Human Health
eISSN: 2583-1194



Radiology Report

INDICATION: pain

FINDINGS

a hyperinflated lungs with no focal alveolar consolidation. No definite pleural effusion seen. Heart size near top normal limits, aortic calcifications and mild ectasia/tortuosity. No typical findings of pulmonary edema. Contour irregularity of the left clavicle appears chronic and suggests old injury.

IMPRESSION

Hyperinflated lungs with no focal alveolar consolidation. No definite pleural effusion seen. Heart size near top normal limits, aortic irregularity of the left clavicle appears chronic and suggests old injury.

RADIOLOGIST REMARKS

Figure 4. Example of a structured radiology report generated by the deployed ClinCrossT5 application, illustrating the automatically generated Findings and Impression sections.

6.3 Interactive User Interface

The web interface allows users to interact with the deployed model using two inputs: a chest X-ray image and the corresponding clinical indication. After report generation, the application displays the Findings and Impression sections separately for easy review.

The application also supports export of the generated report as a Microsoft Word (.docx) document. This feature enables report documentation, sharing, and offline review without additional software or manual formatting.

The interface provides a simple and accessible workflow for generating and reviewing structured radiology reports, making the application suitable for research and demonstration purposes.

6.4 Qualitative Evaluation

A qualitative assessment was performed using representative chest X-ray images to examine the behavior of the deployed application. For the tested examples, the system successfully generated complete radiology reports in the standard Findings–Impression format.

The retrieval module incorporated information from visually similar reference cases, which provided additional context during report generation. The generated reports were well structured and showed consistent correspondence between the Findings and Impression sections.

This assessment focused on the operational behavior of the deployed application rather than diagnostic accuracy. The examples demonstrate that the integrated retrieval-augmented inference pipeline can generate structured radiology reports within the deployed web-based system.

6.5 Deployment Outcomes

The deployment of ClinCrossT5 demonstrated the following capabilities:

integration of image preprocessing, feature extraction, similarity-based retrieval, multimodal inference, and report generation within a single application;

i) browser-based access without local installation;

ii) automatic generation of structured Findings and Impression sections;

iii) interactive report visualization through the web interface;

iv) export of generated reports as Microsoft Word (.docx) documents; and

v) a modular software design that supports future extension of retrieval methods or language models.

6.6 Underlying model performance

ClinCrossT5, the deployed application, is based on a retrieval-augmented multimodal report generation model developed and evaluated separately. The underlying model was evaluated on the IU Chest X-ray dataset using natural language generation metrics. The present study focuses specifically on the software integration, deployment, and operational evaluation of the system; model development and experimental evaluation are beyond the scope of this paper.

Table 2. Performance of the ClinCrossT5 report generation model on IU X-ray dataset

Metric	ClinCrossT5
BLEU-1	0.544
BLEU-4	0.360
ROUGE-L	0.494

6.7 Deployment Latency Evaluation

To assess the responsiveness of the deployed ClinCrossT5 application, we measured the end-to-end inference latency on the production Hugging Face Spaces deployment with NVIDIA T4 GPU. Latency was measured as the wall-clock time between submitting an API request and receiving the generated Findings, Impression, and downloadable report using the gradio_client Python interface.

There were two experiments. To measure the cold-start and warm inference latency, we first submitted a single chest X-ray image for ten consecutive requests.Second, latency was measured on 20 chest X-ray images with file sizes from 19.3 KB to 435.5 KB with three repeated runs per image (60 requests total).

The cold-start latency is the first request that arrives after the application initialization or after a period of inactivity, which requires container and model initialization. One cold-start request in the current deployment took 12.38 s, which reflects this initialization overhead. Additional warm requests for the same image averaged 6.07 s (median=6.11 s, SD=0.19 s, P95=6.32 s), indicating stable performance after initialization.

The average latency for the 20-image benchmark was 8.98 s (median = 9.31 s, SD = 1.87 s, P95 = 11.62 s) and the average latency for individual images ranged from 5.54 s to 11.46 s, as summarized in Table 3. Image file sizes ranged from 19.3 KB to 435.5 KB. We did not find any consistent relationship between image file size and inference latency. These results suggest that latency is affected by more than just image file size, possibly including retrieval operations and autoregressive report generation.

Most of the requests were completed in 5–12 s after the initial warm-up period, which indicates the practical responsiveness of the deployed application in a web-based environment. The cold-start overhead of around 12 s may affect applications requiring constant low latency and could be improved by scheduled warm-up requests or a persistent hosting setup. Future work will investigate the relationship between latency and length of the generated report and evaluate performance for concurrent requests from multiple users.

Table 3. Deployment Latency Summary

Metric	Single-image (repeated, n=9)	Cross-image (20 images, n=60)
Hardware tier	NVIDIA T4 (GPU)	NVIDIA T4 (GPU)
Cold start (s)	12.38	—
Mean (s)	6.07	8.98
Median (s)	6.11	9.31
Min (s)	5.88	5.06
Max (s)	6.40	11.91
P95 (s)	6.32	11.62
SD (s)	0.19	1.87
Distinct images	1	20
n (requests)	9	60

The application was deployed on Hugging Face Spaces using NVIDIA T4 GPU runtime. A cold start is the first request after a period of inactivity, which involves initializing the container and loading the model. The cross-image experiment was performed after the application was initialized, so we do not measure cold-start latency for it. Single-image latency represents warm inference for one repeated image (n = 9). Cross-image latency represents inference across 20 chest X-ray images with three repeated runs per image (n = 60). Image file sizes ranged from 19.3 KB to 435.5 KB. SD denotes standard deviation, and P95 denotes the 95th percentile.

6. LIMITATIONS AND FUTURE DIRECTIONS

While the proposed ClinCrossT5 framework shows successful implementation of a retrieval-augmented multimodal radiology report generation system, there are several limitations to note.

First, the retrieval database is based on the IU Chest X-ray dataset and is limited in size and diversity compared to large clinical repositories. Availability of visually similar reference cases and the retrieval performance can be improved by a larger and more diverse retrieval database.

Second, the current system produces reports from a single chest X-ray image and its associated clinical indication.It excludes multi-view examinations, prior imaging, laboratory results, and other clinical information that may affect radiological interpretation. Incorporating additional patient data could improve report completeness and clinical relevance.

Third, the study assesses the performance of the implemented application in operation, not its clinical impact. Although the system reports successful end-to-end report generation, radiologist reader studies and prospective clinical validation were beyond the scope of this work.

The current implementation is also stand alone web application and not integrated with hospital information systems, electronic health records or PACS.This type of integration would require interoperability with clinical infrastructure, secure authentication, privacy protection and compliance with institutional and regulatory requirements .

Future work will focus on expanding the retrieval database, investigating more advanced retrieval methods, and evaluating recent medical vision-language foundation models. Other enhancements might be scalable cloud deployment, asynchronous inference, monitoring tools and containerized services. Further studies may also explore multilingual report generation and additional imaging modalities. To summarize, ClinCrossT5 is a useful platform for research and evaluation of decision support for retrieval-augmented radiology report generation . Overcoming these limitations will lead to more robust and scalable deployment frameworks.

7. CONCLUSION

In this paper, we introduce ClinCrossT5, a web-based application implementation of a retrieval-augmented multimodal framework for automated chest X-ray report generation. The system extracts features from chest X-ray images and clinical indications using Vision Transformer (ViT-B/16), retrieves reference cases based on similarity, and generates structured Findings and Impression sections using Clinical-T5 language generation.

The main contribution of this work is the practical deployment of a report generation system with retrieval augmentation. The application was deployed successfully using Gradio and Hugging Face Spaces, integrating image preprocessing, feature extraction, similarity-based retrieval, multimodal inference, report generation, and document export into a single web-based interface. The modular design also allows for future improvements in the visual encoder, retrieval module or language model.

The deployment results show that the system is able to generate structured radiology reports successfully and provide an interactive interface for report review and export. We show how to practically implement such an end-to-end retrieval-augmented inference pipeline in a web-based setting.

There are a number of directions for future work. These are improving the retrieval database, testing more advanced medical vision-language models, adding more clinical information, and improving the interoperability with clinical information systems.Studies with radiologists as users will also be crucial to evaluate the system's usability and practical value.

Overall, ClinCrossT5 offers a valuable platform for research and decision-support evaluation of retrieval-augmented radiology report generation. The deployed application shows how a multimodal report generation model can be converted to a web-based accessible system for future research and software development in computer-assisted radiology.

ACKNOWLEDGEMENTS

The authors would like to thank MET IOE, Nashik and SNJB's KBJ COE, Chandwad for providing the computational resources and support necessary for this work. The authors also acknowledge the creators and maintainers of the IU Chest X-ray dataset and the open-source libraries (PyTorch, Hugging Face Transformers, Gradio) that made this deployment possible.

AUTHOR CONTRIBUTIONS

Nilam Sureshrao Khairnar: Conceptualization, Investigation, Methodology, Data Curation, Formal Analysis.

Dr. Shirish S. Sanet: Supervision, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study utilized the publicly available, de-identified Indiana University (IU) Chest X-ray dataset for retrieval database construction and qualitative evaluation. No new patient data were collected, and no direct patient contact or intervention was involved in this work. As the study relied solely on a pre-existing, de-identified, publicly available dataset, formal ethics committee/IRB approval was not required. The deployed application is intended for research and demonstration purposes only and is not validated for direct clinical use or diagnostic decision-making.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Jing, B., Xie, P., & Xing, E. (2018, July). On the automatic generation of medical imaging reports. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers) (pp. 2577-2586).

[2] Li, Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems, 31.

[3] Chen, Z., Song, Y., Chang, T. H., & Wan, X. (2020, November). Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 1439-1449).

[4] Delbrouck, J. B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., & Langlotz, C. (2022, December). Improving the factual correctness of radiology report generation with semantic rewards. In Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. 4348-4360).

[5] Zhou, H. Y., Chen, X., Zhang, Y., Luo, R., Wang, L., & Yu, Y. (2022). Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nature Machine Intelligence, 4(1), 32-40.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[8] Dosovitskiy, A. (2021). An image is worth 16x16 words: Transformers for image recognition at scale; Prof. Int. Conf. on Learning Representations.

[9] Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., ... & Oktay, O. (2022, October). Making the most of text semantics to improve biomedical vision–language processing. In European conference on computer vision (pp. 1-21). Cham: Springer Nature Switzerland.

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[11] Lehman, E., & Johnson, A. (2023). Clinical-t5: Large language models built using mimic clinical text. PhysioNet, 101, 215-220.

[12] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

[13] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).

[14] Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1), 317.

[15] Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., & Lungren, M. (2020, November). Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 1500-1519).

[16] Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., & Lu, Z. (2018). NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings, 2018, 188.

[17] Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., ... & Capurro, D. (2022). A survey on deep learning and explainability for automatic report generation from medical images. ACM Computing Surveys (CSUR), 54(10s), 1-40.

[18] Huang, S. C., Shen, L., Lungren, M. P., & Yeung, S. (2021). 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 (pp. 3942-3951).

[19] Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., & Rajpurkar, P. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature biomedical engineering, 6(12), 1399-1406.