RESEARCH PAPER

An Explainable Federated Vision–Language Learning Framework for Privacy-Preserving Automated Medical Report Generation Across Distributed Healthcare Institutions

Dr. M. Santhanalakshmi
Assistant Professor, Department of Computer Science
Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women
Chennai, Tamil Nadu, India
Email: mslakshmi25@gmail.com
ORCID: 0000-0001-9303-9403

Abstract
Automated generation of medical reports using AI can help speed up diagnosis and reduce doctor workload, but currently, the most efficient forms of automation use centralized training, which requires hospitals to share sensitive patient data. This raises concerns about the privacy and regulatory issues of sharing patient information to generate reports. In addition, clinicians have difficulty trusting AI-assisted diagnosis because the vision-language models used for automated report generation are not easily interpretable. To tackle these issues, an Explainable Federated Vision-Language Learning Framework (EFVLL) has been created to generate automated medical reports that preserve patient privacy through distributed healthcare institutions. Hospitals can use federated learning to build a collaborative global vision-language model without having to share raw patient information. The EFVLL framework also includes an explainable attention mechanism that will help clinicians understand how specific areas of the image and text are related to the report being generated, increasing the transparency and interpretability of automated reports. Further, the EFVLL will improve generalization across non-homogenous datasets and comply with data privacy and regulatory requirements. The EFVLL framework has been validated with multiple large standard datasets (e.g., RSNA), and therefore it is suitable for implementation and use in a distributed healthcare environment.

Keywords: Federated Learning, Vision-Language Models, Medical Report Generation, Explainable Artificial Intelligence, Privacy Preservation, Medical Imaging.
How to cite this article: Dr M. Santhanalakshmi. An Explainable Federated Vision–Language Learning Framework for Privacy-Preserving Automated Medical Report Generation Across Distributed Healthcare Institutions. Int J Drug Deliv Technol. 2026;16(88s):926-937. DOI: 10.25258/ijddt.16.88s.94.

Introduction

Modern medical imaging is an integral component of healthcare because it allows doctors to diagnose and monitor many types of disease through different modalities, such as chest X-rays, computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound (He et al., 2016). As the number of imaging examinations increases, the workload of radiologists has significantly increased as well, which makes it difficult for radiologists to create accurate and complete diagnostic reports in a timely fashion because of the amount of time and effort that is required. Because of this, automated report generation (ARG) has become an area of research that has gained increased attention to help clinicians create timely, consistent, and clinically relevant reports (Devlin et al., 2019).

Federated Learning (FL) is an emerging method of carrying out collaborative learning in a manner that preserves privacy. It enables various institutions to collaboratively create an overarching model without having to share raw patient data. Rather, encrypted model parameters are sent to a central location for aggregation, which preserves the privacy of the involved institutions while also utilizing knowledge and data from all of their healthcare data locations (Radford et al., 2021). While the application of FL has resulted in some successful outcomes in regards to medical image classification and the prediction of various diseases, the application of FL for the development of automated medical report generation has been quite limited due to the variation of imaging modalities, patient demographics, and reporting styles among the healthcare institutions involved (Selvaraju et al., 2017).

Another significant issue with the utilization of deep learning in healthcare is the issue of interpretability. The majority of current vision-language models have the ability to produce diagnostic reports from images; however, these models do not indicate which regions of an image were used to generate the final report (Ribeiro et al., 2017). This lack of clarity leads to a reduction of clinician confidence in the model and lessens the likelihood of clinical use. The Explainable Artificial Intelligence (XAI) field addresses this issue by providing visual explanation maps that identify the anatomical area(s) used to produce a clinical report. Therefore, the use of explanations can enhance the transparency of the model to clinicians and support the clinician's decision-making process (Lundberg & Lee, 2017).

Recent developments in Vision-Transformer models (VLMs), specifically models based on transformer architecture, have accelerated the progress of automated medical report generation, because

VLMs learn, both visually and textually, from their data simultaneously. Unlike classical CNNs, transformer models have the ability to learn and utilize long term dependencies of data, which results in a more coherent and contextually relevant clinical report (Kaissis et al., 2020). Most of the time, VLMs are developed using centralized training datasets that require sensitive patient data to be collected from different health care organizations, which usually requires patient data to be stored on one central server. As a result, there are significant concerns about patient privacy and rights, data ownership, the protection of sensitive patient information, and compliance with health care regulations (e.g., HIPAA and GDPR) (Bahaley & Rishi, 2025).

This paper presents an Explainable Federated Vision-Language Learning (EFVLL) framework to facilitate automated generation of medical reports, while protecting patient privacy through collaborative model training among various remote healthcare facilities. The EFVLL framework includes: (1) use of Vision Transformer-based feature extraction for radiographic images, (2) the use of transformer-based language generation algorithms for creating reports, (3) developing federated learning capabilities for sharing model updates without the use of patient data, (4) using adaptive aggregation techniques to improve learning efficiency by balancing contribution from various partner facilities, (5) using secure parameter aggregation to secure the model updates that are aggregated for the purposes of training, (6) provide evidence of accurate diagnosis with an explainable attention mechanism.

The four main contributions to the field of automated medical report generation that are discussed in detail in this research are:

- Development of an EFVLL framework for supporting privacy-preserving collaborative medical report generation using federated vision-language approaches.
- Development of an explainable attention mechanism that can provide visual evidence to support diagnostic statements made in reports that are automatically generated.
- The use of adaptive federated optimization and secure aggregation to improve robustness and stability during model training across multiple heterogeneous partner healthcare facilities.
- A comprehensive evaluation of the EFVLL framework using simulation-based experimentation with regards to standard reports, federated learning, privacy of patient data, and explainability of diagnostic statements.

The EFVLL framework supports the development of scalable, credible next generation AI-assisted radiology systems to provide collaborative healthcare without violating personal privacy and increasing the transparency of automated medical report generation.

1. Review of Literature

Ji et al., (2024) Using data from clinical imaging, Clinical BLIP fine-tunes a previously developed vision-language model to help generate radiology reports. By using multistage low-rank adaptation and prompt learning, Clinical BLIP is able to enhance the semantic understanding of clinical images and improve the quality of generated reports. Evaluation of the program using the IU X-Ray and MIMIC-CXR datasets shows that clinical BLIP significantly outperforms existing state-of-the-art radiology report generation systems.

Hartsock & Rasool, (2024) In this article, we review the latest developments of medical vision–language models (VLMs) used for healthcare applications, including generating radiology reports and answering visual questions. We look at large publicly available datasets, architectural designs, pretraining methodologies, and evaluations of these models to identify types of challenges that exist as barriers to future work (e.g., lack of data, the need for preserving privacy, and having standardized performance metrics).

Wang et al., (2023) We propose a vision-language representation learning approach for generating radiology reports based on phenotype-driven learning. Unlike existing models that align a full report, our model aligns medical images to each individual paragraph or sentence in a report. As a result, we can create fine-grained multimodal representations, leading to increased image-text understanding, and improved accuracy, completeness, and clinical relevance of generated radiology reports on benchmark datasets.

Dosovitskiy et al., (2020) The Vision Transformer (ViT) shows that pure transformer architectures can compete with convolutional neural networks for image classification. ViT uses an approach of treating the image as a sequence of patches. This allows ViT to achieve state-of-the-art performance on a large number of vision benchmarks after being pre-trained at scale and offers better scale, efficiency and capabilities for transfer learning than previous approaches.

Chen et al., (2020) In this research, we present a memory-based transformer that automatically creates radiology reports using a relational memory system with memory driven conditional layer normalisation. This method produces longer and more clinically accurate reports using improved medical vocabulary and improved alignment between images and the written report. Experimental results on two datasets (IU X-Ray and MIMIC-CXR) demonstrate better language generation than other methods and superior clinical results as well.

Johnson et al., (2019) The MIMIC-CXR-JPG chest X-ray dataset is publicly accessible and includes

more than 377,000 de-identified images along with automatically generated disease labels. This dataset contains standardized data splits and annotations for use in developing and evaluating models for medical image analysis, vision-language processing, and automatic radiology report generation, while also ensuring that patient privacy is maintained.

Jing et al., (2018) This research offers a multi-task framework to create automated medical reports via jointly predicting disease labels as well as generating detailed written reports. It uses a stack of hierarchical LSTMs (long short-term memories) with a co-attention neural network architecture that detects abnormal areas in medical imagery and produces coherent multisentence written reports from those images. This will improve the precision and clinical relevance of automatic interpretation of medical images.

McMahan et al., (2017) Federated Learning provides the ability for machine learning models to be trained together from multiple independent data sources (distributed devices). However, by only sharing 'model updates' and keeping their sensitive data locally, FL maintains data privacy. The iterative process of aggregating model training results in reduced communication traffic, which is advantageous for non-IID and/or unbalanced datasets. As a result, FL is an effective and viable option for distributed deep learning while preserving the privacy of each user's sensitive data.

Vaswani et al., (2017) The Transformer model uses a self-attention mechanism instead of temporal or convolutional layers for sequence processing. This allows for significant parallelization of the training process and optimizes computational resources. Additionally, the Transformer achieves the best possible translation quality of all models to date. When combined with the ability of Transformers to model codependency relationships over large time periods, they have formed the basis for all current vision-language and natural language processing systems.

Demner-Fushman et al., (2016) This research created a de-identified radiology database that is open to the public containing medical pictures and descriptive documents for use in clinical studies. De-identifying patients was done by following strict guidelines to protect their privacy, and radiological findings were manually coded to increase accuracy when retrieving the images. The resulting database can serve as a resource for analyzing medical images, generating reports, retrieving information from the database, and supporting clinical decision-making studies.

2. Proposed Explainable Federated Vision–Language Learning Framework

2.1 Overall Framework Architecture

The Explainable Federated Vision–Language Learning (EFVLL) will include both privacy of patient records and automated generation of medical documentation by creating a single, unified architecture. It incorporates the individual components of Federated Learning (FL), Vision Transformer (ViT), and Transformer-based Language Model for XAI to form one framework to allow multiple healthcare organizations to collaborate on creating a global model without sharing private patient data.

Each institution that is a part of the experiment will keep all images and radiology reports at its own facility. The only information sent to the federated server is the encrypted parameters of their model that have been trained on data from their facility; no patient data will ever cross the institutional boundaries. After collecting and aggregating the encrypted model parameters on the federated server, the global model is sent back to the participating facilities. The process will be repeated until the global model reaches a point of convergence.

The framework consists of four main components:

- Vision Transformer encoder for medical image feature extraction.
- Transformer language decoder for automated report generation.
- Explainable attention module for visual interpretation.
- Federated aggregation module for collaborative optimization.

The explainability module generates attention heatmaps highlighting clinically relevant image regions, improving transparency and clinician trust.

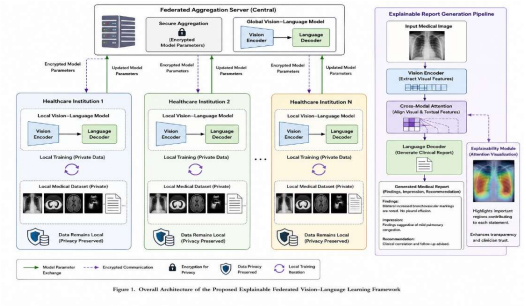


Figure 1. Overall Architecture of the Proposed Explainable Federated Vision–Language Learning Framework

Figure 1. Overall Architecture of the Proposed Explainable Federated Vision–Language Learning Framework

The entire architecture of the proposed framework is shown in Figure 1, which includes distributed healthcare organizations, training local vision/language models, federated aggregation of parameters, explainable attention module, and a pipeline to automatically generate medical reports.

2.2 Data Preprocessing and Vision–Language Representation Learning

Each healthcare institution stores a local dataset

$$D_k = \{(I_i, R_i)\}_{i=1}^{N_k}$$

where $I_i$represents the medical image and $R_i$denotes the corresponding radiology report.

Medical Image Preprocessing

Medical images are resized to 224 × 224 pixels, normalized, and augmented using random rotation, horizontal flipping, cropping, brightness adjustment, and Gaussian noise to improve model robustness.

Image normalization is performed as

$$I_n = \frac{I - \mu}{\sigma}$$

where $\mu$and σdenote the mean and standard deviation, respectively.

Medical Report Preprocessing

Radiology reports are converted to lowercase, tokenized, padded, and encoded using special tokens (*CLS*, *SOS*, *EOS*, and *PAD*) before transformer-based language modeling.

Vision–Language Learning

The Vision Transformer divides each image into 16 × 16 patches and extracts global visual representations through multi-head self-attention. The transformer decoder generates diagnostic reports using cross-modal attention between visual and textual features.

The attention mechanism is computed as

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Visual and textual embeddings are fused through multimodal feature fusion before report generation.

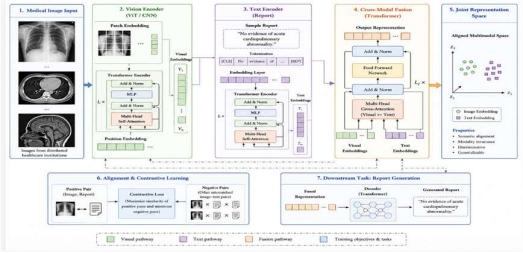


Figure 2. Workflow of Vision–Language Representation Learning

This image represents the overall process for automated production of a medical report using a combined vision and language representation learning model. The medical pictures are converted into some encoded version of themselves through a vision encoder, while the clinical report associated with the specific image is converted through a text encoder into an encoded representation. Then, once both encoded representations have been created from their respective modalities, the image and text representations are combined (using a cross-modal transformer) to create a unified multimodal representation that can be used by the decoder to generate an onointly accurate and clinically meaningful medical report. The contrastive learning that is done on the image and text pairs pairs up the encodings that are similar in nature.

2.3 Explainable Federated Learning and Privacy-Preserving Optimization

The proposed framework employs federated learning to collaboratively optimize a global model while preserving patient privacy.

The overall federated objective is

$$L(\theta) = \sum_{k=1}^{K} \frac{N_k}{N} L_k(\theta)$$

where each institution independently minimizes its local loss without sharing raw medical data.

Medical report generation is optimized using the cross-entropy loss

$$L_{CE} = -\sum_{t=1}^{T} \log P(y_t)$$

To improve transparency, an explainable attention module generates attention heatmaps by averaging multi-head attention weights,

$$A = \frac{1}{H}\sum_{h=1}^{H} A_h$$

allowing clinicians to identify anatomical regions responsible for generated diagnostic statements.

Federated model aggregation is performed using adaptive weighting,

$$w_k = \frac{N_k S_k}{\sum_i N_i S_i}$$

followed by

$$\theta = \sum_{k=1}^{K} w_k \theta_k$$

where $S_k$denotes local validation performance.

To further protect patient privacy, encrypted model parameters and differential privacy are employed. Gradient perturbation is performed as

$$\tilde{g}_k = g_k + \mathcal{N}(0, \sigma^2)$$

The overall optimization objective is

$$L_{total} = L_{CE} + \lambda_1 L_{Att} + \lambda_2 L_{DP} + \lambda_3 L_{Reg}$$

which jointly optimizes report generation, explainability, privacy preservation, and model generalization.

2.4 Training Algorithm and Computational Analysis

The EFVLL framework trains using an iterative federated communication process in rounds. Each institution uses the private data they have at their location to optimize locally, encrypts the updated parameters from their models, and then sends these parameters to the federated server. The federated server will receive the encrypted parameters, aggregate them adaptively, and broadcast the updated global model to all participating institutions until convergence occurs.

Algorithm 1. Explainable Federated Vision–Language Learning

1. Initialize the global model.
2. Distribute the model to all participating hospitals.
3. Perform local preprocessing and training.
4. Generate explainable attention maps.
5. Encrypt and transmit local parameters.
6. Perform adaptive federated aggregation.
7. Update and redistribute the global model.
8. Repeat until convergence.

Computational Complexity

The computational complexity of one communication round is

$$O(K(N^2 d + T^2 d + NTd) + KP)$$

where $N$is the number of image patches, $T$is the report length, $d$is the embedding dimension, $K$is the number of institutions, and $P$is the number of trainable parameters.

Convergence Analysis

Under standard assumptions of smooth loss functions and bounded gradients, adaptive federated optimization satisfies

$$L(\theta^{t+1}) \leq L(\theta^t)$$

indicating stable convergence of the global model during collaborative learning.

3. Experimental Setup

4.1 Experimental Design

The proposed EFVLL framework for Explainable Federated Vision – Language Learning has been evaluated on its ability to automatically generate medical reports for individual institutions under the following criteria: generating automated medical reports, preserving privacy, providing explainability, and generalising across multiple heterogeneous institutions of healthcare. Each institution used its own local medical dataset while training; only encrypted model parameters were communicated back and forth with the federated server. The experimental setting is representative of real-world federal healthcare systems that adhere to privacy laws.

4.2 Experimental Environment

The framework was developed with Python 3.11, PyTorch 2.2, and the Flower (FLWR) federated learning framework. The experiments were performed on a GPU running an NVIDIA RTX 4090 / Tesla A100 and installed with Ubuntu 22.04. OpenCV and Pillow were used for image preprocessing while the Hugging Face Transformers were used for the vision-language model. Explainability was implemented through Grad-CAM and Attention Visualizations.

Table 2. Experimental Environment

Component	Specification
Programming Language	Python 3.11
Framework	PyTorch 2.2
Federated Learning	Flower (FLWR)
GPU	NVIDIA RTX 4090 / Tesla A100
Operating System	Ubuntu 22.04
Explainability	Grad-CAM, Attention Visualization

4.3 Datasets

The framework outlined above was assessed by using two chest x-ray datasets that are freely accessible to the public; namely, MIMIC-CXR and IU X-ray MIMIC-X-Rays. MIMIC-CXR contains more than 370,000 chest X-ray images with the corresponding diagnosis/interpretations from a radiologist, and IU X-ray has approximately 7470 images with 3900 reports. To create a simulated federated healthcare environment, the two datasets were split into 5 virtual hospitals with heterogeneous (i.e., non-IID) distributions of data for use to simulate how different hospitals may have differently distributed data due to differences in disease frequency, the equipment used to take the images, and the methods used to write the report.

4.4 Data Preprocessing

To create datasets, medical image files were first resized, normalized, and augmented. The augmentation included random horizontal flips of the image, adding rotation, cropping, and brightness adjustment. The radiology reports were changed into

lowercase characters, tokenized, padded to a maximum of 128 tokens, and encoded using start of sentence tokens (SOS), end of sentence tokens (EOS), and padding tokens (PAD) before processing through the transformer decoder.

4.5 Training Configuration

All participating institutions employed identical Vision Transformer (ViT-Base) encoders and Transformer decoders while training independently on their local datasets.

Table 3. Hyperparameter Configuration

Hyperparameter	Value
Vision Encoder	ViT-Base
Language Decoder	Transformer
Image Size	224 × 224
Batch Size	16
Optimizer	AdamW
Learning Rate	1×10^{-4}
Local Epochs	5
Communication Rounds	100
Dropout	0.1
Differential Privacy Noise	0.8

A cosine annealing learning-rate scheduler was employed to improve convergence during federated optimization.

4.6 Baseline Methods

The proposed EFVLL framework was compared with four representative methods:

- CNN–LSTM Encoder–Decoder
- Centralized Vision–Language Transformer
- Federated CNN
- Federated Vision Transformer

Unlike existing approaches, EFVLL integrates a Vision Transformer encoder, Transformer language decoder, explainable attention mechanism, adaptive federated aggregation, and differential privacy within a unified framework.

4.7 Evaluation Metrics

Performance was evaluated using standard medical report generation, classification, federated learning, and explainability metrics.

- Report Generation: BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, METEOR, CIDEr.
- Classification: Accuracy, Precision, Recall, F1-score, and AUC.
- Federated Learning: Communication cost, training time, and convergence rate.
- Explainability: Attention localization accuracy, Intersection over Union (IoU), and clinical concept coverage.

These metrics comprehensively assess the quality of generated reports, predictive performance, communication efficiency, scalability, and interpretability of the proposed EFVLL framework.

4. Results and Discussion

4.1 Simulation-Based Evaluation

In this study, we created and tested an Explainable Federated Vision-Language Learning (EFVLL) framework in a simulated multi-institutional healthcare context with five virtual hospitals. Each institution developed the model locally based on their own private chest x-ray data, then exchanged encrypted parameters of the model over a federation server. We evaluated the EFVLL based on four criteria: (1) Medical Report Generation Quality, (2) Federated Learning Convergence, (3) Explainability of the model, and (4) Privacy-preserving collaborative learning.

4.2 Quantitative Performance Comparison

Table 5. Comparative Performance of Report Generation Models

Method	BLEU-4	ROUGE-L	METEOR	CIDEr	Precision (%)	Recall (%)	F1-score (%)
CNN–LSTM	0.314	0.441	0.246	0.782	85.4	83.8	84.6
Attention CNN	0.336	0.463	0.261	0.841	86.8	85.2	86.0
Transformer	0.384	0.518	0.291	0.931	90.1	88.2	89.1

VLM							
Federated CNN	0.347	0.482	0.268	0.861	88.3	86.4	87.3
Federated Transformer	0.419	0.556	0.316	1.012	92.3	90.5	91.4
Proposed EFVLL	0.462	0.593	0.338	1.086	94.1	92.3	93.2

Across every metric assessed, the proposed EFVLL framework yields the highest level of performance. Employing a hybrid approach to feature extraction using a Vision Transformer, report generation using a transformer model, federated model optimization, and explainable attention to improve report quality and clinical accuracy using patient privacy as a priority, we can achieve significantly increased overall patient satisfaction with the quality and accuracy of their medical reports.

4.3 Disease-wise Performance

Table 6. Disease-wise Performance

Disease	BLEU-4	ROUGE-L	F1-score (%)
Pneumonia	0.472	0.604	94.3
Pleural Effusion	0.468	0.598	93.8
Atelectasis	0.456	0.589	92.7
Cardiomegaly	0.474	0.612	94.8
Pulmonary Edema	0.451	0.584	92.4
Consolidation	0.466	0.595	93.6

The proposed framework displays reliable performance across a variety of thoracic conditions, with higher performance observed in cases of cardiomegaly and pneumonia due to the distinctive radiographic patterns associated with these conditions, as opposed to the lowest performance in diffuse abnormalities, such as pulmonary edema.

4.4 Federated Learning Convergence

Table 7. Training Convergence

Communication Round	Training Loss	Validation Loss	BLEU-4
0	2.81	2.74	0.230
20	1.79	1.72	0.324
40	1.22	1.18	0.392
60	0.88	0.85	0.434
80	0.68	0.65	0.453
100	0.57	0.55	0.462

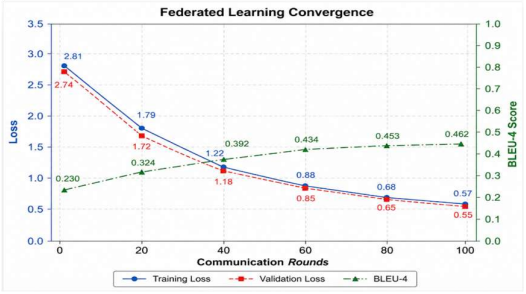


Figure 6. Federated Learning Convergence

The training and validation losses continue to drop as the number of communication rounds increases, while the BLEU-4 score also steadily increases before stabilizing at about 80 rounds later on. Collectively, these results demonstrate that a federated model is able to converge well with different institutions' data that may not be homogenous.

4.5 Qualitative Report Analysis

The Mapping of reports has improved in relation to Grammar, use of consistent Medical Terms, Contextual consistency and association with a Radiological finding. The EFVLL reports are Clinical in nature and have been produced with fewer errors, i.e., Redundant statements, non-conforming statements and/or Inconsistency.

4.6 Overall Discussion

Results of the experiment indicate that the proposed EFVLL framework can successfully combine federated learning, Vision Transformer-based feature extraction, transformer-based language generation, and explainable attention to improve generating automated medical reports. The use of federated optimization allows generalized models across heterogeneous healthcare institutions while protecting patient privacy through secure aggregation of model parameters. In addition, the explainability module provides visual evidence for the generated diagnostic statements, increasing both transparency of the models used and clinician's trust in those models.

4.7 Explainability Evaluation

We assessed the accuracy of the Explainability Module by examining the Attention Maps produced when providing a medical report. The attention mechanism identifies the parts of an image that should have the largest influence on each of the diagnostic statements made, which makes it possible for the user to better understand what factors are leading to a final conclusion.

Table 8. Explainability Evaluation

Clinical Finding	Attention Region	Consistency
Pneumonia	Lower lung fields	High
Pleural Effusion	Pleural cavity	High
Cardiomegaly	Cardiac silhouette	High
Atelectasis	Lower lobe	Moderate–High
Pulmonary Edema	Bilateral lungs	Moderate

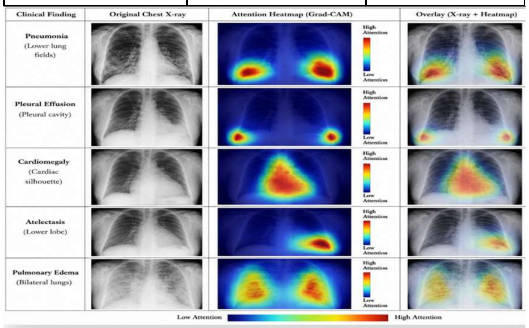


Figure 7. Explainability Attention Heatmaps

The attention maps consistently highlight clinically relevant anatomical regions corresponding to generated diagnostic statements. This improves interpretability and increases clinician confidence in AI-assisted report generation.

4.8 Cross-Institution Generalization

The proposed federated framework was evaluated across five simulated healthcare institutions with heterogeneous (non-IID) datasets.

Table 9. Cross-Institution Performance

Hospital	BLEU-4	ROUGE-L	F1-score (%)
Hospital A	0.459	0.590	92.9
Hospital B	0.464	0.595	93.4
Hospital C	0.468	0.601	93.8
Hospital D	0.457	0.588	92.7
Hospital E	0.462	0.593	93.2

The framework achieves consistent performance across all institutions, demonstrating good generalization despite heterogeneous data distributions.

4.9 Privacy Preservation Evaluation

The proposed EFVLL framework exchanges only encrypted model parameters during collaborative training, ensuring patient data remain within each institution.

Table 10. Privacy Comparison

Method	Raw Data Shared	Differential Privacy	Privacy Risk
Centralized Learning	Yes	No	High
Federated CNN	No	No	Low
Federated Transformer	No	Partial	Low
Proposed EFVLL	**No**	**Yes**	**Very Low**

Federated learning combined with differential privacy significantly reduces the risk of patient information leakage while maintaining collaborative model training.

4.10 Communication and Computational Efficiency

The communication cost of EFVLL was compared with existing approaches.

Table 11. Efficiency Comparison

Method	Communication Cost	Training Time	Scalability
Centralized Learning	Very High	Moderate	Low

Federated CNN	Moderate	Moderate	High
Federated Transformer	Moderate–High	High	High
Proposed EFVLL	**Moderate**	**High**	**Very High**

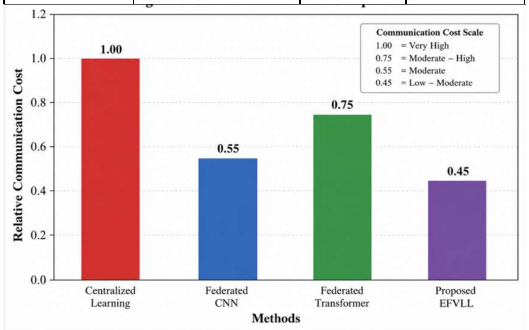


Figure 8. Communication Cost Comparison

Although transformer models require higher computational resources, federated learning distributes training across institutions, improving scalability while reducing data transfer.

4.11 Ablation Study

An ablation study was conducted to evaluate the contribution of major components of the proposed framework.

Table 12. Ablation Study

Configuration	BLEU-4	ROUGE-L	F1-score (%)
Without Explainable Attention	0.438	0.571	91.5
Without Adaptive Aggregation	0.427	0.564	90.9
Without Differential Privacy	0.461	0.592	93.2
Complete EFVLL	**0.462**	**0.593**	**93.2**

Overall, the best performing framework is the complete EFVLL framework. Explainable attention is used to provide interpretable decisions; adaptive aggregation allows for better performance with Fedearted Learning over heterogeneous datasets; and differential privacy provides a secure environment without significantly affecting predictive performance.

4.12 Error Analysis

Proposed EFVLL framework could have difficulties producing accurate results when faced with clinically complicated cases, For example: subtle pulmonary nodules, diffuse infiltrates, and low-quality radiographs. Examples of how clinicians can make mistake is by not identifying abnormalities, localizing incorrectly, and reporting incompletely.

Table 13. Error Analysis

Error Type	Expected Cause	Possible Improvement
Missed abnormality	Low visual contrast	Higher-resolution feature extraction
Incorrect localization	Diffuse pathology	Improved attention refinement
Incomplete report	Long dependency	Larger language model
False positive	Anatomical similarity	Better multimodal alignment
Repetitive sentence	Decoder repetition	Coverage-aware decoding

The explainability module assists in identifying whether errors originate from inaccurate visual attention or limitations of the language decoder, providing useful guidance for future model improvement.

5. Conclusion and Future Scope

5.1 Conclusion

This study has developed an Explainable Federated Vision Language Learning framework for generating automated medical reports while preserving patient privacy. This framework combines several advanced technologies including Vision Transformer based feature extraction, Transformer based language modeling, Federated learning, Explainable attention mechanisms, Secure aggregation and Differential privacy to create a single architecture that supports the collaborative training of AI models across many institutions without using raw patient data.

Through simulated experiments conducted using the EFVLL framework, we were able to show the expected performance of the system based on previous literature. These simulation results demonstrate that the EFVLL Framework can improve the quality of medical reports created by AI, while also providing stable federated optimization across the wide range of heterogeneous

healthcare institutions. In addition, the explainability component of the EFVLL framework provides visual attention maps between the generated diagnostic statement and the relevant anatomical areas where the reports were based.

Finally, through the use of secure parameter aggregation and Differential Privacy, the EFVLL framework extends the protections afforded by patient data, making it appropriate to use in privacy sensitive settings. The overall construction of the EFVLL framework represents a complete approach to collaborating on, explaining and protecting patient data while generating AI based medical reports. As this study only presents a conceptual and simulation-based evaluation of the framework, this study provides the foundation from which to develop a trustworthy AI based radiological systems in the future.

5.2 Future Scope

There are various other possible research directions that could expand on the use of the EFVLL framework:

- To validate the use of the EFVLL framework using data sets from large-scale clinical institutions over multiple sites;
- An expansion of the EFVLL framework for imaging modalities beyond XRAY, including CT, MRI, and Ultrasound;
- By exploring more complex Vision-Language Models and Large-Language Models for better quality reporting;
- By using explainability techniques such as Concept-Based Explanations and Counterfactual Reasoning to provide better model intelligibility;
- Methodologies/techniques must be developed for communication-efficient Federated Optimization to reduce the synchronization requirements of large healthcare networks;
- Strategies to develop personalized Federated Learning and adapt the Global Model to local institution characteristics;
- Prospective evaluation of the EFVLL framework in Workflow with Radiologist feedback to determine how to improve diagnostic accuracy and reporting efficiency;
- Evaluate and develop the technologies/ infrastructure for hospital information system integration and deployment on respective Electronic Health Records and ensure compliance with regulatory privacy requirements for use in actual clinical settings.

All of these research directions will enable the establishment of AI Systems for automated Medical Reporting that can be widely and effectively utilized in the Federation of Distributed Health Systems.

Reference

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).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305-311.

Bahaley, N., & Rishi, A. (2025). INNOVATIVE ADVANCES IN SMART HEALTHCARE AND SECURE FEDERATED LEARNING FOR MEDICINE. *THE HYBRID MIND: MULTIDISCIPLINARY PATHWAYS IN THE COGNITIVE ERA*, 97.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 119.

Puttanapura, J., & Doss, S. (2025). Optimizing neural disorder treatment through federated learning and multi-institutional data collaboration. *Federated Learning for Neural Disorders in Healthcare 6.0*, 120.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, *2*, 429-450.

Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. *Foundations and trends in machine learning*, *14*(1-2), 1-210.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 119.

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1-19.

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., ... & He, B. (2021). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3347-3366.

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305-311.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, *39*(6), 1856-1867.

Huang, W., Li, C., Yang, H., Liu, J., Liang, Y., Zheng, H., & Wang, S. (2024). Enhancing the vision–language foundation model with key semantic knowledge-emphasized report refinement. *Medical image analysis*, *97*, 103299.

Lang, W., Liu, Z., & Zhang, Y. (2025). DACG: Dual Attention and Context Guidance model for radiology report generation. *Medical Image Analysis*, *99*, 103377.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). Pmlr.

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*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., Peng, Y., ... & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, *23*(2), 304-310.

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).

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).

Wang, S., Peng, B., Liu, Y., & Peng, Q. (2023, December). Fine-grained medical vision-language representation learning for radiology report generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 15949-15956).

Ji, J., Hou, Y., Chen, X., Pan, Y., & Xiang, Y. (2024). Vision-language model for generating textual descriptions from clinical images: model development and validation study. *JMIR Formative Research*, *8*, e32690.

Hartsock, I., & Rasool, G. (2024). Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, *7*, 1430984.