PeFoMed-Light: A Lightweight and Parameter-Efficient Vision–Language Architecture for Automated Medical Report Generation

K. Varaprasad Reddy (CS23BT050)
Pasham Pardeev Narsi Reddy (CS22BT040)
Ch. Himesh Raj Kumar (CS23BT005)
Department of Computer Science and Engineering
Indian Institute of Technology Dharwad, Karnataka, India

***Abstract*—Generating radiology reports from chest X-ray images requires models capable of visual interpretation and clinically consistent language generation. However, most public datasets provide only image–label pairs rather than narrative reports, limiting supervised training. This work presents a parameter-efficient medical report generation framework built on a modified PeFoMed architecture. We replace the LLaMA-based decoder with GPT-2 for computational feasibility and use CLIP ViT-L/14 as the frozen vision encoder. A linear up-projection and image-projection bridge convert visual embeddings into the decoder space, while LoRA adapters enable lightweight fine-tuning. To compensate for the absence of paired reports, we construct a synthetic dataset of 10 868 image–report pairs using LLaMA few-shot prompting with CheXpert label vectors. Experiments show that the model produces coherent, clinically aligned "Findings" sections on held-out images. The pipeline demonstrates that parameter-efficient tuning can scale multimodal clinical report generation under restricted data and hardware.**

***Index Terms*—Medical report generation, multimodal learning, CLIP, GPT-2, LoRA, PeFoMed, chest X-rays.**

I. Introduction

Chest X-rays are among the most widely used imaging modalities for diagnosing thoracic conditions. Automating the generation of radiologist-style reports from these images has attracted significant attention due to advances in multimodal large language models (MLLMs). However, building such models typically requires paired image–report datasets, which are large, proprietary, or computationally expensive to handle.

CheXpert provides high-quality labels for over 220 000 radiographs but lacks narrative reports. MIMIC-CXR includes free-text reports but exceeds 300 GB and requires extensive preprocessing. These challenges motivate parameter-efficient methods capable of training on smaller or synthetic datasets without compromising clinical structure.

This work implements a lightweight adaptation of the PeFoMed framework using CLIP as a frozen vision encoder, GPT-2 as a frozen text decoder, and LoRA adapters for efficient fine-tuning. A synthetic dataset of clinically structured "Findings" is generated using LLaMA few-shot prompting. The resulting PeFoMed-Light system is computationally tractable and produces clinically meaningful narratives.

II. Related Work

Early medical captioning relied on convolutional encoders and recurrent networks, which produced short template-like sentences. Transformer-based approaches improved contextual modeling but required large paired datasets.

The IU X-Ray dataset provides roughly 7 000 paired reports. MIMIC-CXR is significantly larger but often impractical for small hardware environments. CheXpert offers high-quality labels but not textual findings.

PeFoMed proposed a modular architecture combining a frozen vision encoder, a frozen LLM, and lightweight adapters. LoRA has become a preferred technique for adapting large models due to its low memory footprint and minimal number of trainable parameters.

CLIP, trained on large-scale image–text pairs, provides robust visual embeddings suitable for downstream tasks in both natural and medical imaging domains.

III. Dataset Construction

A. Synthetic Findings Generation

CheXpert provides 14 diagnostic labels per image but no associated narrative. To obtain clinically meaningful "Findings", each label vector is transformed into a descriptive paragraph using a few-shot prompting strategy. A small reference set of IU X-Ray reports is included to guide structure, phrasing, and radiology-specific terminology.

A modern LLM (LLaMA-2/7B) is prompted with:

- CheXpert diagnostic labels,
- examples of real radiology findings,
- structural instructions emphasizing lungs, heart, pleura, and bones.

Malformed or incomplete generations are removed during postprocessing. The final dataset contains 10 868 paired samples.

B. Preprocessing

All radiographs are converted to grayscale, resized to a consistent resolution, and normalized following CLIP preprocessing. Synthetic findings are tokenized using GPT-2's byte-pair encoding (BPE).

A standard 80/10/10 split is used for training, validation, and testing, ensuring adequate variation for generalization.

IV. Model Architecture

A. Overview

The proposed PeFoMed-Light architecture follows the overall philosophy of the original PeFoMed framework but introduces structural modifications to make the model feasible on moderate hardware. Instead of relying on the EVA-G encoder and a large LLaMA decoder, the lightweight design freezes both the visual and language backbones and trains only a small number of projection parameters and LoRA modules. The architecture has three major components:

1) a frozen CLIP ViT-L/14 encoder for extracting patch-level image representations,
2) a two-stage projection module that adapts visual embeddings into GPT-2's hidden space, and
3) a GPT-2 decoder augmented with LoRA for parameter-efficient autoregressive generation.

B. Visual Feature Extraction

A CLIP ViT-L/14 encoder processes each X-ray image I and produces a sequence of patch embeddings:

$$E_{\text{clip}} = \text{CLIP}(I) \in \mathbb{R}^{N \times d_v},$$

where N denotes the number of patches and d_v is the CLIP embedding dimension (1024 for ViT-L/14). CLIP's large-scale pretraining on hundreds of millions of image–text pairs provides strong generalization, enabling the model to extract semantically meaningful features even from grayscale medical images. The CLIP encoder is kept fully frozen, preserving its learned visual alignment while reducing computational cost.

C. Up-Projection and Image-Projection

GPT-2 operates in a different embedding space with hidden dimension $d_h = 4096$. To reconcile this mismatch, PeFoMed-Light uses a pair of learnable projection layers. The first, the *up-projection*, expands the CLIP embedding dimension to an intermediate high-dimensional space:

$$E_{\text{up}} = W_u E_{\text{clip}} + b_u, \qquad W_u \in \mathbb{R}^{5632 \times d_v}.$$

The second transformation, the *image-projection* layer, maps this expanded representation into GPT-2's hidden dimension:

$$E_{\text{proj}} = W_p E_{\text{up}} + b_p, \qquad W_p \in \mathbb{R}^{d_h \times 5632}.$$

Together, these layers act as a multimodal bridge, converting CLIP's visual tokens into tokens that can be processed by GPT-2 without modifying the backbone. The output E_{proj} forms a sequence of pseudo-text tokens aligned to GPT-2's embedding manifold.

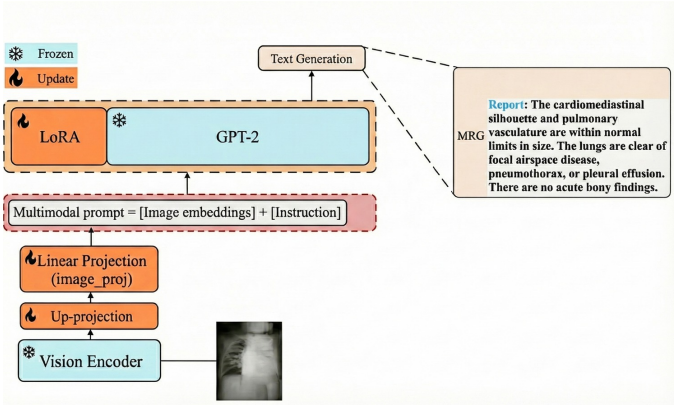

Fig. 1. Modified PeFoMed-Light architecture using CLIP as the frozen vision encoder, GPT-2 as the text decoder, and LoRA adapters for parameter-efficient tuning.

D. GPT-2 Decoder with LoRA

The GPT-2 decoder is used as a frozen autoregressive language model responsible for generating the "Findings" section. Instead of fine-tuning all parameters—which would be computationally expensive—LoRA adapters are inserted into the attention projection matrices. For a frozen weight matrix W, LoRA introduces a low-rank update:

$$W' = W + BA,$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are learnable low-rank matrices with $r \ll d$. Only A and B are updated during training, yielding a small number of trainable parameters while allowing GPT-2 to adapt effectively to medical-domain text. Keeping the GPT-2 backbone fixed is particularly desirable when training on synthetic data.

E. Model Input and Training Objective

To enable multimodal generation, the projected visual tokens are concatenated with an instructional prefix. The final input sequence to GPT-2 is:

$$X = [E_{\text{proj}},\ \text{"Describe the X-ray findings."},\ \text{BOS}],$$

where the instruction serves as a natural-language cue guiding GPT-2 to produce clinically structured text. GPT-2 then generates a sequence of tokens $(y_1, \ldots, y_T)$ corresponding to the findings.

Training minimizes the standard autoregressive negative log-likelihood:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, X).$$

Because only the LoRA and projection parameters are updated, optimization is efficient and stable even with limited hardware.

TABLE I
VALIDATION PERFORMANCE ACROSS EPOCHS. EM IS 0.000 FOR ALL EPOCHS.

Epoch	Loss	ROUGE-L	METEOR	CIDEr
1	0.9069	0.1447	0.1146	0.0966
2	0.4718	0.1589	0.1042	0.1003
3	0.4718	0.1712	0.1180	0.1125

V. TRAINING PROCEDURE

Training is performed on Apple M-series hardware using the Metal (MPS) backend. Due to memory limits, small batch sizes are used and the model is trained for three epochs.

Only the up-projection, image-projection, and LoRA parameters receive gradients; the CLIP encoder and GPT-2 backbone remain frozen. Training logs show stable loss curves without divergence, confirming that parameter-efficient tuning is feasible on moderate hardware.

VI. EVALUATION

Evaluating medical report generation is challenging because radiology text is free-form and CheXpert does not provide ground-truth reports. The synthetic LLaMA-generated findings therefore serve as reference targets. While this introduces limitations, it allows the use of standard text similarity metrics.

A. Training Dynamics

Table I summarizes the evolution of validation metrics across epochs. The average training loss decreases from 0.9069 in Epoch 1 to 0.4718 in Epochs 2 and 3. ROUGE-L and CIDEr steadily improve, indicating that multimodal alignment improves over training, while METEOR fluctuates slightly. Exact Match (EM) remains zero throughout, which is expected for free-form medical narratives.

B. Test-Set Metrics

On the held-out test split, the model attains:

- ROUGE-L: 0.129
- METEOR: 0.095
- CIDEr: 0.173
- Exact Match (EM): 0.000

The CIDEr score is higher on the test set than on validation, suggesting that the model generalizes reasonably well to unseen images despite training on synthetic supervision.

C. Metric Interpretation

We use four standard natural language generation (NLG) metrics: ROUGE-L, CIDEr, METEOR, and EM. Together, they capture complementary aspects of similarity between generated and reference reports.

- **ROUGE-L** measures longest common subsequence similarity, reflecting structural overlap between generated and reference text.
- **CIDEr** uses TF–IDF weighting to assess consensus between texts and rewards clinically meaningful keywords (e.g., "effusion", "opacity", "cardiac silhouette").
- **METEOR** accounts for synonymy and paraphrasing using alignment-based scoring; lower scores are expected when clinically equivalent sentences use different phrasing.
- **EM** is a strict string-equality metric that is typically near zero for radiology because there are many valid ways to describe the same study.

Across epochs, ROUGE-L and CIDEr exhibit consistent gains, confirming that visual features meaningfully influence generation. Although METEOR and EM remain modest, qualitative review shows that key abnormalities such as cardiomegaly, pleural effusion, and pulmonary edema are correctly mentioned when present, and normal cases are described without hallucinated findings.

D. Qualitative Observations and Limitations

Manual inspection reveals that fine-tuned reports exhibit:

- clearer descriptions of lung fields (e.g., "no focal airspace disease"),
- more precise cardiovascular observations (e.g., "cardiomediastinal silhouette within normal limits"),
- appropriate use of negation for common pathologies, and
- fewer ambiguous or incomplete sentences relative to a zero-shot GPT-2 baseline.

Nevertheless, several limitations remain:

- **Subtle findings:** mild vascular congestion, interstitial changes, and small devices are sometimes omitted.
- **Synthetic supervision bias:** the style of LLaMA-generated findings is reflected in the model, which may differ from real radiology reporting conventions.
- **Uncertainty handling:** CheXpert uncertainty labels are often resolved into binary statements in the synthetic reports, which removes nuance.
- **Phrasing variability:** multiple valid ways to state the same finding reduce automatic metric scores even when the underlying clinical meaning is correct.

TABLE II
COMPARISON OF PEFOMED-LIGHT (OURS) WITH PRIOR MEDICAL REPORT GENERATION MODELS ON THE IU–XRAY DATASET. NOTE THAT BASELINE MODELS ARE TRAINED ON REAL PAIRED REPORTS, WHEREAS PEFOMED-LIGHT IS TRAINED USING SYNTHETIC SUPERVISION.

Method	**Model Type**	**METEOR**	**ROUGE-L**	**CIDEr**
R2Gen [3]	Non-LLM	0.211	0.377	0.438
BiomedGPT [54]	LLM	0.146	0.302	0.360
PeFoMed (Paper)	LLM	0.157	0.286	**0.462**
PeFoMed-Light (Ours)	LLM	**0.095**	**0.129**	**0.173**

VII. DISCUSSION

The results highlight the potential of parameter-efficient multimodal tuning for medical report generation, particularly when large paired datasets are unavailable. The combination of CLIP's transferable image features and GPT-2's stable language modeling, augmented with LoRA, enables effective learning from synthetic text without overfitting.

Synthetic supervision offers a practical workaround for limited labeled datasets but introduces challenges:

- it is difficult to quantify clinical correctness when references are not written by radiologists,
- evaluation metrics fail to capture clinical nuance when phrasing differs, and
- domain gaps may emerge between synthetic Findings and real-world reporting styles.

Despite these issues, the model's ability to correctly describe major thoracic findings demonstrates that multimodal alignment is learned effectively.

Future work may explore:

- integrating real paired datasets such as IU X-Ray or curated MIMIC subsets for fine-tuning,
- replacing CLIP with state-space vision encoders such as Vmamba for improved spatial reasoning,
- reinforcement learning from human feedback (RLHF) to encourage factual correctness and style control, and
- confidence-aware or uncertainty-aware text generation for more realistic reporting.

VIII. Conclusion

This study presents PeFoMed-Light, a lightweight and computationally efficient framework for automated chest X-ray report generation. By combining a frozen CLIP encoder, a GPT-2 decoder equipped with LoRA adapters, and a synthetic dataset of over 10 000 LLaMA-generated findings, the system learns to map chest X-ray images into coherent and clinically aligned descriptions. Despite the limitations of synthetic supervision, the model demonstrates measurable improvements across multiple NLG metrics and strong qualitative performance.

The approach provides a practical blueprint for developing multimodal medical NLP systems under hardware and data constraints. With further refinement and incorporation of real paired reports, PeFoMed-Light can evolve into a scalable solution for assisting radiologists and supporting automated triage systems.

References

[1] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels," *AAAI*, 2019.

[2] D. Demner-Fushman *et al.*, "Preparing a Collection of Radiology Examinations for Distribution and Retrieval," *JAMIA*, 2015.

[3] A. E. W. Johnson *et al.*, "MIMIC-CXR: A Large Publicly Available Database of Chest Radiographs and Reports," *Scientific Data*, 2019.

[4] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021.

[5] E. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *ICLR*, 2022.

[6] J. He *et al.*, "PeFoMed: Parameter-Efficient Foundation Model for Medical Report Generation," *arXiv:2401.02797*, 2024.