

A Disease-Memory Guided Framework for Radiology Report Generation from Multi-View Chest X-rays

Prajakta Mandar Dhamanskar¹, Chintan Thacker², Saurabh Shah³

Department of Computer Science and Engineering, Parul University, Vadodara, Gujarat, India

Corresponding author: Prajakta Mandar Dhamanskar (e-mail: prajakta.dhamanskar@fragnel.edu.in), **Author(s)**
Email: Chintan Thacker (e-mail: chintanthacker450@gmail.com), Saurabh Shah (email: saurabh.shah37287@paruluniversity.ac.in)

Abstract Chest X-rays are widely used, low-cost diagnostic tools. Writing chest X-ray reports manually is a time-consuming and tedious task. The increasing number of X-rays burdens radiologists, creating a need for a clinically accurate, fast and automated report generation system. To solve this challenge, we propose a new disease-memory-guided vision-language framework for radiology report generation. Frontal and lateral chest X-rays (multi-view) are encoded using a Swin-Transformer. These encoded visual features are then passed through a multi-label classification head that predicts disease probabilities, from which the top three labels are converted into learnable disease-memory embeddings. These are concatenated with projected visual tokens and fixed task instruction embeddings to form a multimodal prompt which is fed into a frozen decoder. A LLaMA-2-7B model adapted to the radiology domain using LoRA fine-tuning on the train-only MIMIC-CXR subset is used as an autoregressive decoder. The disease-memory embeddings serve as disease guidance during report generation. An extensive experimental analysis of different experimental variants is conducted on the benchmark Indiana University Chest X-ray dataset using three aspects: 1) statistical n-gram overlap, 2) semantic similarity, and 3) clinical label-based metrics. Experimental results of the proposed system show improvements in semantic evaluation metrics, with BERTScore precision, recall, and F1 values of 0.898, 0.891 and 0.894, respectively, and ROUGE-L and METEOR scores of 0.327 and 0.281, indicating improved contextual consistency between generated and reference reports. However, performance on statistical metrics (e.g., BLEU4=0.068) and clinical evaluation metrics remains limited, with Micro-F1, Macro-F1, Precision and Recall values of 0.4211, 0.0640, 0.4272, and 0.4151, respectively, indicating only partial abnormality description. This framework improves abnormality grounding while maintaining computational efficiency and can be used as a supportive tool for preliminary report drafting in research or resource-constrained settings.

Keywords Radiology Report Generation; Disease-Memory Prefix; Large Language Model; Chest X-ray reporting; Vision Language Models.

1. Introduction

Chest X-rays are among the most widely used and cost-effective diagnostic tools in medical practice. Writing radiology reports manually is a time-consuming, tedious task that can lead to reporting errors [1]. To overcome these challenges such as increasing imaging needs, workload in reporting, and expert shortages, automated radiology report generation (R2Gen) systems have been researched extensively. A study presented in [2] introduced the role of keyword-based AI in this regard. Radiology report generation using chest X-rays can be termed a multimodal problem, as it uses both image and textual modalities. This task differs from regular image captioning in terms of domain-specific terminology needed, fine anomaly detection, and accurate

localization. Furthermore, the publicly available chest X-ray datasets contain a high proportion of normal studies, which biases models toward frequent normal-report templates and reduces sensitivity to abnormal findings [4], [22]. Rule-based or template retrieval strategies were used by early methods for the chest X-ray report generation task. These techniques were computationally efficient but often produced rigid and repetitive reports which often failed to describe complex abnormalities [3], [5]. Later, encoder-decoder frameworks improved report fluency by learning image-to-text mappings but often lacked clinical consistency [6]. More recent transformer-based and memory-guided models, such as memory-driven transformers, cross-modal memory networks, and retrieval-enhanced systems, have enhanced contextual

reasoning and long-range dependencies [7], [8], [9], [10], [11], [12]. In addition, vision-language models and large language models (LLMs) have further advanced the field by combining powerful visual encoders with pre-trained language models [13], [14], [15], [16], [17].

Despite these advancements, there are still several important limitations. First, commonly used evaluation metrics such as BLEU, ROUGE, and CIDEr do not always indicate true clinical consistency [4], [18], [19]. Second, many models tend to generate generic normal-style reports, often missing clinically significant abnormalities such as pleural effusion or masses. Third, in practical scenarios, fine-tuning large language models can be computationally expensive. Finally, successful integration of disease-aware clinical context into report generation remains an open challenge. To overcome the limitations, this study aims to improve abnormality-sensitive chest X-ray report generation by incorporating disease-aware semantic guidance into a vision-language framework while preserving report fluency and computational efficiency. We propose a disease-guided report generation pipeline that combines structured disease predictions and visual features before decoding. This framework contrasts with previous methods such as R2GenCMN that incorporates feature retrievals from external memory banks and R2GenGPT, which adopts text prefix prompting through a pre-trained LLM; our framework employs a structured disease-memory embedding strategy where the disease predictions are first embedded as learnable embeddings and are subsequently incorporated into the visual feature spaces before decoding, allowing for continuous cross-modal conditioning compared to discrete memory retrievals or text promptings. This design allows more explicit incorporation of disease-specific information into the generation process. The main contributions of this study are as follows

1. **Disease-Guided Multimodal Framework:** We present a vision-language architecture that integrates predicted disease labels as semantic guidance, enabling the model to better focus on clinically significant abnormalities rather than relying exclusively on visual embeddings.
2. **Structured Disease-Memory Embedding (Key Novelty):** In contrast to R2GenCMN, which utilizes external memory lookups, and R2GenGPT, which depends on textual prefixes, our approach converts disease predictions into learnable embeddings and integrates them with visual features prior to decoding. This enables continuous cross-modal conditioning, representing a distinct architectural contribution

compared to existing memory-based and prompting-based methods.

3. **Auxiliary Disease Supervision for Alignment:** The model incorporates an auxiliary multi-label disease classification objective to enhance alignment between visual features and generated text, improving semantic consistency between reports and underlying pathologies.
4. **Parameter-efficient Training Strategy:** The proposed framework uses a frozen Large Language Model (LLM) along with LoRA fine-tuning, which helps to reduce computing expenses at the same time maintaining strong language generation capabilities.
5. **Comprehensive Evaluation on IU X-ray Dataset:** The proposed method is evaluated on the IU X-ray dataset using both quantitative metrics and qualitative analysis, demonstrating improved sensitivity to abnormalities.

II. Related Work

Automated chest X-ray report generation has progressed from rule-based systems to advanced vision-language models. Early studies relied on template-based methods [3], which used predefined report templates or adapted based on detected findings. These methods can generate grammatically correct reports for normal reports but lack the flexibility needed for generating complicated or patient-specific abnormal findings. Later, encoder-decoder models emerged as a leading approach, employing an image encoder to extract visual features and a language decoder to generate reports. The conditioned transformers framework using pre-trained image and language modules, and ChestBioX-Gen that uses BioGPT with co-attention, were proposed in [6] and [13], respectively. Recent multimodal transformer-based approaches have further integrated visual features with language decoders [20]. Warm-starting the decoder with medical knowledge improves performance and training efficiency [21]. However, the difficulty of learning strong image-grounded report generation using such methods is discussed in [22]. In order to improve visual-text alignment, many prior works have explored cross-modal fusion and attention mechanisms, including multi-scale feature fusion [23]. The multimodal transformer for radiology report generation is used in [9]. Visual prior-based alignment is adopted in [24], while [25] has used a cross-attention technique for diagnosis report generation. Recent approaches such as KIA introduce knowledge-guided implicit vision-language alignment [26], while RadAlign

explicitly aligns visual features with medical diagnostic concepts before report generation [27]. Region-aware and knowledge-guided techniques have been proposed to enhance interpretability and localization. Anatomical structure using region masks and an interactive region-guided approach is used in [28] and [29], respectively. In order to improve contextual reasoning, [10] developed a learned knowledge base with multimodal alignment, whereas [11] and [12] used memory-based models. Segmentation-aware multimodal LLMs have recently incorporated anatomical masks and mask-aware prompting for report generation [30]. These approaches improve interpretability and consistency, but they often require additional annotations or external supervision.

Recent advances include LLMs and vision-language models. GPT-based models are used in [15], and [16] introduced R2GenGPT, which incorporates a frozen LLM decoder to maintain fluent language generation with minimal training cost. The use of SERPENT-VLM for iterative generation of reports is demonstrated in [17]. CXR-LLaVA further applies a multimodal LLM to chest X-ray interpretation [31]. The increasing role of foundation models in medical imaging is discussed in [32] and [33]. Clinical evaluation of such models has also revealed clinically significant reporting errors, highlighting the need for clinician oversight [34]. Recent uncertainty-estimation studies further emphasize the need to assess the reliability of LVLM-generated reports beyond conventional text-generation metrics [35]. However, these models can generate fluent but clinically irrelevant reports when disease-specific visual evidence is not incorporated, as most approaches rely on implicit prompting or textual conditioning rather than structured disease-aware embeddings. This limitation has been addressed in [36] through contrastive learning. Overall, previous studies emphasize ongoing research into improving the alignment between visual features and text, making it more interpretable, and applying the large pre-trained models for radiology report generation. However, reliable abnormality grounding and clinical accuracy remain challenging, particularly for subtle and localized findings.

III. Materials and Methods

A. Dataset

The Indiana University (IU) Chest X-ray dataset [37], is used. It contains 7,470 images (approximately two per report: frontal and lateral) from 3,955 patients, as shown in Table 1. Samples without images are excluded. Each training sample consists of a pair of multi-view chest X-rays and its corresponding findings report. Since the

MIMIC-CXR dataset [38] is huge in size, a subset of the dataset is used for pre-training. The dataset consists of more than 370,000 chest X-rays accompanied by their respective free-text radiology reports. Paired image-report data is used along with consistent preprocessing for the reports. There are 14 commonly used clinical labels in the dataset; however, ‘No Finding’ is excluded in this study since the focus is on generating reports based on abnormalities, leading to the use of 13 disease classes in the disease-memory module. These disease classes include Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. Uncertain labels are considered negative when training the model. Training, validation, and test datasets are patient-wise separated with the ratio of 70% / 15% / 15%, respectively (2,768 / 593 / 594 patients), to prevent data leakage. No explicit anatomical segmentation or external annotations are used in the proposed framework. The key novelty of the proposed method lies in the disease-memory embeddings, which convert predicted disease labels into semantic prompts to generate reports. This allows for better grounding of abnormalities than the purely visual-based methods.

Table 1. Statistics and data-split characteristics of the Indiana University Chest X-ray dataset.

Description	Value
Number of Images	7,470
Number of Patients	3,955
Average Words per Report	50-70 words
Report Vocabulary Size	6,000 tokens
Radiology Report Sections	Findings, Impression
Number of Common Clinical Labels	14 (13 used in this study)
Typical Image Size	224 x 224 pixels (resized)
Train/Validation/Test Split	70 %/15% / 15% (patient-level)

Fig. 1 illustrates the overall architecture, and the algorithm is presented in Algorithm 1. A Swin-Transformer extracts features from frontal and lateral views, which are fused and projected into the language embedding space. Disease probabilities are predicted using a classification head, and the top-k disease labels are encoded into disease-memory embeddings. These embeddings are concatenated with visual tokens and task instruction embeddings to form a multimodal prompt. This prompt is processed by a frozen MedLLaMA decoder with LoRA fine-tuning, to generate reports auto-regressively. The framework

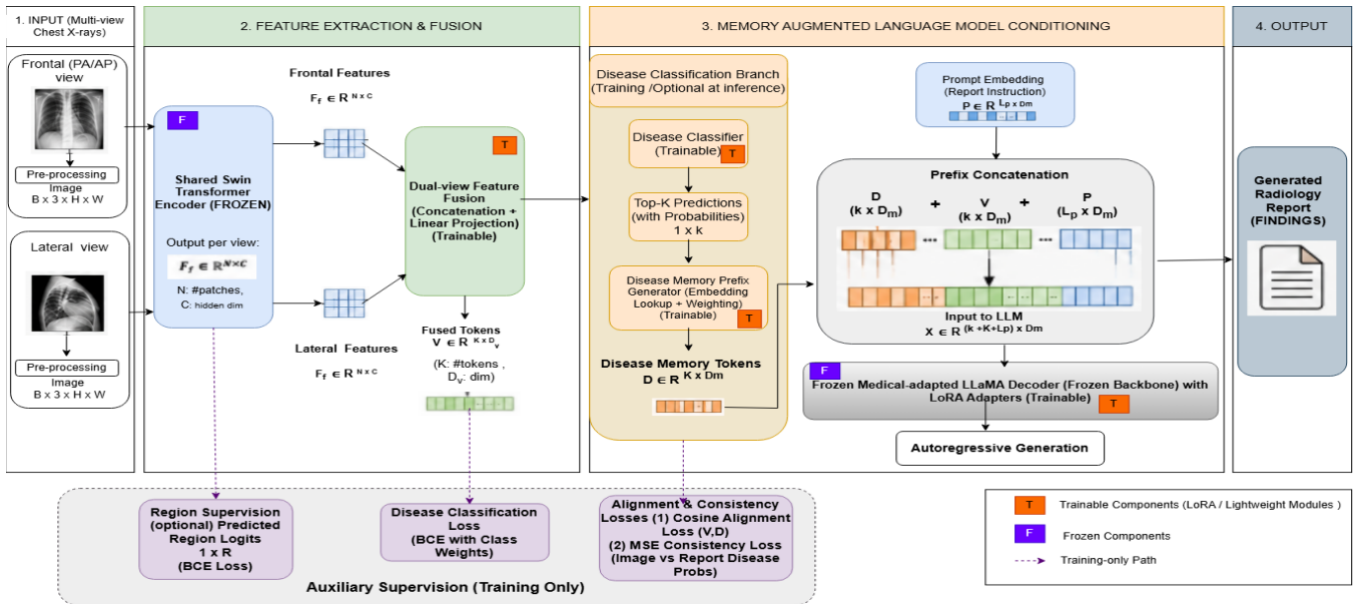


Fig. 1. Proposed disease-memory guided framework showing multi-view chest X-ray encoding, feature fusion, disease prediction and memory embedding, multimodal prompt construction, and autoregressive radiology report generation, disease prediction and memory embedding, multimodal prompt construction, and autoregressive radiology report generation.

improves clinical consistency while maintaining computational efficiency.

B. Visual Encoder

We use a pre-trained Swin-Transformer (swin_base_patch4_window7_224) as a visual feature extractor. The encoder is selected for its effectiveness in extracting hierarchical visual features and is initialized with ImageNet-pre-trained weights. Frontal (PA/AP) and lateral chest X-ray images are provided as dual-view inputs to a shared Swin-Transformer encoder as shown in Eq. (1) [8]. Let B denote the batch size, and H and W denote the image height and width, respectively. Let N denote the number of visual tokens, d_v the visual feature dimension and d_l the hidden dimension of the language model.

$$I_f \in R^{B \times H \times W \times 3}, I_l \in R^{B \times H \times W \times 3} \quad (1)$$

Since chest X-ray images are originally greyscale, each image is converted to a 3-channel format by duplicating the single channel and resized to 224×224. Eq. (2) gives the image features extracted by the encoder [8].

$$F_f = E(I_f), \quad F_l = E(I_l) \quad (2)$$

where $E(\cdot)$ denotes the encoder and $F_f, F_l \in R^{B \times N \times d_v}$. Final-stage features (dimension 1024) are extracted and fused as shown in Eq. (3) [23].

$$F = [F_f; F_l] W_f + b_f \quad (3)$$

where $[F_f; F_l] \in R^{B \times N \times 2d_v}$, $W_f \in R^{2d_v \times d_v}$ and $b_f \in R^{d_v}$. Only the final stage is fine-tuned, while earlier

layers remain frozen to preserve pre-trained knowledge and reduce computational cost.

C. Feature Projection Layer

To align visual features in Eq. (4) with the decoder embedding space, a learnable projection is applied [6], [9].

$$V \in R^{B \times N \times d_v} \quad (4)$$

where F denotes fused visual features, B is the batch size, N is the number of visual tokens, and d_v is the feature dimension. The projected visual features are obtained as shown in Eq. (5).

$$\hat{V} = V W_p + b_p \quad (5)$$

where $W_p \in R^{d_v \times d_l}$ is a trainable projection matrix, $b_p \in R^{d_l}$ is a bias vector (broadcast across tokens). The multiplication $F W_p$ denotes a linear transformation applied independently to each token along the feature dimension. The projected visual features $\hat{V} \in R^{B \times N \times d_l}$ are treated as visual prefix tokens and concatenated with disease-memory embeddings and task prompts to form a unified multimodal prefix.

D. Disease-Memory Prefix Module

To incorporate structured clinical knowledge, a disease-memory prefix module is introduced. A multi-label classifier predicts disease probabilities as shown in Eq. (6) [1].

$$D = [d_1, d_2, \dots, d_M] \in [0,1]^M \quad (6)$$

Algorithm 1. Training Procedure for the Proposed Disease-Memory-Guided Multimodal Report Generation Framework**Input:**

Training Dataset $D = \{(I_f, I_l, Y, Y_{\text{labels}})\}$, Disease-memory embedding matrix E_d , Top-K value k , Loss weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, Learning rate η , Maximum number of training epochs MaxEpochs

Output:

Optimized trainable parameters θ^*

Initialization

1. Load ImageNet-pretrained Swin Transformer encoder θ_{enc} ; freeze earlier stages and fine-tune the final stage.
2. Load radiology-adapted MedLLaMA parameters θ_{LM} ; freeze the base language-model parameters and initialize trainable LoRA parameters θ_{LoRA} .
3. Initialize the trainable disease-classification parameters θ_{cls} , projection parameters θ_{proj} , report-disease classification parameters θ_{rep} , disease-memory embeddings E_d , cross-modal alignment parameters, and token-gating parameters.

Training Procedure

4. FOR epoch = 1 to MaxEpochs do
5. FOR each mini-batch $(I_f, I_l, Y, Y_{\text{labels}}) \in D$ do
6. Extract frontal and lateral visual features using Swin encoder.
7. Fuse the multi-view visual features using Eq. (3).
8. Project the fused visual features into the language-model embedding space using Eq. (5).
9. Predict disease probabilities using Eq. (6).
10. Select the Top-(K) diseases and their probabilities using Eq. (7) and Eq. (8).
11. Retrieve the corresponding disease-memory embeddings using Eq. (9).
12. Apply visual token compression and gating using Eq. (21) and Eq. (22).
13. Construct the multimodal prompt by concatenating disease-memory embeddings, gated visual tokens, and task-prompt embeddings using Eq. (10).
14. Apply the cross-modal alignment module using Eq. (11).
15. Generate the report using the adapted LLaMA-2 decoder with teacher forcing according to Eq. (12) and Eq. (13).
16. Compute the report-level disease probabilities from the decoder hidden states.
17. Compute the language-generation, image-disease, report-disease, alignment, and consistency losses using Eq. (16), Eq. (17), Eq. (18), Eq. (19) and Eq. (20) respectively.
18. Compute the total loss using Eq. (15).
19. Update all trainable parameters using gradient-based optimization: $\theta = \theta - \eta \nabla \theta L_{\text{total}}$
20. END FOR
21. END FOR
22. Select the checkpoint corresponding to the lowest validation loss obtained during training.
23. Return the optimized trainable parameters θ^* .

where $M=13$, represents the total number of disease classes and $d_i \in [0,1]$ represents the predicted probability of i^{th} disease. The classifier head is first trained with the help of the disease labels automatically extracted from radiology reports using CheXbert. It is then optimized with binary cross-entropy loss. This guarantees consistency between training and inference. To focus on the most relevant clinical findings, the top-K diseases are selected based on their

confidence scores. Formally, given D , the Top-K selection operator returns the index set I_K as shown in Eq. (7) corresponding to the K highest-confidence disease categories. K is chosen based on validation performance to balance relevant abnormality coverage and noise reduction.

$$I_K = \text{TopK}(D, K), I_K \subseteq \{1, \dots, M\}, |I_K| = K \quad (7)$$

The selected probabilities are given in Eq. (8).

Table 2. Training hyperparameters, loss weights, and inference settings of the proposed framework.

Parameter	Value
Vision Encoder	Swin-Base Transformer
Language Decoder	LLaMA-2-7B (MedLLaMA)
Fine-tuning Strategy	Frozen Decoder + LoRA
Number of Disease Classes	13
Top-k Disease Prefix Tokens	3
Epochs	10
Learning Rate	1e-5
Batch Size	4
Disease Loss Weight	0.2
Alignment Loss Weight	0.05
Decoding Strategy	Beam Search
Number of Beams	2
Max Generated Tokens	80
No Repeat N-gram Size	3
Repetition Penalty	1.1

E. Prompt Construction and Multimodal Conditioning

A unified multimodal prompt is constructed by combining three components: projected visual features, disease-memory embeddings, and task-specific instruction embeddings. Let the projected visual features be: $\hat{V} \in \mathbb{R}^{B \times N_v \times d_l}$, where B is batch size, N_v is the number of visual tokens and d_l is the embedding dimension of the language model. The disease-memory embeddings are denoted as $D_{\text{mem}} \in \mathbb{R}^{B \times k \times d_l}$ where k is the number of disease tokens and task-specific instruction embeddings are denoted as $P \in \mathbb{R}^{B \times N_t \times d_l}$, where N_t is the number of instruction tokens. The multimodal prompt is constructed by concatenating these components along the token dimension as shown in Eq. (10) [39].

$$S = [D_{\text{mem}}; \hat{V}; P] \in \mathbb{R}^{B \times (N_v + K + N_t) \times d_l} \quad (10)$$

where $[\cdot; \cdot]$ denotes sequence-wise concatenation along the token axis. The combined sequence is further refined through a cross-modal alignment module as per Eq. (11).

$$S_f = \text{CrossModalAlign}(S), \quad S_f \in \mathbb{R}^{B \times (N_v + k + N_t) \times d_l} \quad (11)$$

The CrossModalAlign(.) is a module having a stack of transformer layers that perform self-attention over the entire multimodal sequence [23], [24]. During training, the aligned multimodal prefix S_f is prepended to the ground-truth report tokens. The language model is

Table 3. Component-wise configuration of experimental variants (E1–E6), showing progressively enabled supervision, alignment, and knowledge augmentation modules

Variant	Vision	Disease	Region	Alignment	Gating	Knowledge Augmentation
E1	✓	X	X	X	X	X
E2	✓	✓	X	X	X	X
E3	✓	✓	✓	X	X	X
E4	✓	✓	✓	Shallow	X	X
E5	✓	✓	✓	Learned	✓	X
E6	✓	✓	X*	Learned	✓	✓

Note: ✓ = component enabled; X = component disabled. *Region-level supervision was disabled for E6. Shallow = direct projection/fusion; Learned = trainable cross-modal alignment.

$$D_{\text{top-K}} = \{d_{i_1}, d_{i_2}, \dots, d_{i_K}\} \in \mathbb{R}^K, \quad i_j \in I_K \quad (8)$$

-where indices i_j correspond to the K highest predicted probabilities [12]. Corresponding embeddings are retrieved from a learnable table $W_e \in \mathbb{R}^{M \times d_l}$. The resulting disease prefix embeddings are defined as shown in Eq. (9).

$$P = [e_i | i \in I_K], \quad e_i = W_e[i], \quad P \in \mathbb{R}^{K \times d_l} \quad (9)$$

These embeddings encode high-level clinical priors and are integrated into the multimodal prefix.

optimized using teacher forcing to predict subsequent tokens. The generation loss is defined as shown in Eq. (12).

$$L_{\text{gen}} = -\sum_{t=1}^T \log P(y_t | y < t, S_f) \quad (12)$$

-where $Y = \{y_1, y_2, \dots, y_T\}$ denotes the target report sequence, $y < t = \{y_1, \dots, y_{t-1}\}$, and T is the sequence length. The model generates reports auto-regressively conditioned on S_f during inference [16].

F. Language Model Decoder

The proposed framework uses a radiology-domain adapted LLaMA-2-7B model as the language decoder.

To integrate domain-specific knowledge, the base model is fine-tuned on a subset of the MIMIC-CXR dataset [38] derived from the official training split. This subset contains only radiology report text (Findings and Impression parts), and no images are used at this stage. Samples with missing text are removed from the dataset. The dataset is split into the training and validation sets in the ratio of 95:5. The resulting

in about 6 million trainable parameters. No samples from the IU dataset or test set are used at this stage. During multimodal training, base model parameters are frozen, and parameter-efficient tuning is performed using LoRA. Trainable components include LoRA layers, projection layers, disease-memory embeddings, and classification heads. The conditional probability of generating a report Y is defined as shown

Table 4: Quantitative comparison of experimental variants E1–E6 on the IU Chest X-ray test set using natural-language generation and semantic similarity metrics. Bold values indicate the best result for each metric.

Model variant	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	BERT Score-F1	BERT Score-P	BERT Score-R
Vision-only baseline (E1)	0.289	0.176	0.121	0.085	0.316	0.247	0.892	0.893	0.891
Auxiliary Disease supervision (E2)	0.226	0.145	0.102	0.073	0.340	0.249	0.900	0.923	0.880
Auxiliary Disease + Region supervision (E3)	0.256	0.159	0.110	0.077	0.313	0.246	0.896	0.912	0.880
Shallow alignment (beam search) (E4)	0.306	0.184	0.124	0.087	0.316	0.255	0.892	0.891	0.891
Learned Alignment with token gating (E5)	0.308	0.179	0.117	0.079	0.311	0.253	0.891	0.891	0.891
proposed Knowledge-Augmented Model (E6)	0.300	0.170	0.105	0.068	0.327	0.281	0.894	0.898	0.891

domain-adapted model is referred to as MedLLaMA.

For domain adaptation, the model is fine-tuned for 2 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 4 (batch size 1 with gradient accumulation). Fine-tuning is done using the parameter-efficient approach LoRA, where attention projection layers (q_proj , v_proj) are tuned with rank $r = 8$, scaling factor $\alpha = 16$, and dropout = 0.05, resulting

in Eq. (13).

$$P(Y|S_f) = \prod_{t=1}^T P(y_t|y_{<t}, S_f; \theta) \quad (13)$$

where $Y = \{y_1, y_2, \dots, y_T\}$ denotes the generated token sequence, T is the report length, $y_{<t} = (y_1, y_2, \dots, y_{t-1})$ denotes previously generated tokens and θ represents trainable parameters, including LoRA weights and auxiliary modules as shown in Eq. (14). During inference, beam search decoding with a beam

Table 5: Clinical performance of experimental variants E1–E6 on the IU Chest X-ray test set using CheXbert-based label extraction. Bold values indicate the best performance for each metric.

Model variant	Micro-F1	Macro-F1	Precision	Recall
Vision-only baseline (E1)	0.4091	0.0406	0.3969	0.4222
Auxiliary Disease supervision (E2)	0.4229	0.0560	0.4102	0.4363
Auxiliary Disease + Region supervision (E3)	0.4182	0.0419	0.4144	0.4222
Shallow alignment (beam search) (E4)	0.4055	0.0403	0.3942	0.4175
Learned Alignment with token gating (E5)	0.4091	0.0406	0.3969	0.4222
proposed Knowledge-Augmented Model (E6)	0.4211	0.0640	0.4272	0.4151

size of 2 is used for report generation for all experimental variants.

$$y_t = \underset{v \in V}{\operatorname{argmax}} P(y_t = v \mid y < t, S_f; \theta) \quad (14)$$

where V is the output vocabulary. Until the end-of-sequence token is generated, this autoregressive decoding mechanism continues [6], [16].

G. Loss Functions

The proposed framework is trained using a composite objective that jointly optimizes language generation, disease prediction, cross-modal alignment, and consistency between visual and textual representations [23], [24]. The total loss is defined as shown in Eq. (15).

$$L_{total} = L_{lm} + \lambda_1 L_{img-dis} + \lambda_2 L_{align} + \lambda_3 L_{rep-dis} + \lambda_4 L_{cons} \quad (15)$$

where L_{lm} is the language modeling loss, $L_{img-dis}$, and $L_{rep-dis}$ are disease classification losses for image and report modalities respectively. L_{align} enforces alignment between disease-memory and visual representations, and L_{cons} promotes consistency between image and report-based disease predictions. The coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}_+$ are scalar hyperparameters controlling the contribution of each

Table 6. Per-disease clinical performance of the proposed model on the IU Chest X-ray test set using CheXbert-based evaluation.

Disease	P	R	F1	Prev	FN%
Enlarged Cardiomeastinum	0.50	0.03	0.06	0.07	97
Cardiomegaly	0.75	0.12	0.21	0.11	88
Lung Opacity	1.00	0.02	0.04	0.12	98
Lung Lesion	0.00	0.00	0.00	0.04	100
Edema	0.00	0.00	0.00	0.01	100
Consolidation	0.00	0.00	0.00	0.00	100
Pneumonia	0.00	0.00	0.00	0.01	100
Atelectasis	0.00	0.00	0.00	0.06	100
Pneumothorax	0.00	0.00	0.00	0.00	100
Pleural Effusion	0.00	0.00	0.00	0.03	100
Pleural Other	0.00	0.00	0.00	0.01	100
Fracture	0.00	0.00	0.00	0.03	100
Support Devices	0.00	0.00	0.00	0.04	100
No Finding	0.43	0.94	0.59	0.40	6

Note: P = precision; R = recall; F1 = F1-score; Prev = prevalence; FN = false-negative rate.

Table 7. Representative qualitative examples comparing reports generated by the proposed framework with ground-truth reports from the IU Chest X-ray dataset.

UID	Case Type	Ground Truth Report	Generated Report	Observation
407	Normal	Lungs are clear. No pleural effusions or pneumothoraces. Heart size normal.	Heart size is normal. Lungs are clear. No pleural effusion or pneumothorax.	Highly consistent normal report
2528	Normal	Heart and mediastinum normal. Lungs clear.	Heart is normal in size. No focal airspace opacity. Lungs clear.	Clinically accurate normal case
945	Abnormal	Consolidation developed in left upper lobe.	Focal pulmonary opacity noted. No pleural effusion.	Partial grounding (missed location specificity)
1834	Abnormal	Mild cardiomegaly with pulmonary edema.	Cardiac silhouette mildly enlarged with vascular congestion.	Correct detection with appropriate attribute representation
1287	Abnormal	Small right pleural effusion.	Small pleural effusion is present.	Partial grounding (missing laterality)
2210	Challenging	Bilateral patchy infiltrates with left basilar atelectasis.	Mild bibasal opacity.	Partial grounding (incomplete severity and laterality)

loss component and $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x > 0\}$

The language modeling loss is defined using teacher-forcing as shown in Eq. (16) [16].

$$L_{lm} = -\sum_{t=1}^T \log P(y_t \mid y < t, S_f) \quad (16)$$

-where y_t denotes the token at time step t , $y < t = (y_1, y_2, \dots, y_{t-1})$ represents the preceding tokens, T is the sequence length, and $P(y_t \mid y < t, S_f)$ is the conditional probability of token y_t predicted by the decoder given prior tokens and the multimodal prefix S_f .

H. Disease & Alignment Losses

Disease prediction is formulated as a multi-label classification problem over M predefined disease categories [1]. The classification losses are defined using binary cross-entropy as shown in Eq. (17) and Eq. (18).

$$L_{img-dis} = -\sum_{i=1}^M [D_i^* \log \hat{D}_i^{img} + (1 - D_i^*) \log(1 - \hat{D}_i^{img})] \quad (17)$$

$$L_{rep_dis} = - \sum_{i=1}^M [D_i^* \log \hat{D}_i^{rep} + (1 - D_i^*) \log (1 - \hat{D}_i^{rep})] \quad (18)$$

-where, $\hat{D}_{img}, \hat{D}_{rep} \in [0,1]^M$ denote predicted disease probabilities from the image and report branches, respectively, $D^* \in \{0,1\}^M$ represent the ground-truth disease labels, and M is the number of disease categories. Image-side predictions are obtained from the auxiliary classifier operating on visual features.

$$L_{cons} = \frac{1}{M} \|\hat{D}_{img} - \hat{D}_{rep}\|_2^2 \quad (20)$$

which enforces agreement between visual and textual disease predictions and improves cross-modal coherence. All of these loss components are jointly optimized in an end-to-end manner. The weighting coefficients λ_1 to λ_4 were determined empirically based on validation performance and training stability. In our experiments, we use: $\lambda_1=0.20$, $\lambda_2=0.05$, $\lambda_3 = 0.10$, and $\lambda_4 = 0.05$. These values balance the contribution of each objective and ensure that no single loss

Table 8. Qualitative comparison of abnormality grounding across experimental variants for a representative abnormal chest X-ray case.

Model Variant	Generated Report	Abnormality Grounding Analysis
Ground Truth	Mild cardiomegaly with bilateral pulmonary vascular congestion and small pleural effusion.	Reference findings
Vision-only baseline	Heart size is normal. Lungs are clear. No pleural effusion or pneumothorax.	Missed all abnormalities; default normal report
Disease auxiliary supervision	Cardiac silhouette is mildly enlarged. No focal consolidation.	Correct cardiomegaly detected; missed congestion and effusion
Disease + region supervision	Mild cardiomegaly. Mild bibasal hazy opacity. No pneumothorax.	Captured cardiomegaly; partial pulmonary finding
Shallow alignment (beam search)	Cardiac silhouette mildly enlarged with vascular prominence.	Captured cardiomegaly and vascular congestion
Learned disease token	Mild enlargement of the cardiac silhouette. Small pleural effusion is present.	Correct cardiomegaly and effusion
Proposed disease-memory prefix	Mild cardiomegaly with pulmonary vascular congestion and small pleural effusion.	Most complete alignment with reference findings among compared methods

Report-side predictions are derived from a lightweight classifier applied to mean-pooled decoder hidden states. The ground-truth labels D^* are obtained using CheXbert, which extracts structured labels from radiology reports. These labels are pre-computed during data preprocessing. CheXbert provides contextual handling of negation and uncertainty, ensuring a unified and learned labeling strategy across datasets. The alignment loss is defined using cosine similarity between visual and disease-memory representations as shown in Eq. (19) [24], [27].

$$L_{align} = 1 - \frac{\bar{v}^T \bar{d}}{\|\bar{v}\|_2 \|\bar{d}\|_2} \quad (19)$$

where $\bar{v} = \text{GAP}(\hat{V})$ is the global visual representation obtained via global average pooling, and $\bar{d} = \text{Mean}(D_{mem})$ is the mean-pooled disease-memory embedding. D_{mem} , representing aggregated disease-memory priors selected via Top-K prediction. This loss encourages consistency between visual evidence and disease-level priors. The consistency loss is defined as the mean squared error between predicted disease probability distributions from the image-side and report-side classifiers as shown in Eq. (20).

dominates training while maintaining stable convergence. During training, report-side disease predictions are generated using the classifier on mean-pooling of the decoder hidden states using teacher forcing. This branch is used only for training supervision and does not run during inference.

IV. Experimental Setup

All experiments are conducted using multi-view chest X-rays and corresponding expert-written reports from the IU X-ray dataset. The dataset is split into training, validation and test subsets using a patient-independent split with a standard ratio of 70:15:15 to prevent data leakage across patients. The model is implemented in PyTorch using a single, 40 GB NVIDIA A100 GPU. The Swin-Transformer encoder is initialized with pre-trained weights. The MedLLaMA decoder remains frozen to preserve pre-trained knowledge, while parameter-efficient fine-tuning is achieved using LoRA. Trainable components include projection layers, disease-memory prefix embeddings, auxiliary classification heads, and the final stage of the Swin-Transformer. All reported results are generated using beam search decoding with beam size = 2 unless otherwise stated.

A. Pre-processing

Images are resized to (224 x 224) to match Swin-Transformer requirements and normalized according to ImageNet statistics. No substantial augmentation is performed to maintain clinically meaningful information. Radiology reports are tokenized via the LLaMA Tokenizer (SentencePiece). All samples are padded or truncated to a maximum of 80 tokens based on dataset statistics.

B. Training and Validation Protocol

Models are trained for a fixed number of epochs, with validation performed after each epoch. The checkpoint with the lowest validation loss is selected for testing. Early stopping is not used to maintain consistency across experiments. All experiments use a fixed random seed (42) across PyTorch, NumPy, and CUDA to ensure reproducibility. In E1 to E5, models are trained directly on the IU Chest X-ray dataset using image-report pairs, whereas E6 (proposed) uses a two-stage training strategy, i.e. MIMIC-CXR pre-training followed by IU fine-tuning.

C. Hyper-parameter Justification and Inference Settings

Important hyper-parameters are selected through a combination of existing literature and empirical considerations. The learning rate (1e-5) maintains optimization stability while fine-tuning a partially frozen architecture. A small batch size (4) with gradient accumulation achieves a stable effective batch size of 8 is used due to GPU memory constraints. Top-k disease prefix tokens (k=3) capture the most relevant disease semantics while avoiding noise from low-confidence predictions. The disease loss weight (0.2) ensures effective supervision for multi-label classification without dominating the language modeling objective and the alignment loss weight (0.05) is kept small to softly align image and report representations without destabilizing the training process. A beam size of 2 balances generation quality and computational efficiency, as larger beam sizes yield diminishing returns for short medical reports. A maximum length for generation of 80 tokens is selected based on statistics of the dataset and will prevent outputs that are too long and redundant. No-repeat n-gram size = 3, repetition penalty = 1.1 are used to optimize linguistic coherence and reduce redundancy in generated reports. Table 2 summarizes key model and training hyper-parameters, and the loss and inference settings.

D. Evaluation Metrics

The quality of generated reports is assessed using textual similarity and clinically grounded metrics. All evaluations are conducted on the findings sections of the IU X-ray dataset.

1. Textual Similarity Metrics:

We report commonly used lexical and semantic evaluation metrics for radiology report generation as used in prior works. BLEU-1 to BLEU-4 calculate n-gram precision at different lengths. ROUGE-L evaluates structural similarity by finding the longest common subsequence, METEOR finds semantic similarity using stemming and synonym matching, whereas CIDEr measures consensus using TF-IDF-weighted n-grams. In addition, contextual similarity is evaluated using BERTScore [40], which is presented in terms of Precision, Recall, and F1. BERTScore computes contextual embeddings unlike the n-gram based metrics. All metrics are computed using a single reference report per study. Before evaluation, both generated and reference texts are normalized (lowercasing and punctuation removal) prior to evaluation.

2. Clinical Efficacy Metrics:

Since textual metrics do not fully capture clinical correctness, we additionally evaluate disease-level performance using CheXbert. We report Micro-F1 (overall performance), Macro-F1 (performance on less frequent diseases), and Precision/Recall to analyze false positives and false negatives.

E. Ablation Study and Experimental Variants

To analyze the contribution of each component, we design a series of experimental variants by progressively enabling supervision signals, alignment mechanisms, and knowledge augmentation.

1. Notations Consistency with Section II

Let $V = f_{enc}(I_f, I_l) \in R^{B \times N_v \times d_v}$, denote visual features extracted by the Swin-Transformer encoder and fusion module. Let $Y = (y_1, \dots, y_T)$ represent the generated report. Let $D \in \{0, 1\}^M$ represent multi-label disease annotations, and $R \in \{0, 1\}^{M \times P}$ represent region-level supervision. Let $W_e \in R^{M \times d_l}$, represent the disease embedding (memory) table.

2. Token Definitions

Full visual tokens are given in Eq. (21) [8], [23].

$$Z_{full} = V \in R^{B \times N_v \times d_v} \quad (21)$$

Compressed Tokens (Patch aggregation/ gating), as shown in Eq. (22).

$$Z_{comp} = g(V) \quad (22)$$

where $g(\cdot)$ is a token compression function.

$Z_{comp} \in R^{B \times N_c \times d_v}$, with $N_c \ll N_v$

Disease-memory prefix tokens as shown in Eq. (23).

$$Z_{mem} = [e_i | i \in \text{TopK}(D, K)], \quad e_i = W_e[i] \quad (23)$$

3. Experimental Variants

We define the following variants:

a) Baseline and Supervision Variants

E1: Vision-only Baseline: This experiment uses only visual tokens, without auxiliary supervision. The model is trained using only the language modeling objective as shown in Eq. (24) [6], [16].

$$L = L_{lm}(Y|V) \quad (24)$$

where S_f corresponds to the decoder input derived from visual features without learned alignment, and L_{lm} is defined in Eq. (16).

E2: Auxiliary Disease Supervision: This experiment adds disease classification loss L_{img_dis} as shown in Eq. (25) [1], and encourages disease-aware visual encoding but does not inject disease tokens into the decoder. This evaluates whether supervision alone improves clinical grounding.

$$L = L_{lm} + \lambda_1 L_{img_dis} \quad (25)$$

where L_{img_dis} is defined in Eq. (17).

E3: Auxiliary Disease + Region Supervision: This variant extends E2 by incorporating region-level supervision into the visual encoder as shown in Eq. 26, improving spatial grounding [28], [29], [30].

$$L = L_{lm} + \lambda_1 L_{img_dis} + \lambda_r L_{region} \quad (26)$$

where L_{region} corresponds to region-level supervision and $\lambda_r \in \mathbb{R}^+$.

b) Alignment Mechanisms

E4: Shallow Alignment (Inference-time): This experiment introduces a shallow alignment mechanism during decoding without modifying model inputs. The scoring function incorporates disease-level alignment externally as shown in Eq. (27).

$$\log P(Y|I) = \log P(Y|S_f) + \alpha S_{align}(Y, D) \quad (27)$$

where $S_{align}(Y, D)$ is an alignment scoring function and $\alpha \in \mathbb{R}^+$. This does not involve the learned alignment module in Eq. (11).

E5: Learned Alignment with Token Gating: This variant introduces **learned cross-modal alignment** (Eq. 11) together with a **token gating mechanism** applied to visual features as in Eq. (28).

$$\tilde{F} = g(F') \quad (28)$$

where $g(\cdot)$ denotes a learnable gating function that selects salient tokens. The multimodal sequence becomes: $S = [\tilde{F}; E_T]$ and is processed via the learned cross-modal alignment module (Eq. (11)) to obtain S_f . The model is trained using the full composite loss (Eq. (15)), excluding knowledge augmentation.

c) Knowledge-Augmented Model

E6: Knowledge-Augmented Model (proposed): This experiment introduces external knowledge augmentation by combining a) disease-memory embeddings D_{mem} , b) learned alignment (Eq. (11)), c) token gating $g(\cdot)$, and d) knowledge augmentation via MIMIC-CXR pretraining. Region-level supervision is disabled in this setting. The multimodal prompt is constructed as $S = [D_{mem}; \tilde{V}g; P]$ as defined in Eq. (10) where $\tilde{V}g$ denotes projected (and gated) visual tokens,

D_{mem} represents disease-memory embeddings, and P corresponds to task-specific prompt embeddings. The full training objective is: $L = L_{total}$ where L_{total} is defined in Eq. (15). The aim of this experiment is to transfer radiology-domain knowledge from a larger radiology corpus to the target benchmark dataset, which is similar to prior transfer learning techniques. Because of computational limitations, disease-memory prompting and MIMIC-CXR pre-training are evaluated together in E6. Their combination represents an effective implementation scenario where structured priors and domain adaptation work together. All variants share the same backbone and training settings unless specified. Each experiment isolates a single factor (e.g., supervision, token compression, or conditioning) to ensure fair comparison. Component-wise breakdown of all experiment variants is provided in Table 3. The progression from E1 to E6 indicates the stepwise incorporation of disease supervision, region grounding, alignment techniques, and knowledge enhancement.

V. Results

A. Quantitative Results

All experimental variants (E1-E6) are evaluated on the IU Chest X-ray test set. Table 4 presents the quantitative comparison of all experimental variants in terms of BLEU, METEOR, ROUGE-L, and BERTScore (F1, Precision, and Recall). Best results are highlighted in bold. Table 5 presents clinical evaluation metrics like Micro-F1, Macro-F1, Precision and Recall for disease prediction from generated reports. Table 6 presents per disease performance of the proposed model. These metrics are computed using CheXbert-based label extraction.

B. Qualitative Analysis of Generated Reports

Qualitative analysis is done in order to evaluate the fluency, clinical relevance, and the abnormality grounding of the generated reports. Instead of a formal or statistical assessment of the model, this is meant to give illustrative information about the model behavior. A qualitative review of a selection of 50 cases which contain normal, single, and multiple abnormalities from IU Chest X-ray dataset has been done. This set of cases has been selected for illustration purpose and hence, is neither random nor representative. Thus, this form of evaluation is likely to be subjected to selection bias (Tables 7 & Table 8). To improve clarity, we define 'correct detection' as precise identification of the primary abnormality described in the ground-truth report, while 'partial grounding' refers to cases in which the model captures the general abnormality but omits specific attributes such as severity, anatomical location e.g., laterality, or coexisting findings. We acknowledge that this qualitative study does not have blinded review by multiple radiologists, nor does it consider

parameters such as accuracy, completeness, hallucinations, severity, laterality, or clinical usefulness on the basis of scoring protocols. Blinded review, especially with inter-rater reliability measures, would provide a more thorough examination of the clinical reliability and is an important area to explore in the future.

C. Comparison with State-of-the-Art (Indicative)

Table 9 presents an indicative comparison with representative prior approaches, including TMRGM [3], R2Gen [12], R2GenCMN [11], CDGPT2 [6], and R2GenGPT [16]. TMRGM utilizes template-based multi-attention, whereas R2Gen and R2GenCMN apply memory-based techniques for report generation and cross-modal interaction. CDGPT2 applies conditioned Transformer-based generation, and R2GenGPT adopts a frozen large language model decoder. Since the reported results were obtained under different datasets, splits, preprocessing, and training configurations, the comparison is intended for contextual analysis rather than direct performance ranking.

VI. Discussion

A. Quantitative and Semantic Performance Analysis

The differences between lexical, semantic, and clinical evaluation metrics are highlighted in Table 4 and Table 5. Small changes in BLEU ($\pm 0.01\text{--}0.02$) do not reflect meaningful improvements. On the other hand,

comparable lexical and semantic scores (e.g., BLEU-4 = 0.085, ROUGE-L = 0.316), as well as Micro-F1 of 0.4091, which indicates some alignment with clinical labels without any explicit supervision. However, when disease-level supervision (E2) is incorporated, semantic and clinical performance is improved, achieving the best BERTScore-F1 (0.900) and Micro-F1 (0.4229), suggesting better alignment with frequent disease patterns. The region-level supervision (E3) slightly reduces performance (ROUGE-L = 0.313, Micro-F1 = 0.4182), which means that signals at this level are noisy. The decoding with beam search (E4) does not contribute to better performance in the clinical setting (Micro-F1 = 0.4055), which confirms that lexical enhancement is not enough. The learned alignment with token gating (E5) does not show any significant difference from the baseline (Micro-F1 = 0.4091). The proposed model (E6) performs well semantically (METEOR score = 0.281, ROUGE-L score = 0.327) with a relatively similar but a little bit lower BERTScore-F1 (0.894) than E2. In clinical terms, E6 scores 0.4211 for Micro-F1, 0.0640 for Macro-F1 and 0.4272 for precision. Although there is an improvement in Macro-F1, the value of Macro-F1 is still very small, meaning that it is not very effective at identifying rare diseases.

Per-disease analysis (Table 6) demonstrates the concentration of performance in a few cases like Cardiomegaly (F1 = 0.21) and "No Finding" (F1 = 0.59); while other diseases demonstrate poor recall and high false-negative rates. This indicates that although the model produces clinical-sounding reports, it misses out on less common or co-occurring abnormalities. This is

Table 9. Indicative comparison of the proposed framework with existing radiology report-generation methods on the IU Chest X-ray dataset. Results from previous studies use different data splits, preprocessing, and evaluation protocols and should not be interpreted as direct performance rankings.

Method	BLEU 1	BLEU 4	ROUGE-L	METEOR	CIDEr	BERT Score-F1	BERT Score-P	BERT Score-R
TMRGM [3]	0.442	0.145	0.368	0.186				
R2Gen [12]	0.470	0.165	0.371	0.187				
R2GenCMN [11]	0.475	0.170	0.375	0.191				
CDGPT2 [6]	0.387	0.111	0.289	0.164	0.257			
R2GenGPT [16]	0.488	0.173	0.377	0.211	0.438			
R2GenMedLLa MA (Ours)	0.300	0.068	0.327	0.281	0.273	0.894	0.898	0.891

METEOR, ROUGE-L, and BERTScore are more suitable for semantic similarity analysis. Yet, an improvement in these semantic metrics does not always reflect clinical correctness since they measure linguistic similarity and not the ability to detect abnormalities. Therefore, we further use CheXbert-based clinical evaluation metrics to evaluate diagnostic relevance. The vision-only model (E1) provides

because the semantic scores depend on fluency and similarity, while the clinical scores depend on accurate detection of abnormalities. In effect, improvements in linguistic competence do not necessarily improve diagnostic ability. Although this study does not include statistical significance testing or multi-seed evaluation, the findings should be treated as initial empirical trends rather than statistical improvements. The consistency

of the direction of the changes for several lexical, semantic, and clinical metrics can serve as supporting evidence of the model's behavior, but it still cannot serve as proof of improvement. In particular, although the clinical scores demonstrate considerable correspondence to the disease labels in all the models considered, the differences observed between variants are relatively small and may fall within the range of random variation. Nevertheless, the obtained findings emphasize the gap between the lexical similarity and the clinical correctness of the generated reports.

B. Qualitative Analysis, Clinical Relevance and Failure Cases

Table 7 shows representative qualitative examples comparing reports generated by the proposed framework with ground-truth reports from the IU Chest X-ray dataset. Qualitative analysis shows that disease-aware conditioning enhances abnormality grounding compared to vision-only generation. While baseline models usually generate generic normal reports, disease-aware variants produce more clinically relevant findings, including “focal opacity”, “pleural effusion”, and “cardiomegaly”. The proposed method demonstrates improved alignment with reference reports in terms of content and specificity. Illustrative examples in Table 8 show that methods utilizing disease-specific signals show gradual improvements in abnormality grounding. However, several limitations still exist. The system is not able to handle minor abnormalities, laterality-specific abnormalities, and multi-condition cases. Occasionally, secondary findings are also omitted, and normal-style predictions are also observed in some abnormal cases. It is important to emphasize that this analysis is based on illustrative examples rather than a structured evaluation protocol and may be subject to selection bias. Therefore, it should not be interpreted as evidence of general clinical effectiveness. Future work will include comprehensive clinical evaluation across multiple dimensions, such as correctness, completeness, hallucination, laterality, severity, and clinical usefulness, ideally conducted by multiple blinded radiologists with inter-rater reliability analysis.

C. Comparison with State-of-the-Art Methods (Indicative)

The methods in Table 9 use different strategies for automated report generation. TMRGM [3] uses a template-based multi-attention approach, while R2Gen [12] introduces a memory-driven Transformer for exploiting recurring report patterns. R2GenCMN [11] further incorporates cross-modal memory to strengthen image-text interaction. CDGPT2 [6] uses conditioned Transformer-based generation with pretrained visual and language components, whereas R2GenGPT [16] employs a frozen LLM decoder to leverage pretrained language knowledge. In contrast, the proposed method

combines multi-view visual features, disease-memory conditioning, and parameter-efficient adaptation of a radiology-domain LLaMA-2 decoder.

There exists a trade-off between lexical similarity and semantic/clinical alignment compared to previous works, as shown in Table 9. The BLEU score of the proposed method is lower (BLEU-4 = 0.068), which indicates less dependency on exact phrase matching. In contrast, existing methods often report higher BLEU scores due to template-based generation strategies. Despite lower lexical scores, the proposed approach attains the highest METEOR (0.281) and strong BERTScore values ($P = 0.898$, $R = 0.891$, $F1 = 0.894$), as shown in Table 9, reflecting improved semantic similarity. These metrics capture paraphrasing ability and contextual alignment. However, it does not ensure the clinical accuracy of the generated reports. Moreover, the BERTScore metric is provided only for the proposed methodology, since this metric is not available for other works. The results suggest improved alignment of generated reports with extracted clinical attributes compared to baseline methods. This needs to be considered cautiously, as CheXbert-based evaluation serves as an automated proxy and cannot replace expert clinical assessment. The methodology of TMRGM is template-based multi-attention, R2Gen and R2GenCMN are memory-based encoder-decoder architectures, CDGPT2 is transformer-based generation, and R2GenGPT is frozen LLM-based. These differences also lead to different training strategies and computational requirements. Computationally, the methodology benefits from parameter-efficient fine-tuning of the frozen MedLLaMA model and can be trained using a single A100 GPU. It is significant to acknowledge the fact that these comparisons are indicative only because it is impossible to rank these methods directly due to discrepancies in the ways that datasets are split, processed, and evaluated in different research. Comparisons presented in Table 9 are representative as they include experiments that utilize different IU X-ray dataset splits, preprocessing, and evaluation protocols; hence, performance ranking is not appropriate. Therefore, the results highlight broader trends particularly the trade-off between lexical fidelity and semantic flexibility rather than definitive superiority.

D. Limitations

Many limitations still exist despite the improvements observed. It relies on automatically extracted CheXbert labels, which may introduce noise, and lacks expert radiologist evaluation. The model shows poor performance for subtle, laterality-specific, and multi-condition abnormalities, having a risk of false negatives and hallucinations. Also, the problem with the imbalanced dataset and the small size of the IU Chest X-ray dataset can influence generalization. Therefore,

the system is currently suitable only for research or preliminary report drafting and requires further clinical validation.

VII. Conclusion

A disease-memory guided framework for radiology report generation from multi-view chest X-rays is presented in this work. Disease-aware conditioning is integrated into a parameter-efficient large language model by adapting a frozen MedLLaMA backbone with lightweight trainable components. The experimental findings reveal better semantic coherence with METEOR (0.281), ROUGE-L (0.327), and BERTScore-F1 (0.894). Evaluation with CheXbert reveals fair performance (Micro-F1: 0.4211, Macro-F1: 0.0640), reflecting the difficulty of accurate label prediction. While improvements are observed, the proposed method does not outperform all baselines across all metrics. There is a trade-off between lexical overlap and semantic consistency. Qualitative examples demonstrate that the model can ground common diseases such as "cardiomegaly" and "pleural effusion" in some cases, although subtle findings, laterality-specific, and multiple-disease findings remain challenging. The system is not suitable for autonomous clinical reporting and is intended only for research or preliminary report generation. Some limitations remain, such as using automated labelling without verification by experts and testing on one particular dataset, which may limit generalizability. Future work will focus on improving abnormality grounding through disease reasoning and structured conditioning. Evaluation will be extended to larger multi-institutional datasets (e.g., MIMIC-CXR), along with expert-based clinical assessment.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets used in this study are publicly available from their respective repositories, subject to their applicable access and licensing conditions.

Author Contribution

Prajakta Mandar Dhamanskar: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Visualization, WritingOriginal Draft. Chintan Thacker: Supervision, Validation, WritingReview and Editing. Saurabh Shah: Supervision, Investigation, Project Administration, Writing Review and Editing. All authors reviewed and approved the final manuscript.

Declarations

Ethical Approval

This study involved secondary analysis of publicly available or credentialed de-identified datasets and did not involve direct interaction with human participants. According to the applicable institutional and dataset governance requirements, additional institutional ethical approval and individual informed consent were not required.

Consent for publication: Not applicable.

Competing Interests

The authors declare that they have no competing interests.

AI-assisted writing or generative AI use:

During preparation of this work, the authors used ChatGPT for language editing. The authors subsequently reviewed and edited the content and take full responsibility for the publication.

Code Availability:

The source code and trained model checkpoints are not publicly available as the work is part of an ongoing research project.

References

- [1] Z. He, A. N. N. Wong, and J. S. Yoo, 'Radiology report generation using automatic keyword adaptation, frequency-based multi-label classification and text-to-text large language models', *Computers in Biology and Medicine*, vol. 196, p. 110625, Sep. 2025, doi: 10.1016/j.combiomed.2025.110625.
- [2] F. Dong, S. Nie, M. Chen, F. Xu, and Q. Li, 'Keyword-based AI assistance in the generation of radiology reports: A pilot study', *npj Digit. Med.*, vol. 8, no. 1, p. 490, Aug. 2025, doi: 10.1038/s41746-025-01889-4.
- [3] X. Wang, Y. Zhang, Z. Guo, and J. Li, 'TMRGM: A Template-Based Multi-Attention Model for X-Ray Imaging Report Generation', *JoAIMS*, vol. 2, no. 1–2, pp. 21–32, 2021, doi: 10.2991/jaims.d.210428.002.
- [4] F. Yu et al., 'Evaluating progress in automatic chest X-ray radiology report generation', *Patterns*, vol. 4, no. 9, p. 100802, Sep. 2023, doi: 10.1016/j.patter.2023.100802.
- [5] K. Kale, P. Bhattacharyya, and K. Jadhav, 'Replace and Report: NLP Assisted Radiology Report Generation', in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 10731–10742. doi: 10.18653/v1/2023.findings-acl.683.

- [6] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, 'Automated radiology report generation using conditioned transformers', *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 2021, doi: 10.1016/j.imu.2021.100557.
- [7] D. Gao et al., 'Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning', *Medical Image Analysis*, vol. 91, p. 102982, Jan. 2024, doi: 10.1016/j.media.2023.102982.
- [8] Z. Liu et al., 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.
- [9] Y. Tang, Y. Yuan, F. Tao, and M. Tang, 'Cross-Modal Augmented Transformer for Automated Medical Report Generation', *IEEE J. Transl. Eng. Health Med.*, vol. 13, pp. 33–48, 2025, doi: 10.1109/JTEHM.2025.3536441.
- [10] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, 'Radiology report generation with a learned knowledge base and multi-modal alignment', *Medical Image Analysis*, vol. 86, p. 102798, May 2023, doi: 10.1016/j.media.2023.102798.
- [11] Z. Chen, Y. Shen, Y. Song, and X. Wan, 'Cross-modal Memory Networks for Radiology Report Generation', in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, 2021, pp. 5904–5914. doi: 10.18653/v1/2021.acl-long.459.
- [12] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, 'Generating Radiology Reports via Memory-driven Transformer', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 1439–1449. doi: 10.18653/v1/2020.emnlp-main.112.
- [13] M. Y. Ouis and M. A. Akhloufi, 'ChestBioX-Gen: contextual biomedical report generation from chest X-ray images using BioGPT and co-attention mechanism', *Front. Imaging*, vol. 3, p. 1373420, Apr. 2024, doi: 10.3389/fimag.2024.1373420.
- [14] J. Ji, Y. Hou, X. Chen, Y. Pan, and Y. Xiang, 'Vision-Language Model for Generating Textual Descriptions From Clinical Images: Model Development and Validation Study', *JMIR Form Res*, vol. 8, p. e32690, Feb. 2024, doi: 10.2196/32690.
- [15] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, 'Medical image captioning via generative pretrained transformers', *Sci Rep*, vol. 13, no. 1, p. 4171, Mar. 2023, doi: 10.1038/s41598-023-31223-5.
- [16] Z. Wang, L. Liu, L. Wang, and L. Zhou, 'R2GenGPT: Radiology Report Generation with frozen LLMs', *Meta-Radiology*, vol. 1, no. 3, p. 100033, Nov. 2023, doi: 10.1016/j.metrad.2023.100033.
- [17] M. Kapadnis, S. Patnaik, A. Nandy, S. Ray, P. Goyal, and D. Sheet, 'SERPENT-VLM: Self-Refining Radiology Report Generation Using Vision Language Models', in *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 283–291. doi: 10.18653/v1/2024.clinicalnlp-1.24.
- [18] C.-Y. Lin, 'ROUGE: A Package for Automatic Evaluation of Summaries', in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. Accessed: Sep. 01, 2026. [Online]. Available: <https://aclanthology.org/W04-1013/>
- [19] R. Vedantam, C. L. Zitnick, and D. Parikh, 'CIDEr: Consensus-based image description evaluation', in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 4566–4575. doi: 10.1109/CVPR.2015.7299087.
- [20] P. Singh and S. Singh, 'ChestX-Transcribe: a multimodal transformer for automated radiology report generation from chest x-rays', *Front. Digit. Health*, vol. 7, p. 1535168, Jan. 2025, doi: 10.3389/fdgth.2025.1535168.
- [21] A. Nicolson, J. Dowling, and B. Koopman, 'Improving chest X-ray report generation by leveraging warm starting', *Artificial Intelligence in Medicine*, vol. 144, p. 102633, Oct. 2023, doi: 10.1016/j.artmed.2023.102633.
- [22] Z. Babar, T. Van Laarhoven, and E. Marchiori, 'Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines', *PLoS ONE*, vol. 16, no. 11, p. e0259639, Nov. 2021, doi: 10.1371/journal.pone.0259639.
- [23] Y. Pan, L.-J. Liu, X.-B. Yang, W. Peng, and Q.-S. Huang, 'Chest radiology report generation based on cross-modal multi-scale feature fusion', *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 1, p. 100823, Mar. 2024, doi: 10.1016/j.jrras.2024.100823.
- [24] S. Zhang, C. Zhou, L. Chen, Z. Li, Y. Gao, and Y. Chen, 'Visual prior-based cross-modal

- alignment network for radiology report generation', *Computers in Biology and Medicine*, vol. 166, p. 107522, Nov. 2023, doi: 10.1016/j.combiomed.2023.107522.
- [25] J. Zhao et al., 'Automated Chest X-Ray Diagnosis Report Generation with Cross-Attention Mechanism', *Applied Sciences*, vol. 15, no. 1, p. 343, Jan. 2025, doi: 10.3390/app15010343.
- [26] H. Yin, S. Zhou, P. Wang, Z. Wu, and Y. Hao, 'KIA: Knowledge-Guided Implicit Vision-Language Alignment for Chest X-Ray Report Generation', in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds, Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 4096–4108. Accessed: Sep. 01, 2026. [Online]. Available: <https://aclanthology.org/2025.coling-main.276/>
- [27] D. Gu, Y. Gao, Y. Zhou, M. Zhou, and D. Metaxas, 'RadAlign: Advancing Radiology Report Generation with Vision-Language Concept Alignment', in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, vol. 15966, J. C. Gee, D. C. Alexander, J. Hong, J. E. Iglesias, C. H. Sudre, A. Venkataraman, P. Golland, J. H. Kim, and J. Park, Eds, Cham: Springer Nature Switzerland, 2026, pp. 484–494. doi: 10.1007/978-3-032-04981-0_46.
- [28] T. Gu, D. Liu, Z. Li, and W. Cai, 'Complex Organ Mask Guided Radiology Report Generation', in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2024, pp. 7980–7989. doi: 10.1109/WACV57701.2024.00781.
- [29] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, 'Interactive and Explainable Region-guided Radiology Report Generation', in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 7433–7442. doi: 10.1109/CVPR52729.2023.00718.
- [30] H. Sharma et al., 'MAIRA-Seg: Enhancing Radiology Report Generation with Segmentation-Aware Multimodal Large Language Models', in *Proceedings of the 4th Machine Learning for Health Symposium*, PMLR, Feb. 2025, pp. 941–960. Accessed: Sep. 01, 2026. [Online]. Available: <https://proceedings.mlr.press/v259/sharma25a.html>
- [31] S. Lee, J. Youn, H. Kim, M. Kim, and S. H. Yoon, 'CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images', *Eur Radiol*, vol. 35, no. 7, pp. 4374–4386, Jan. 2025, doi: 10.1007/s00330-024-11339-6.
- [32] J. S. Ryu, H. Kang, Y. Chu, and S. Yang, 'Vision-language foundation models for medical imaging: a review of current practices and innovations', *Biomed. Eng. Lett.*, vol. 15, no. 5, pp. 809–830, Sep. 2025, doi: 10.1007/s13534-025-00484-6.
- [33] I. Hartsock and G. Rasool, 'Vision-language models for medical report generation and visual question answering: a review', *Front. Artif. Intell.*, vol. 7, p. 1430984, Nov. 2024, doi: 10.3389/frai.2024.1430984.
- [34] R. Tanno et al., 'Collaboration between clinicians and vision-language models in radiology report generation', *Nat Med*, vol. 31, no. 2, pp. 599–608, Feb. 2025, doi: 10.1038/s41591-024-03302-1.
- [35] J. Xu, 'Uncertainty Estimation in Large Vision Language Models for Automated Radiology Report Generation', in *Proceedings of the 4th Machine Learning for Health Symposium*, PMLR, Feb. 2025, pp. 1039–1052. Accessed: Sep. 01, 2026. [Online]. Available: <https://proceedings.mlr.press/v259/xu25a.html>
- [36] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, 'Contrastive Learning of Medical Visual Representations from Paired Images and Text', 2020, *Proc. Mach. Learn. Res.*, vol. 182, pp. 2–25. doi: 10.48550/ARXIV.2010.00747.
- [37] D. Demner-Fushman et al., 'Preparing a collection of radiology examinations for distribution and retrieval', *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, Mar. 2016, doi: 10.1093/jamia/ocv080.
- [38] A. E. W. Johnson et al., 'MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports', *Sci Data*, vol. 6, no. 1, p. 317, Dec. 2019, doi: 10.1038/s41597-019-0322-0.
- [39] W. Li, X. Wang, W. Li, and B. Jin, 'A Survey of Automatic Prompt Engineering: An Optimization Perspective', 2025, *arXiv*. doi: 10.48550/ARXIV.2502.11560.
- [40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, 'BERTScore: Evaluating Text Generation with BERT', in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. doi: 10.48550/ARXIV.1904.09675.

Author Biography



Prajakta Dhamanskar received an M.E. degree in information technology from V.E.S.I.T. Chembur, Mumbai, Maharashtra, India and is pursuing a Ph.D. in computer engineering from Parul University, Vadodara, Gujarat.

She has 12+ years of experience in Academia. Currently, she serves as an assistant professor in the Department of Computer Engineering at Fr. Conceicao Rodrigues College of Engineering, Bandra, Mumbai, Maharashtra, India. Her research interests are in the areas of machine learning, deep learning, computer vision, design and algorithms. She had guided many major and minor projects of undergraduate students. She is eager to develop digital study materials to provide her students with easy access during their studies.



Dr. Chintan Thacker received his Ph.D. in Artificial Intelligence & Computer Vision from Gujarat Technological University in 2021. He previously served as the Head of the Department of Computer Science and Engineering at HJD Institute of Technical Education and Research,

Kera, India. He has over 15 years of experience in academia and more than 1 year of experience in the industry. Currently, he is an Associate Professor in the Department of Computer Science and Engineering at Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujarat. He has mentored several Ph.D. students & received grants from SSIP & GUJCOST. He regularly conducts workshops and leads projects in the field of computer vision and AI.



Prof. Saurabh Shah is a Professor of Computer Science & Engineering and a passionate Academician with more than 24 Years of experience in Academics, Administration and Research. He is presently working as a Professor, Director

(Talent and Career Development) and Director (Faculty Augmentation) at Parul University, Gujarat, India. His major Research areas are Medical Image Analysis, Computer Vision, Data Analytics, Machine Learning and Pattern Recognition. He has supervised 08 PhD candidates who have been awarded PhD under his supervision.