www.ijarp.com ISSN 2456-9992 Page: 01-18

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 04

GENERATIVE AI FOR MEDICAL IMAGE REPORT GENERATOR

[*1]Md. Faiz Jamal, [2]Minhaj Hasan, [3]Mohd.Ehtesham Khan, [4]Dr. Shaba Irram Mam

[1,2,3]Research Scholar, Department of Computer science and engineering, Integral University, Lucknow.

[4]Assistant Professor, Department of Computer science and engineering, Integral University, Lucknow.

Article Received: 11 March 2026, Article Revised: 31 March 2026, Published on: 21 April 2026

***Corresponding Author: Md. Faiz Jamal**

Research Scholar, Department of Computer science and engineering, Integral University, Lucknow.

DOI: https://doi-doi.org/101555/ijarp.3110

ABSTRACT

The rapid advancement of generative artificial intelligence has revolutionized the field of medical imaging, particularly in the domain of automated radiology report generation. With radiologists facing unprecedented workloads due to the exponential growth in medical imaging volumes, generative AI systems offer a promising solution to alleviate clinical burden, reduce diagnostic variability, and improve patient care. This research paper presents a comprehensive analysis of state-of-the-art generative AI approaches for medical image report generation, examining the technological foundations, architectural frameworks, evaluation methodologies, and clinical translation challenges. The paper systematically reviews deep learning architectures including Transformer-based models, vision-language models (VLMs), contrastive learning frameworks, and large language models (LLMs) that have been applied to generate diagnostic narratives from medical images across multiple modalities chest X-ray, CT, MRI, and ultrasound. Through a detailed examination of recent advances including disease-aware dual-stage frameworks, retrieval-augmented generation, category-wise contrastive decoding, and generalist foundation models, this study provides quantitative performance comparisons across benchmark datasets (MIMIC-CXR, IU-Xray, CheXpert) using standardized evaluation metrics including BLEU, ROUGE, CHEXBERT, RadGraph F1, and GREEN. The paper concludes with a proposed clinical validation framework and offers recommendations for future research directions in multimodal integration, hallucination mitigation, and real-world deployment.

KEYWORDS: Generative AI, Medical Image Report Generation, Radiology Report Generation, Vision-Language Models, Large Language Models, Transformer Architectures, Multimodal Learning, Contrastive Learning, Clinical NLP, Medical Imaging

1. INTRODUCTION

1.1 The Growing Burden on Radiology Services

Medical imaging has become indispensable to modern healthcare, serving as a cornerstone for diagnosis, treatment planning, and disease monitoring across virtually all medical specialties. Chest radiography alone remains the most frequently performed diagnostic procedure worldwide, valued for its low cost, minimal radiation exposure, and ability to provide substantial clinical information. However, the exponential growth in medical imaging volumes has placed unprecedented strain on radiology services globally. The demand for interpreting chest X-rays has outpaced the supply of radiologists, leaving many clinicians overworked and vulnerable to fatigue-related diagnostic errors.

The challenge is multidimensional: radiologists must not only detect and characterize abnormalities across increasingly complex imaging studies but also produce comprehensive, structured, and clinically accurate narrative reports that communicate findings to referring physicians. This report generation process is cognitively demanding and time-consuming, often requiring several minutes per study even for experienced specialists. The growing volume-pressure has led to longer turnaround times, potential declines in reporting consistency, and increased burnout among radiology professionals.

1.2 The Promise of Generative AI for Automated Reporting

Generative artificial intelligence offers a transformative solution to these challenges. Automatic radiology report generation can alleviate the workload for physicians, minimize regional disparities in medical resources, and improve diagnostic consistency across clinical settings. By leveraging deep learning to mimic the cognitive processes of radiologists extracting information from medical images, integrating clinical context and medical knowledge, and producing comprehensive diagnostic narratives generative AI systems have the potential to augment rather than replace human expertise.

The impact of such automation extends beyond mere efficiency. In a landmark evaluation of the generalist foundation model MedVersa, radiologists judged AI-generated and human-written chest radiograph reports to be clinically equivalent in 64% of all cases, with equivalence rising to 91% for normal studies. User studies demonstrated notable reductions in report writing time and clinically relevant discrepancies compared with standard reporting

workflows. These findings suggest that generative AI can support radiologist efficiency while maintaining or even improving diagnostic accuracy.

1.3 Research Objectives and Scope

This paper addresses the following research objectives: (1) to systematically analyze the core generative AI technologies enabling medical image report generation; (2) to present quantitative performance data from state-of-the-art systems evaluated on benchmark datasets; (3) to provide a mathematical framework for understanding multimodal learning and report generation; (4) to propose a comprehensive evaluation taxonomy encompassing both natural language and clinical accuracy metrics; (5) to examine clinical validation studies and real-world deployment considerations; and (6) to identify future research directions in multimodal integration, hallucination mitigation, and generalist foundation models.

The scope encompasses multiple imaging modalities including chest X-ray, CT, MRI, breast ultrasound, and orbital MRI, with particular emphasis on chest X-ray report generation as the most extensively studied domain. The paper addresses both the technical architecture of generative models and the clinical, regulatory, and ethical considerations surrounding their deployment.

2. Core Technologies and Architectural Foundations

2.1 Overview of Generative AI Approaches for Medical Report Generation

Generative AI for medical image report generation sits at the intersection of computer vision and natural language processing. The task requires a model to accept medical images as input and produce coherent, clinically accurate textual descriptions of observed findings and their diagnostic implications. Recent surveys have proposed a general workflow comprising five main components: multi-modality data acquisition, data preparation, feature learning, feature fusion and interaction, and report generation.

The evolution of approaches can be traced through several generations of architectural innovation. Early systems adapted convolutional–recurrent (CNN–RNN) architectures originally developed for natural image captioning. While these models demonstrated feasibility, they struggled to capture long-range dependencies and complex clinical semantics. The introduction of visual attention mechanisms enabled models to focus on fine-grained image regions associated with pathological findings, representing a significant advance.

The emergence of Transformer-based architectures marked a paradigm shift, enabling better modeling of global contextual relationships and significantly enhancing report coherence and

clinical relevance. More recently, vision-language models (VLMs) and large language models (LLMs) have positioned multimodal foundation models at the forefront of automated radiology report generation.

Table 1: Evolution of Generative AI Architectures for Medical Report Generation.

Generation	Architecture	Key Innovation	Limitations	Representative Work
First Generation	CNN-RNN	Image-to-text mapping	Poor long-range dependency modeling	Encoder-decoder baselines
Second Generation	CNN-Attention-RNN	Visual attention mechanisms	Limited clinical semantics	R2Gen, R2GenCMN
Third Generation	Transformer-based	Global context modeling	Computational intensity	M2-Transformer, RGRG
Fourth Generation	Vision-Language Models (VLMs)	Cross-modal alignment	Hallucination risk	RadAlign, OAE, CLALA-Net
Fifth Generation	Multimodal LLMs + Foundation Models	Unified medical AI	Deployment complexity	MedVersa, LLaVA-Rad, Maira-2

2.2 Vision-Language Models for Report Generation

Vision-language models represent the current state-of-the-art for medical report generation. Unlike earlier approaches that treated visual feature extraction and language generation as separate stages, VLMs learn joint representations of images and text in a shared embedding space. This enables the model to align visual features with medical diagnostic criteria, introducing core knowledge supervision and creating interpretable intermediate diagnosis results.

The RadAlign framework exemplifies this paradigm. By aligning visual features with medical diagnostic criteria in a shared representation space, RadAlign achieves superior disease classification on MIMIC-CXR (average AUC: 0.885) and enables accurate report generation (GREEN score: 0.678 vs. state-of-the-art 0.634). Critically, RadAlign also demonstrates exceptional generalization capabilities, outperforming state-of-the-art foundation and specialized models on the external OpenI dataset (AUC: 0.923 vs. 0.836).

Cross-Modal Alignment Objective

The core training objective for vision-language alignment can be formulated as a contrastive loss:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(v_i, t_j)/\tau)}$$

Where:

- v_i = visual feature vector for image i
- t_i = text feature vector for report i
- $\text{sim}(v, t) = \frac{v \cdot t}{\|v\|\|t\|}$ = cosine similarity
- τ = temperature parameter controlling distribution sharpness
- N = batch size

2.3 Retrieval-Augmented Generation

A significant advancement in medical report generation is the incorporation of retrieval-augmented mechanisms. The Observe, Align, and Enhancement (OAE) framework proposes a hierarchical retrieval-enhanced approach comprising three stages: Observe, which leverages retrieval techniques to enhance visual feature comprehension by identifying similar images and associated reports; Align, where retrieved contextual reports guide the generation of an initial diagnostic report ensuring semantic consistency; and Enhance, an iterative refinement process that incorporates additional textual information to improve semantic coherence and diagnostic precision.

Retrieval augmentation addresses a fundamental limitation of purely generative models: the difficulty of producing thorough diagnostic narratives without reference to prior cases. By retrieving relevant exemplars based on visual and disease-specific similarities, models can generate reports that are not only linguistically fluent but also clinically grounded.

Retrieval-Augmented Generation

The probability of generating a report R given an image I and a set of retrieved exemplars $E = \{(\hat{I}_k, \hat{R}_k)\}_{k=1}^{K}$ can be expressed as:

$$P(R|I, E) = \prod_{t=1}^{T} P(r_t|r_{<t}, I, E)$$

Where the retrieval-conditioned probability is modeled as:

$$P(r_t|r_{<t}, I, E) = \text{softmax}(W \cdot \text{Transformer}(h_t, [\text{enc}(I); \text{enc}(E)]))$$

Here, h_t represents the hidden state at decoding step t, and the encoder processes both the query image and retrieved exemplars.

2.4 Disease-Aware and Lesion-Aware Architectures

A critical insight in medical report generation is that radiologists interpret images by integrating both global patterns and regional lesion cues. The Contrastive Lesion Attention and LLM Aggregation Network (CLALA-Net) embodies this insight through three key modules: Lesion Cross-Attention (LCA), which injects lesion-level cues derived from full-image classification into each region of interest; Lesion-Level Contrastive Learning (LLCL), which enhances discriminability by aligning lesion representations across chest X-rays; and Image-Text Contrastive Learning (ITCL), which improves visual-textual semantic alignment.

Similarly, the Disease-Aware Dual-Stage Framework (Wu et al., 2026) introduces Disease-Aware Semantic Tokens (DASTs) corresponding to specific pathology categories, learned through cross-attention mechanisms and multi-label classification. Stage 2 incorporates a Disease-Visual Attention Fusion (DVAF) module and a Dual-Modal Similarity Retrieval (DMSR) mechanism that combines visual and disease-specific similarities to retrieve relevant exemplars.

These disease-aware architectures address the fundamental limitation of conventional approaches, which often lack sufficient disease-awareness in visual representations and struggle to generate clinically accurate reports.

3. Major Model Architectures and Benchmark Performance

3.1 Transformer-Based Architectures

Transformers have become the dominant architecture for medical report generation due to their ability to model long-range dependencies and capture global contextual relationships. The multimodal Transformer-based framework for breast ultrasound report generation employs a Vision Transformer (ViT) to extract rich spatial and morphological features from ultrasound images, while pretrained language models (BERT, BioBERT, and GPT-2) handle textual embedding in various encoder–decoder configurations.

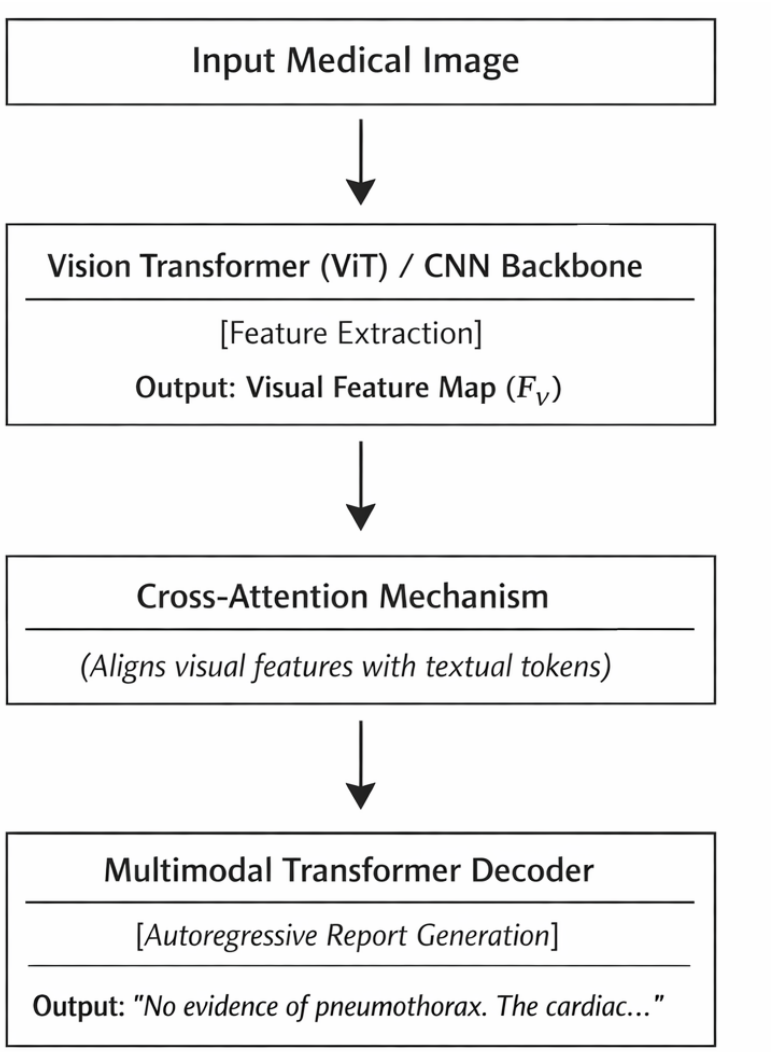

Figure 1: Transformer-Based Medical Report Generation Architecture.

The transformer decoder employs beam search to select the optimal token sequence. Beam search is applied to select the best token at each timestep, which is then fed back into the decoder for the next timestep, enabling the generation of coherent, clinically relevant reports.

3.2 Large Language Model Integration

The integration of large language models has substantially advanced the capabilities of medical report generation systems. The CLALA-Net framework utilizes an LLM-based aggregator to consolidate region-of-interest (ROI)-level descriptions into a clinically coherent report. An LLM-driven label extraction pipeline generates fine-grained lesion annotations for training and evaluation, addressing the challenge of limited annotated data in medical imaging.

The use of LLMs as optimizable orchestrators is exemplified by MedVersa, which employs an LLM to unlock generalist capabilities for processing multimodal inputs and outputs across diverse tasks. MedVersa is trained on tens of millions of compiled medical instances from diverse imaging scenarios and can accept heterogeneous inputs including images and clinical text, generating outputs such as classifications, segmentations, and narrative reports.

3.3 Contrastive Decoding for Hallucination Mitigation

A persistent challenge in medical report generation is the tendency of models to produce "hallucinations" clinically incorrect statements that are not supported by the underlying image. Current multimodal large language models generate full radiology reports in a single forward pass. As the report grows, the model increasingly relies on its own output rather than the image, a pattern that can produce clinically incorrect pairings.

The Category-Wise Contrastive Decoding (CWCD) framework addresses this limitation by generating category-wise reports through contrasting normal X-rays with masked X-rays using category-specific visual prompts. This decoding strategy maintains attention to visual tokens throughout generation, reducing reliance on language priors that introduce spurious pathology co-occurrences. Experimental results demonstrate that CWCD consistently outperforms baseline methods across both clinical efficacy and natural language generation metrics.

Equation 3: Contrastive Decoding Formulation

The contrastive decoding objective can be formulated as:

$$P_{\text{CWCD}}(y_t|y_{<t}, I) = \text{softmax}(\log P_{\text{LLM}}(y_t|y_{<t}, I) - \alpha \cdot \log P_{\text{LLM}}(y_t|y_{<t}, I_{\text{masked}}))$$

Where:

- I = original medical image
- I_{masked} = masked version of the image
- α = contrastive strength hyperparameter
- The subtraction amplifies tokens that are truly grounded in the image

3.4 Generalist Foundation Models

The emergence of generalist foundation models represents a paradigm shift in medical AI. Unlike task-specific systems that can only perform narrow functions, generalist models can learn from multimodal inputs and produce flexible outputs across imaging workflows. MedVersa demonstrates that a single, flexible, multimodal foundation model can match or

surpass task-specialized medical imaging systems while supporting radiologist efficiency and consistency.

The UniX model further advances this paradigm by unifying autoregression and diffusion within a single architecture an autoregressive branch for understanding and a diffusion branch for high-fidelity generation. This unified approach addresses the limitation of existing models that cannot simultaneously handle both image understanding and generation tasks.

Table 2: Comparative Performance of State-of-the-Art Models.

Model	Architecture	Dataset	Key Metric	Performance
RadAlign	VLM + Cross-modal retrieval	MIMIC-CXR	Average AUC	0.885
RadAlign	VLM + Cross-modal retrieval	MIMIC-CXR	GREEN Score	0.678 (SOTA: 0.634)
RadAlign	VLM + Cross-modal retrieval	OpenI (external)	AUC	0.923 (SOTA: 0.836)
CLALA-Net	LLM + Contrastive learning	Chest-Imagenome	Lesion F1-score	0.40
CLALA-Net	LLM + Contrastive learning	Chest-Imagenome	Total Score (20 max)	14.32
Disease-Aware Framework	Dual-stage + DAST + DVAF	CheXpert Plus, IU X-ray, MIMIC-CXR	Clinical accuracy & linguistic quality	State-of-the-art
MedVersa	Generalist foundation model	Multi-task (9 tasks)	Clinical equivalence	64% (91% for normal studies)

4. Benchmark Datasets and Evaluation Metrics

4.1 Public Datasets for Medical Report Generation

The development of deep learning-based radiology report generation has been significantly enabled by the availability of large-scale, publicly accessible datasets. Three datasets have emerged as benchmarks for the field: MIMIC-CXR, IU-Xray, and CheXpert.

MIMIC-CXR is a de-identified publicly available database of chest radiographs with free-text reports, representing one of the largest collections of radiology image-text pairs. The dataset contains over 377,000 images corresponding to 227,000 radiographic studies, each accompanied by a structured free-text radiology report. The scale of MIMIC-CXR has enabled training of large foundation models capable of learning rich visual-textual representations.

IU-Xray (Indiana University Chest X-ray Collection) provides a smaller but well-annotated dataset of chest X-rays with corresponding reports. Despite its smaller size compared to MIMIC-CXR, IU-Xray remains a valuable benchmark for evaluating model generalization and performance on diverse imaging protocols.

CheXpert and **CheXpert Plus** offer labeled datasets with 14 common chest radiographic observations, enabling supervised learning of disease classification alongside report generation. CheXpert Plus extends the original dataset with additional annotations and report pairs.

Table 3: Summary of Public Radiology Report Generation Datasets.

Dataset	Modality	Number of Images	Number of Reports	Annotation Type	Primary Use
MIMIC-CXR	Chest X-ray	377,000+	227,000+	Free-text reports	Large-scale training
IU-Xray	Chest X-ray	7,470	3,955	Free-text reports	Benchmark evaluation
CheXpert	Chest X-ray	224,316		14-label classification	Disease classification
CheXpert Plus	Chest X-ray	243,000+	243,000+	14 labels + reports	Combined training
Chest-Imagenome	Chest X-ray	242,072		Anatomical location + labels	Lesion-level analysis
OpenI (Indiana)	Chest X-ray	3,216	3,216	Free-text reports	External validation

4.2 Evaluation Metrics: From Lexical to Clinical

The evaluation of medical report generation systems requires a multifaceted approach that captures both linguistic quality and clinical accuracy. Traditional natural language generation metrics such as BLEU and ROUGE measure lexical overlap between generated and reference reports, but these metrics often fail to detect clinically significant errors such as incorrect negation or anatomical misclassification.

RadEval, a unified open-source framework for evaluating radiology texts, consolidates a diverse range of metrics from classic n-gram overlap (BLEU, ROUGE) and contextual measures (BERTScore) to clinical concept-based scores (F1CheXbert, F1RadGraph, RaTEScore, SRR-BERT, TemporalEntityF1) and advanced LLM-based evaluators (GREEN).

Table 4: Taxonomy of Evaluation Metrics for Medical Report Generation.

Metric Category	Specific Metrics	What It Measures	Clinical Relevance
Lexical Overlap	BLEU-1 to BLEU-4, ROUGE-L	Word/phrase overlap with reference	Low (can miss clinical errors)
Semantic Similarity	BERTScore	Contextual embedding similarity	Moderate
Clinical Entity Recognition	F1CheXbert, F1RadGraph	Detection of clinical concepts (findings, anatomy, uncertainty)	High
Temporal Consistency	TemporalEntityF1	Correct ordering of temporal events	High
LLM-based Evaluation	GREEN, G-Eval	Holistic clinical quality assessment	High
Radiologist Judgment	Expert review	Gold standard clinical equivalence	Highest

In a comparative assessment of ChatGPT-4 performance in orbital MRI reporting, evaluation included established NLP metrics (BLEU-4, ROUGE-L, BERTScore), clinical content recognition scores (RadGraph F1, CheXbert), and expert human judgment. Among automated metrics, BERTScore demonstrated the highest language similarity, while RadGraph F1 best captured clinical entity recognition.

Equation 4: RadGraph F1 Calculation

$$\text{RadGraph F1} = \frac{2 \cdot \text{Precision}_{\text{entities}} \cdot \text{Recall}_{\text{entities}}}{\text{Precision}_{\text{entities}} + \text{Recall}_{\text{entities}}}$$

Where entities include clinical findings, anatomical locations, and uncertainty modifiers extracted from both generated and reference reports using the RadGraph information extraction system.

4.3 Clinical Validation and Human Evaluation

Automated metrics, while valuable for model development and benchmarking, cannot substitute for clinical validation. The ultimate test of a medical report generation system is whether radiologists judge its outputs to be clinically equivalent to human-written reports.

The MedVersa study represents the most rigorous clinical validation to date. In blinded evaluations of chest radiograph reports, radiologists judged AI-generated and human-written reports to be clinically equivalent in 64% of all cases, with equivalence rising to 91% for normal studies. User studies demonstrated notable reductions in report writing time and clinically relevant discrepancies compared with standard reporting workflows.

In the ChatGPT-4 orbital MRI study, clinician assessment revealed moderate agreement with LLM outputs, with performance decreasing in complex or infiltrative cases. The study highlighted both the promise and current limitations of LLMs in radiology, particularly regarding their inability to process volumetric data and maintain spatial consistency.

5. Clinical Applications Across Imaging Modalities

5.1 Chest X-Ray Report Generation

Chest X-ray report generation remains the most extensively studied application of generative AI in medical imaging, due to the modality's prevalence, standardization, and the availability of large-scale datasets. Recent advances have focused on addressing the fundamental challenges of low-contrast images, subtle pathologies, and the requirement to generate long, unconstrained textual reports.

The Disease-Aware Dual-Stage Framework achieves state-of-the-art performance through its two-stage design. In Stage 1, the model learns Disease-Aware Semantic Tokens (DASTs) corresponding to specific pathology categories through cross-attention mechanisms and multi-label classification, while simultaneously aligning vision and language representations via contrastive learning. In Stage 2, the Disease-Visual Attention Fusion (DVAF) module integrates disease-aware representations with visual features, and the Dual-Modal Similarity Retrieval (DMSR) mechanism combines visual and disease-specific similarities to retrieve relevant exemplars.

5.2 CT and MRI Report Generation

Beyond chest X-ray, generative AI has been applied to more complex 3D imaging modalities. The MS-VLM (Multi-Slice Vision-Language Model) mimics radiologists' approach to 3D medical images by examining individual slices sequentially and synthesizing information across slices and views. In both slice-level and 3D scenario evaluations, MS-VLM surpasses existing methods in radiology report generation, producing more coherent and clinically relevant reports.

For emergency head CT reporting, a self-attentive deep fusion framework has been developed that incorporates clinically validated preprocessing for head CTs, a multimodal neural network refined through fine-grained regularization, and intra-sequence self-attention to enhance context modeling.

The Glio-LLaMA-Vision model specifically addresses adult-type diffuse gliomas, integrating molecular status prediction with radiology report generation. This demonstrates the potential

of generative AI to combine diagnostic classification with narrative report generation for specialized oncological imaging.

5.3 Breast Ultrasound Report Generation

Breast ultrasound imaging presents unique challenges due to the variability in image acquisition, the subtlety of morphological features, and the need for precise anatomical localization. A multimodal Transformer-based framework for automatic breast ultrasound report generation has been developed, employing a Vision Transformer (ViT) to extract rich spatial and morphological features. For textual embedding, pretrained language models (BERT, BioBERT, and GPT-2) are implemented in various encoder–decoder configurations. Experimental results demonstrate that BioBERT-based models consistently outperform general domain counterparts in clinical specificity, while GPT-2-based decoders improve linguistic fluency.

6. CHALLENGES AND LIMITATIONS

6.1 Hallucination and Clinical Factual Accuracy

The most significant challenge facing generative AI for medical report generation is the phenomenon of hallucination generation of clinically incorrect statements not supported by the underlying image. Token-based metrics (BLEU, ROUGE), semantic metrics (BERTScore), and rule-based metrics (CheXpert) often fail to detect clinically significant errors, such as incorrect negation or anatomical misclassification.

The root cause of hallucination lies in the autoregressive decoding strategy used by most models. As the report grows, the model increasingly relies on its own output rather than the image, a pattern that can produce clinically incorrect pairings. The CWCD framework addresses this through category-wise contrastive decoding, but complete mitigation of hallucination remains an open challenge.

6.2 Data Scarcity and Domain Shift

Despite the availability of large datasets like MIMIC-CXR, data scarcity remains a limitation for specialized imaging modalities and rare diseases. The lack of diverse, annotated datasets limits the generalizability of existing models across different imaging protocols, patient populations, and healthcare settings. Domain shift the degradation in model performance when applied to data from a different distribution than the training set remains a significant barrier to clinical deployment.

6.3 Multi-Modality Integration

Current models typically operate on a single imaging modality, yet clinical diagnosis often requires integration of multiple imaging studies, laboratory results, and clinical history. The Argus benchmark for 3D radiology report generation and the DiA-gnostic VLVAE framework for handling missing modalities represent early steps toward true multi-modality integration, but substantial work remains.

6.4 Uncertainty Modeling

Radiology reports inherently contain uncertainty findings may be described as "possible," "likely," or "cannot exclude" and the reasoning process for reaching clinical impressions is seldom explicitly modeled in current systems. The CURV (Coherent Uncertainty-Aware Reasoning in Vision-Language Models) framework addresses this gap by explicitly modeling uncertainty in findings and the reasoning process for reaching clinical impressions. However, uncertainty modeling in generative medical AI remains an underdeveloped area.

6.5 Computational Constraints and Deployment

The computational demands of large foundation models pose challenges for real-world deployment, particularly in resource-constrained healthcare settings. Lightweight models, such as the Lightweight Hybrid Foundation Model for lung cancer prognosis based on low-dose chest X-ray images, represent efforts to reduce computational intensity while maintaining clinical utility. Nevertheless, the balance between model capacity and deployability remains a key consideration.

7. Future Directions

7.1 Multimodal Integration with Clinical Data

Future systems must integrate medical images with broader clinical context electronic health records, laboratory results, prior imaging studies, and clinical notes to generate truly comprehensive diagnostic reports. The retrieval-augmented mechanisms in frameworks like OAE and RadAlign provide a foundation for incorporating external knowledge, but the integration of heterogeneous clinical data types remains an open research challenge.

7.2 Explainability and Trust

For generative AI to be adopted in clinical practice, radiologists must trust its outputs. Explainability mechanisms that highlight which image regions contributed to specific textual findings are essential. The visual grounding represented by bounding boxes on the image, as implemented in CXRReportGen, represents one approach to explainability. Future systems

should provide interpretable intermediate representations that allow clinicians to verify the model's reasoning.

7.3 Real-Time and Interactive Systems

Current systems operate in batch mode, generating complete reports from static images. Future systems should support interactive workflows where radiologists can query the model, request clarifications, or correct errors in real-time. The integration of LLMs with vision models, as demonstrated in NaviGPT and similar systems, points toward interactive diagnostic assistants.

7.4 Federated Learning and Privacy Preservation

Medical imaging data is highly sensitive and subject to strict privacy regulations. Federated learning training models across multiple institutions without sharing raw data offers a pathway to develop robust models while preserving patient privacy. Future research should explore federated approaches to medical report generation that leverage distributed data assets without compromising confidentiality.

7.5 Clinical Integration and Workflow Design

The successful deployment of generative AI for medical report generation requires careful integration into clinical workflows. MedVersa's user studies demonstrated notable reductions in report writing time, but optimal workflow integration whether as a draft generator, a dictation assistant, or a quality assurance tool remains to be determined. Future research should explore human-AI collaboration models that maximize the complementary strengths of radiologists and AI systems.

8. CONCLUSION

Generative AI for medical image report generation has advanced dramatically over the past five years, evolving from proof-of-concept CNN-RNN architectures to sophisticated vision-language models, retrieval-augmented frameworks, and generalist foundation models capable of matching or surpassing task-specialized systems across multiple clinical tasks. The field has benefited from the availability of large-scale datasets (MIMIC-CXR, IU-Xray, CheXpert), standardized evaluation frameworks (RadEval), and rigorous clinical validation studies demonstrating clinical equivalence in the majority of cases.

Key technical advances include disease-aware dual-stage frameworks that learn pathology-specific semantic tokens, lesion-aware architectures that integrate global and regional representations through contrastive learning, retrieval-augmented mechanisms that ground generation in relevant exemplars, and contrastive decoding strategies that mitigate

hallucination by maintaining visual grounding throughout generation. Generalist foundation models like MedVersa and UniX demonstrate that a single, flexible architecture can handle diverse medical imaging tasks, suggesting a future where unified systems replace fragmented task-specific tools.

However, significant challenges remain. Hallucination, while reduced by contrastive decoding, has not been eliminated. Data scarcity limits generalizability across modalities and rare diseases. Uncertainty modeling critical for clinical decision-making remains underdeveloped. And the computational demands of large foundation models pose deployment challenges in resource-constrained settings.

The path forward requires continued advances in multimodal integration, explainability, uncertainty quantification, and clinical workflow design. As these systems mature, they promise not to replace radiologists but to augment their capabilities reducing burnout, improving consistency, and ultimately enabling more timely and accurate diagnoses for patients worldwide.

REFERENCES

1. Wang, X., Figueredo, G., Li, R., & Zhang, W. E. (2025). A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*, 103, 103627.
2. Wu, P., et al. (2026). A Disease-Aware Dual-Stage Framework for Chest X-ray Report Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40), 33953-33961.
3. Srivastava, S., Bhosale, M., Doermann, D., & Gao, M. (2026). CWCD: Category-Wise Contrastive Decoding for Structured Medical Report Generation. *arXiv:2604.10410v2*.
4. Khatoon, S., & Mahmood, A. (2026). Automated Radiological Report Generation from Breast Ultrasound Images Using Vision and Language Transformers. *Journal of Imaging*, 12(2), 68.
5. Hybrid framework for lesion-aware, clinically coherent chest X-ray report generation using contrastive learning and large language models. (2026). *Scientific Reports*, 16, 4645.
6. Wong, K. K. (2025). RadAlign: Advancing Radiology Report Generation with Vision-Language Concept Alignment. *MICCAI 2025*, LNCS 15966, 484-494.

7. Chen, K., et al. (2025). Observe, align, and enhance: a hierarchical retrieval-augmented vision-language model for generating radiology reports. *Health Information Science and Systems*, 13(1), 72.
8. MedVersa: A Generalist Foundation Model for Diverse Medical Imaging Tasks. (2026). *NEJM AI*, 3(4).
9. Boodoo, R. (2025). RadEval: A framework for radiology text evaluation. RadEval consolidates metrics from n-gram overlap (BLEU, ROUGE) to clinical concept-based scores (F1CheXbert, F1RadGraph, RaTEScore) and LLM-based evaluators (GREEN).
10. Assessment of ChatGPT performance in orbital MRI reporting with multimetric evaluation of transformer based language models. (2025). *Scientific Reports*, 15, 35654.
11. Gyeong Jung. (2025). From Images to Reports: The Future of Deep Learning in Radiology Report Generation. This systematic review covers key datasets including MIMIC-CXR, IU-XRay, and CheXpert.
12. A multimodal framework for explainable chest X-ray report generation. (2026). Generated reports show improved factual completeness and clinically relevant region-level attention.
13. Read like a radiologist: Efficient vision-language model for 3D medical imaging interpretation. (2026). MS-VLM surpasses existing methods in radiology report generation.
14. Generative AI in different imaging modalities for disease diagnosis: A review. (2026). Generative AI methods show accuracy improvements of between 5 and 20 per cent when synthetically enhanced data is utilized.
15. CURV: Coherent Uncertainty-Aware Reasoning in Vision-Language Models for X-Ray Report Generation. (2025). *NeurIPS*. Explicitly models uncertainty in findings and reasoning processes.
16. DiA-gnostic VLVAE: Disentangled Alignment-Constrained Vision Language Variational AutoEncoder for Robust Radiology Reporting with Missing Modalities. (2025). Achieves robust radiology reporting through Disentangled Alignment with Mixture-of-Experts.
17. TRRG: Towards Truthful Radiology Report Generation With Cross-modal Disease Clue Enhanced Large Language Models. (2025). *MICCAI 2025*. Proposes stage-wise training for cross-modal disease clue injection.

18. Argus: Benchmarking and Enhancing Vision-Language Models for 3D Radiology Report Generation. (2025). *ACL 2025*, 16448-16460. Provides benchmarking framework for 3D radiology report generation.
19. Self-Attentive Deep Fusion Framework with Transformer-Based Semantics for Emergency Head CT Reporting. (2026). Incorporates clinically validated preprocessing and intra-sequence self-attention.
20. UniX: Unifying Autoregression and Diffusion for Chest X-Ray Understanding and Generation. (2026). Decouples understanding and generation into autoregressive and diffusion branches.