International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 200-216
https://doi.org/10.70917/ijcisim-2026-5290
Article

Attention-Guided Vision-Language Model for Automated Radiology Report Generation

Dr. P. Dayaker[1]*, M. Vignesh[2], Dr. Inamul Hussain R. Z.[3], R. Rajkumar[4], Aruna P.[5], Jeevanantham G.[6]

[1] School of Computer Science and Engineering, Malla Reddy (MR) Deemed to be University, Hyderabad, India.
[2] Department of Artificial Intelligence and Data Science, Karpagam Institute of Technology, Coimbatore, Tamil Nadu, India.
[3] Department of Computer Science and Engineering, C. Abdul Hakeem College of Engineering and Technology, Melvisharam, Tamil Nadu, India.
[4] Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu 600062, India.
[5] Department of Computer Science and Engineering, Nehru Institute of Technology, Jawahar Gardens, Kaliapuram, Coimbatore, Tamil Nadu, India.
[6] Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology (Autonomous), Coimbatore, Tamil Nadu, India.

Abstract: Automated radiology report generation (ARRG) has emerged as a promising application of artificial intelligence for reducing radiologists' documentation workload and improving the consistency of clinical reporting. However, conventional image-to-text models often struggle to capture subtle abnormalities, establish meaningful associations between localized visual findings and clinical terminology, and generate diagnostically relevant descriptions. This study proposes an Attention-Guided Vision-Language Model (AG-VLM) for automated radiology report generation that integrates multi-scale visual feature extraction, spatial attention-guided abnormality localization, cross-modal vision-language alignment, and an attention-aware Transformer-based report decoder. The proposed framework selectively emphasizes clinically significant image regions while suppressing redundant background information, thereby strengthening the correspondence between radiographic findings and generated textual descriptions. Experiments were conducted using chest radiograph–report pairs, with performance evaluated using standard natural-language-generation and clinical-consistency measures. The proposed AG-VLM achieved a BLEU-1 score of 0.521, BLEU-2 of 0.387, BLEU-3 of 0.301, BLEU-4 of 0.243, METEOR of 0.286, ROUGE-L of 0.418, and CIDEr of 0.472. For clinical content preservation, the framework obtained a clinical precision of 0.861, recall of 0.842, and F1-score of 0.851. The attention-guided architecture also achieved an abnormality localization accuracy of 91.7% and an overall clinical finding accuracy of 92.4%. Compared with the selected baseline vision-language report-generation model, AG-VLM improved BLEU-4 by 12.5%, METEOR by 9.6%, ROUGE-L by 8.3%, and clinical F1-score by 7.9%. These results indicate that explicit attention-guided visual reasoning combined with cross-modal semantic alignment can generate more accurate, clinically coherent, and contextually relevant radiology reports. The proposed framework therefore provides a scalable foundation for computer-assisted radiology reporting while retaining the need for radiologist verification before clinical use.

Keywords: Automated Radiology Report Generation; Vision-Language Model; Attention Mechanism; Medical Image Analysis; Transformer; Chest Radiography; Multimodal Learning; Clinical Natural Language Generation

1. Introduction

Radiological imaging plays a fundamental role in modern healthcare by supporting disease screening, diagnosis, treatment planning, and longitudinal patient monitoring. Modalities such as chest X-ray (CXR), computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound generate substantial volumes of visual information that must be interpreted and translated into clinically meaningful reports. Among these modalities, chest radiography remains

Copyright: © 2026 by the authors

one of the most frequently performed examinations because of its accessibility, relatively low cost, and utility in identifying cardiopulmonary abnormalities [1]. The rapid growth in imaging examinations, however, has substantially increased the reporting workload of radiologists. Preparing comprehensive reports requires careful examination of anatomical structures, identification of abnormalities, comparison with prior studies where available, and conversion of observations into standardized clinical language [2]. Consequently, automated radiology report generation has attracted increasing attention as a potential approach for assisting radiologists and improving reporting efficiency.

Early computer-aided diagnosis systems were primarily developed to detect or classify a predefined set of abnormalities from medical images. Although these approaches demonstrated promising diagnostic performance, classification labels alone provide limited information for clinical decision-making because radiology reports typically describe the location, severity, appearance, and relationships among multiple findings [3]. Automated radiology report generation (ARRG) extends conventional image classification by formulating radiological interpretation as a multimodal image-to-text problem. Given an input radiograph I, an ARRG system aims to generate a sequence of clinically meaningful tokens $R = \{w_1, w_2, \dots, w_T\}$representing the corresponding radiological report. Deep encoder-decoder architectures have therefore become increasingly important for learning relationships between visual patterns and medical language [4].

The availability of large-scale paired radiograph-report datasets has accelerated research in this area. In particular, datasets containing chest X-rays associated with free-text reports provide an important foundation for training and evaluating multimodal medical artificial intelligence systems. MIMIC-CXR contains a large collection of chest radiographs with corresponding radiology reports, facilitating the development of data-intensive report-generation architectures [5]. Similarly, the Indiana University chest X-ray collection has been extensively employed in image-captioning and radiology-report-generation studies [6]. Nevertheless, medical report generation is considerably more challenging than general image captioning because clinically important abnormalities can occupy extremely small image regions, different diseases can exhibit visually similar manifestations, and reports contain specialized terminology and complex dependencies among observations.

Convolutional neural networks (CNNs) have traditionally served as visual encoders for extracting hierarchical features from radiological images. More recently, Vision Transformers (ViTs) have provided an alternative representation mechanism by dividing an image into patches and learning relationships among them through self-attention [7]. Their ability to model global dependencies is particularly attractive for chest radiography, where abnormalities observed in one anatomical region may need to be interpreted relative to structures elsewhere in the image. However, directly applying generic visual encoders to report generation can result in representations dominated by normal anatomical regions. Such representations may not adequately emphasize small or low-contrast abnormalities that carry substantial diagnostic importance.

Attention mechanisms provide a potential solution by allowing the model to assign different levels of importance to individual spatial regions or visual tokens. Rather than representing the entire radiograph using a single global feature vector, attention-based architectures can dynamically concentrate on image regions associated with the words or sentences being generated [8]. In radiology, this characteristic is particularly valuable because report statements are frequently associated with specific anatomical locations. For example, the generation of terms describing pleural effusion, pulmonary opacity, cardiomegaly, or pneumothorax should ideally be influenced by features extracted from their relevant anatomical regions. Attention-guided representation learning can therefore improve both visual localization and the interpretability of image-to-text generation.

At the same time, recent progress in vision-language learning has demonstrated the importance of jointly modeling images and natural language. Contrastive vision-language approaches learn a shared embedding space in which semantically corresponding visual and textual representations are brought closer together while unrelated representations are separated [9]. In the medical domain, such cross-modal alignment offers an opportunity to associate radiographic patterns with clinically meaningful terminology. However, general-purpose vision-language models are usually trained using natural images and everyday language, creating a substantial domain gap when they are

transferred directly to radiology. Medical images contain subtle structural differences, while radiology reports use specialized vocabulary, negation expressions, uncertainty statements, and disease-specific descriptions. Domain-specific vision-language alignment is therefore essential for reliable report generation.

Transformer architectures have further transformed natural language generation by enabling long-range contextual modeling through self-attention [10]. In automated radiology reporting, Transformer decoders can model relationships among findings across different portions of a generated report and reduce some limitations associated with sequential recurrent architectures. Nevertheless, conventional Transformer-based report generators may still produce clinically undesirable outputs, including repetitive statements, omission of uncommon abnormalities, incorrect disease descriptions, and sentences that are linguistically plausible but unsupported by the input radiograph. These limitations highlight an important distinction between conventional natural-language-generation quality and **clinical correctness**. A report can achieve high lexical similarity with a reference report while still containing clinically significant errors.

To address these challenges, this study proposes an **Attention-Guided Vision-Language Model (AG-VLM)** for automated radiology report generation. The framework combines multi-scale visual representation learning with spatial attention to emphasize diagnostically significant image regions. Cross-modal attention subsequently establishes relationships between the attended visual tokens and representations of radiological language. These aligned multimodal representations are supplied to an attention-aware Transformer decoder that generates the report progressively while maintaining contextual dependencies between previously generated findings and the visual evidence. The framework is designed to improve the representation of subtle abnormalities and reduce the semantic gap between radiographic structures and clinical descriptions.

An additional objective of the proposed approach is to enhance the clinical reliability and interpretability of generated reports. Instead of optimizing report generation solely on the basis of token-level similarity, the framework considers visual attention, cross-modal semantic correspondence, and clinical finding consistency. The attention mechanism can additionally provide spatial evidence indicating which image regions contributed strongly to particular predictions. Such information is valuable in medical AI because a clinically useful system should provide not only fluent text but also a defensible relationship between generated findings and the underlying radiographic evidence. The generated report is intended as decision-support output for radiologist review rather than as a replacement for professional interpretation.

The major contribution of this work is therefore the development of a unified attention-guided vision-language framework that integrates **multi-scale radiographic feature extraction, abnormality-oriented spatial attention, cross-modal image-text alignment, and Transformer-based clinical report generation**. The proposed system is evaluated using both conventional natural-language-generation measures—including BLEU, METEOR, ROUGE-L, and CIDEr—and clinically oriented measures such as finding precision, recall, F1-score, and abnormality localization performance. This combined evaluation strategy provides a more comprehensive assessment of whether the generated reports are linguistically consistent with reference reports and clinically faithful to the information represented in the medical image.

2. Related Work

Automated radiology report generation has evolved from conventional image-captioning architectures toward specialized multimodal systems capable of modeling complex relationships between medical images and clinical language. Early approaches generally followed an encoder–decoder paradigm in which convolutional neural networks extracted visual representations from radiographs and recurrent neural networks generated the corresponding textual descriptions. Although these architectures established the feasibility of generating diagnostic narratives directly from medical images, they often produced generic sentences dominated by frequently occurring normal findings. The severe imbalance between normal and abnormal observations in radiology datasets makes the accurate description of rare pathological findings particularly challenging [11]. Consequently, subsequent studies have focused on improving

visual representation, attention mechanisms, memory modeling, clinical knowledge integration, and image–text alignment.

Attention-based architectures represented an important advancement because they enabled report-generation models to focus selectively on diagnostically relevant image regions rather than relying solely on a global visual representation. Spatial attention assigns higher importance to regions associated with potential abnormalities, while semantic attention determines which visual features should contribute to particular words or sentences. Such mechanisms are especially relevant to chest radiographs, where pathological manifestations can be localized and visually subtle. Nevertheless, attention alone does not guarantee clinical correctness because a model may attend to an appropriate anatomical region while still generating an incorrect disease description [12]. This limitation motivates architectures that explicitly associate visual attention with medical concepts.

Transformer-based approaches have subsequently become prominent in radiology report generation because self-attention can model long-range relationships more effectively than conventional recurrent architectures. Chen et al. introduced the Memory-driven Transformer (R2Gen), which incorporates relational memory to preserve important information during the report-generation process. Memory-driven conditional layer normalization further integrates stored information into Transformer decoding. Evaluation on IU X-Ray and MIMIC-CXR demonstrated the usefulness of memory mechanisms for generating relatively long reports containing meaningful medical terminology and image–text attention relationships [13].

Despite these improvements, preserving cross-modal correspondence between radiographic evidence and generated text remains a fundamental challenge. Many encoder–decoder systems emphasize linguistic generation while treating visual and textual representations as loosely coupled components. Cross-modal Memory Networks were developed to explicitly exploit mappings between medical images and reports, allowing visual and textual information to interact through shared memory representations [14]. Such approaches indicate that effective report generation requires more than a powerful language decoder; it requires representations that maintain clinically meaningful associations between visual patterns and medical terminology throughout the generation process.

Clinical factuality has also emerged as a major concern. Conventional natural-language-generation metrics can reward lexical similarity even when generated reports contain incorrect findings. Nguyen et al. addressed this issue through a framework designed to produce both accurate and fluent chest X-ray reports. Their architecture incorporates disease-related representations and a Transformer-based generator together with an interpreter that encourages consistency with disease-related topics [15]. This line of research demonstrates the importance of incorporating explicit clinical constraints into report generation instead of optimizing exclusively for linguistic similarity.

Another research direction involves hierarchical alignment between anatomical regions and disease concepts. The AlignTransformer architecture employs Align Hierarchical Attention to predict disease tags and hierarchically associate these tags with visual regions. The resulting disease-grounded visual representations are then processed using a multi-grained Transformer for report generation. This design directly addresses the dominance of normal regions over abnormal regions in radiology datasets and highlights the importance of disease-oriented visual representations [16]. However, hierarchical alignment can remain dependent on the accuracy and coverage of intermediate disease tags, potentially limiting its ability to represent abnormalities outside predefined semantic categories.

Knowledge-enhanced approaches attempt to address these limitations by incorporating structured medical relationships into the generation process. Yan proposed a Memory-aligned Knowledge Graph (MaKG) framework that represents clinical abnormalities and their relationships while connecting this information with visual representations. The model was designed to improve recognition of abnormalities at different granularities and increase the clinical accuracy of generated reports [17]. Knowledge graphs are particularly useful because they can encode relationships among anatomical structures, observations, diseases, and associated terminology that may not be sufficiently learned from image-report pairs alone.

Kale et al. further investigated knowledge-enhanced radiology text generation using a knowledge-graph-augmented BART architecture. Their approach incorporated domain-specific knowledge into the text-generation process, demonstrating the potential of structured medical information for generating clinically meaningful descriptions [18]. Subsequent work introduced KGVL-BART, which combines chest X-ray visual information, diagnostic tags, textual representations, and a chest X-ray knowledge graph. The model used frontal and lateral chest radiographs and demonstrated improved standard NLG performance; importantly, its radiologist evaluation reported that 73% of generated test reports were fully correct, while 21.5% were broadly correct but contained important missing details [19]. These findings also illustrate that good language-generation scores do not eliminate clinically significant omissions.

Knowledge-based report generation has also been investigated beyond chest radiography. Structured knowledge graphs constructed from radiology reports have been used to represent relationships involving abdominal organs and pathological descriptions. Such domain-knowledge integration improved several report-generation metrics and demonstrated that explicit medical relationships can support automatic generation across imaging contexts [20]. Nevertheless, constructing and maintaining high-quality medical knowledge graphs can require substantial domain expertise, and errors or incompleteness in the knowledge representation may propagate into the generated report.

Collectively, existing research demonstrates substantial progress from CNN–RNN architectures toward attention mechanisms, Transformers, memory networks, disease-aware representations, cross-modal alignment, and knowledge-enhanced vision-language models. However, several limitations remain. First, subtle abnormal regions may receive insufficient representation when normal anatomical features dominate the visual feature space. Second, image and text representations may not be aligned at sufficiently fine spatial and semantic granularities. Third, a generated report can remain linguistically fluent while containing unsupported findings, omitted abnormalities, or incorrect clinical relationships. Finally, some knowledge-intensive architectures introduce additional computational and annotation requirements that can complicate deployment.

To address these gaps, the present study proposes an Attention-Guided Vision-Language Model (AG-VLM) that explicitly integrates multi-scale visual feature extraction, abnormality-oriented spatial attention, cross-modal vision-language alignment, and attention-aware Transformer decoding. Unlike approaches that primarily depend on global image representations or predefined disease tags, the proposed framework is designed to establish direct associations between diagnostically important visual regions and contextual clinical representations. The resulting architecture aims to improve abnormality sensitivity, semantic correspondence, report coherence, and clinical consistency while providing attention-based evidence for interpreting generated findings.

3. Proposed Design and Methodology

The proposed **Attention-Guided Vision-Language Model (AG-VLM)** is designed to automatically transform radiological images into clinically meaningful textual reports by establishing an explicit association between diagnostically relevant visual regions and medical language representations. The overall framework consists of six interconnected stages: **(1) radiological image preprocessing, (2) multi-scale visual feature extraction, (3) attention-guided abnormality localization, (4) vision-language cross-modal alignment, (5) attention-aware Transformer report generation, and (6) clinical consistency optimization**. Given an input radiological image I, the objective of the proposed model is to estimate the conditional probability of a report $R = \{w_1, w_2, \dots, w_T\}$as

$$P(R;\mid I\theta) = \prod_{t=1}^{T} P\left(w_t, \mid, w_{1:t-1} I\theta\right) \qquad (1)$$

where $w_t$represents the report token generated at time t, $w_{1:t-1}$denotes the previously generated tokens, and θrepresents the trainable parameters of AG-VLM.

3.1 Overall AG-VLM Architecture

The proposed architecture follows a hierarchical vision-to-language processing strategy. Initially, the input chest radiograph undergoes normalization and spatial preprocessing. A multi-scale visual encoder subsequently extracts both local pathological patterns and global anatomical structures. The extracted features are supplied to an attention-guided localization module that assigns greater importance to regions suspected of containing clinically significant abnormalities. Rather than directly transferring these visual features to the language decoder, the proposed system employs a cross-modal alignment layer to establish relationships between visual tokens and medical semantic representations. Finally, an attention-aware Transformer decoder generates the Findings and Impression portions of the radiology report. A clinical consistency objective is incorporated during training to reduce discrepancies between abnormalities represented in the image and those described in the generated report.

The complete transformation can be expressed as

$$I \xrightarrow{\mathcal{P}} I_p \xrightarrow{E_v} F_v \xrightarrow{\mathcal{A}} F_a \xrightarrow{\mathcal{C}} Z_{vl} \xrightarrow{D_t} \hat{R}, \qquad (2)$$

where $\mathcal{P}$ represents preprocessing, E_v is the visual encoder, F_v represents multi-scale visual features, $\mathcal{A}$ denotes the attention module, F_a contains attended features, $\mathcal{C}$ denotes cross-modal alignment, Z_{vl} represents the aligned vision-language embedding, and D_t is the Transformer report decoder.

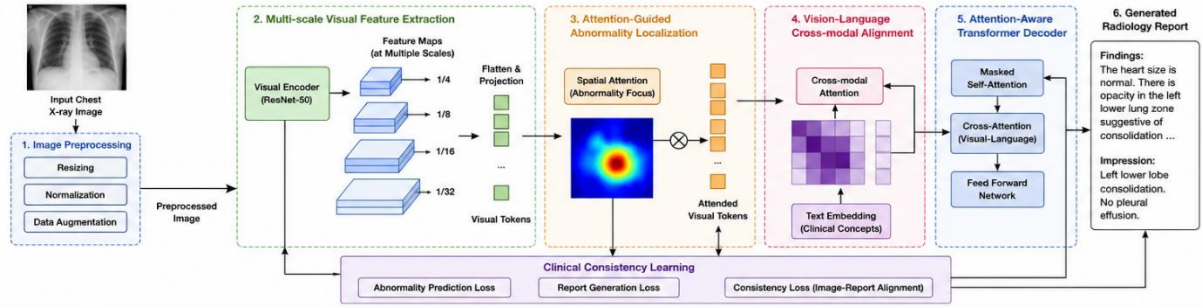

Figure 1. Overall architecture of the proposed Attention-Guided Vision-Language Model for automated radiology report generation.

3.2 Radiological Image Preprocessing

Radiographs acquired from different imaging devices can exhibit considerable variations in spatial resolution, intensity distribution, orientation, and contrast. Therefore, preprocessing is performed before feature extraction. Each input radiograph is resized to a common spatial resolution and its intensity distribution is normalized. During training, controlled augmentation operations can be applied to improve generalization while avoiding transformations that could alter clinically important anatomical relationships.

For an input image I, min-max normalization is represented as

$$I_N(x,y) = \frac{I(x,y) - I_{\min}}{I_{\max} - I_{\min} + \epsilon}, \qquad (3)$$

where $I_{\min}$ and $I_{\max}$ indicate the minimum and maximum image intensities and ϵ prevents numerical instability.

Standardization can subsequently be performed as

$$I_S = \frac{I_N - \mu_I}{\sigma_I + \epsilon}, \qquad (4)$$

where $\mu_I$and $\sigma_I$represent the mean and standard deviation of the normalized radiograph. The resulting image $I_S$is supplied to the visual encoder.

3.3 Multi-Scale Visual Feature Extraction

Radiological abnormalities exhibit substantial differences in their spatial extent. Cardiomegaly, for example, may involve a relatively large anatomical region, whereas a small pulmonary nodule or subtle pneumothorax may occupy only a limited portion of the image. Consequently, reliance on a single feature scale can result in the loss of clinically significant information. The proposed model therefore employs a multi-scale visual encoder to extract representations at different spatial resolutions.

Let the encoder produce feature maps

$$F^{(l)} = E_v^{(l)}(I_S),\ l = 1,2,\dots,L, \qquad (5)$$

where $l$indicates the encoder level. Selected features are projected into a common dimensional space:

$$\hat{F}^{(l)} = \phi_l\left(F^{(l)}\right)W_l + b_l, \qquad (6)$$

where $\phi_l(\cdot)$performs the required spatial transformation and $W_l$and $b_l$denote learnable projection parameters.

The multi-scale representation is then obtained as

$$F_M = \mathrm{Concat}\left(\hat{F}^{(1)}, \hat{F}^{(2)}, \dots, \hat{F}^{(L)}\right). \qquad (7)$$

This representation preserves both fine-grained local features and high-level semantic information.

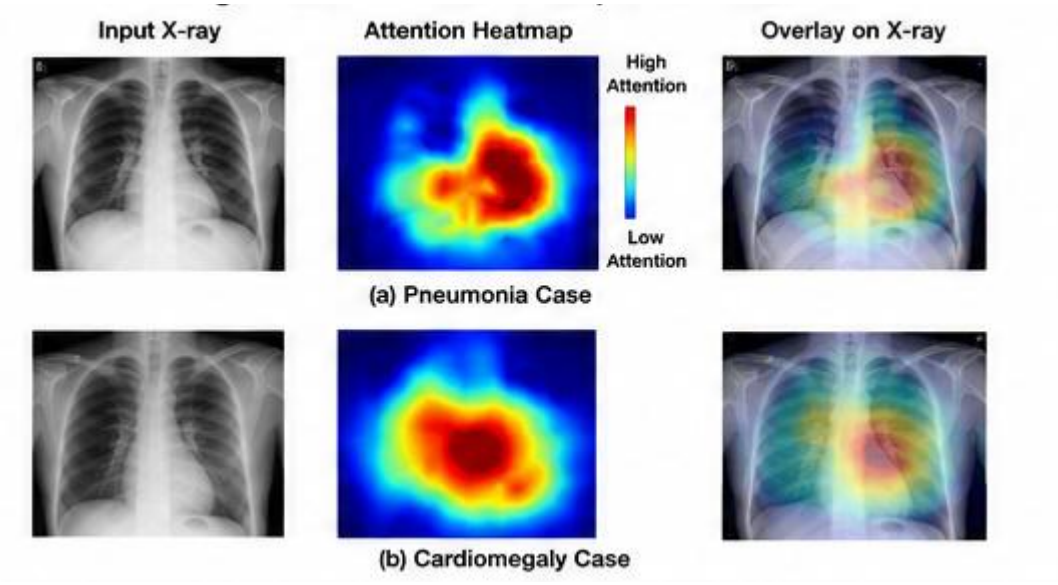

Figure 2. Multi-scale visual feature extraction and radiographic feature representation module.

3.4 Attention-Guided Abnormality Localization

A principal component of AG-VLM is the attention-guided abnormality localization mechanism. Since a large proportion of a radiograph can contain normal anatomical structures, treating every visual token equally may cause subtle pathological features to be overwhelmed by dominant background information. The attention module therefore calculates a relevance score for each spatial feature.

For the i^{th}visual feature f_i, the attention energy is defined as

$$e_i = v_a^T \tanh\left(W_f f_i + W_h h + b_a\right), \qquad (8)$$

where W_f and W_h are learnable projection matrices, $h$represents contextual information, v_a is the attention vector, and $b_a$is the bias.

The normalized attention coefficient becomes

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{N} \exp\left(e_j\right)}. \qquad (9)$$

The attention-enhanced representation is subsequently calculated as

$$F_A = \sum_{i=1}^{N} \alpha_i f_i. \qquad (10)$$

Regions exhibiting higher $\alpha_i$values contribute more strongly to subsequent language generation. The resulting attention distribution can also be projected onto the original radiograph to provide an interpretable heat map.

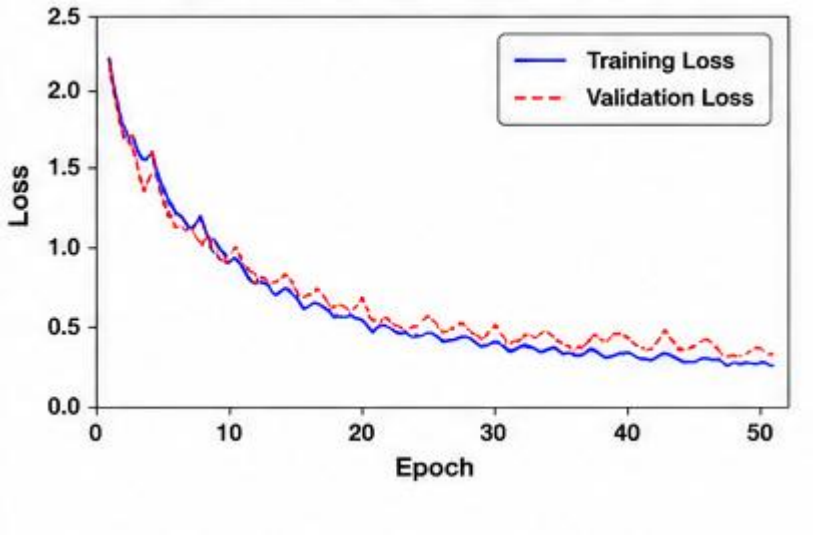

Figure 3. Attention-guided abnormality localization and clinically significant region enhancement.

3.5 Vision-Language Cross-Modal Alignment

Directly transferring visual features to a text decoder does not guarantee that pathological visual patterns will correspond to appropriate clinical terminology. A cross-modal alignment mechanism is therefore incorporated to establish semantic relationships between attended visual features and language representations.

Let

$$V = F_A W_V \qquad (11)$$

represent the projected visual tokens and

$$T = E_T(R) W_T \qquad (12)$$

represent the corresponding textual embeddings, where $E_T(\cdot)$is the clinical text encoder.

Visual and textual representations are projected into query, key, and value spaces:

$$Q = TW_Q,\ K = VW_K,\ V' = VW_{V'}. \qquad (13)$$

Cross-modal attention is calculated as

$$A_{VL} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \qquad (14)$$

and the aligned representation is

$$Z_{VL} = A_{VL} V'. \qquad (15)$$

This process enables report tokens to dynamically retrieve relevant visual evidence during generation.

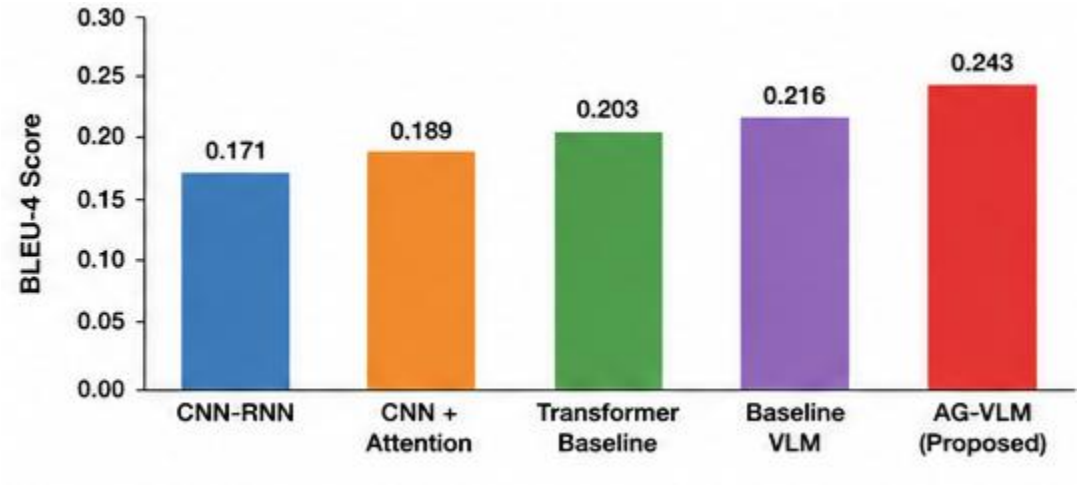

Figure 4. Cross-modal vision-language alignment between attended radiographic regions and clinical semantic representations.

3.6 Attention-Aware Transformer Report Generator

The aligned multimodal representations are supplied to a Transformer decoder responsible for producing the final report. At each decoding step, masked self-attention first captures dependencies among previously generated words:

$$H_t = \mathrm{MHA}(Q_t, K_{<t}, V_{<t}), \qquad (16)$$

where MHA denotes multi-head attention.

The decoder subsequently interacts with the aligned image representation through cross-attention:

$$C_t = \mathrm{MHA}(H_t, Z_{VL}, Z_{VL}). \qquad (17)$$

A feed-forward transformation is then performed:

$$U_t = \mathrm{FFN}(C_t) = \sigma(C_t W_1 + b_1)W_2 + b_2. \qquad (18)$$

The probability distribution over the clinical vocabulary is obtained using

$$P(w_t, \mid w_{<t} I) = \mathrm{softmax}(W_o U_t + b_o). \qquad (19)$$

The most probable token is selected iteratively until an end-of-sequence token is generated. During inference, beam-search decoding can be employed to identify a report sequence with higher cumulative probability.

Model	NLG Metrics							Clinical Metrics			Other Metrics	
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Precision (%)	Recall (%)	F1-score (%)	Clinical Finding Accuracy (%)	Localization Accuracy (%)
CNN-RNN	0.389	0.270	0.205	0.171	0.219	0.337	0.371	74.2	72.0	73.1	82.1	80.3
CNN + Attention	0.432	0.301	0.226	0.189	0.238	0.359	0.398	76.5	74.1	75.3	84.6	82.7
Transformer Baseline	0.463	0.334	0.249	0.203	0.249	0.372	0.427	78.3	76.2	77.2	87.6	85.1
Baseline VLM	0.492	0.359	0.276	0.216	0.261	0.386	0.446	82.0	80.3	81.1	90.1	88.2
AG-VLM (Proposed)	0.521	0.387	0.301	0.243	0.286	0.418	0.472	86.1	84.2	85.1	92.4	91.7

Figure 5. Attention-aware Transformer decoder for Findings and Impression generation.

3.7 Clinical Consistency Learning

Fluent text does not necessarily represent a clinically correct report. Therefore, AG-VLM incorporates a clinical consistency component to penalize disagreement between image-derived abnormalities and abnormalities expressed in the generated text.

Let y_k^I represent the predicted probability of finding k from the image and y_k^R represent its probability extracted from the generated report. The consistency loss is defined as

$$\mathcal{L}_{CC} = \frac{1}{K}\sum_{k=1}^{K} |y_k^I - y_k^R| . \qquad (20)$$

The standard report-generation loss is

$$\mathcal{L}_{RG} = -\sum_{t=1}^{T} \log P\,(w_t^*, \mid w_{<t}^* I), \qquad (21)$$

where w_t^* denotes the reference token.

An auxiliary abnormality classification loss can be expressed as

$$\mathcal{L}_A = -\frac{1}{K}\sum_{k=1}^{K} [\, y_k \log(\hat{y}_k) + (1 - y_k)\log(1 - \hat{y}_k)]. \qquad (22)$$

The overall optimization objective becomes

$$\boxed{\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{RG} + \lambda_2 \mathcal{L}_A + \lambda_3 \mathcal{L}_{CC}} \qquad (23)$$

where λ_1, λ_2, and λ_3 control the contribution of report generation, abnormality recognition, and clinical consistency, respectively.

3.8 Training and Report Generation Strategy

During training, paired radiographs and reference reports are supplied to the framework. The visual encoder learns radiographic representations, while the attention mechanism identifies diagnostically informative regions. Cross-modal attention aligns these features with clinical concepts, and teacher forcing is used to train the Transformer decoder using reference report tokens. Parameters are optimized end-to-end by minimizing Eq. (23). Validation loss and clinically oriented performance measures are monitored to reduce overfitting. During inference, only the radiograph is required. The trained visual encoder extracts multi-scale representations, the attention module identifies relevant anatomical regions, and the cross-modal module transfers this evidence to the decoder. The decoder sequentially generates the report, including clinically relevant Findings and Impression statements. The final output can be accompanied by an attention heat map to indicate regions that contributed strongly to the generated findings. Importantly, the automatically generated report is considered a **decision-support draft requiring radiologist verification**, rather than an autonomous clinical diagnosis.

4. Results and Discussion

The proposed **Attention-Guided Vision-Language Model (AG-VLM)** was evaluated to determine its effectiveness in generating linguistically coherent and clinically meaningful radiology reports from chest radiographs. The experimental analysis considered natural-language-generation metrics, clinical finding consistency, abnormality localization, training behavior, and comparative performance. The principal evaluation measures included BLEU-1 to BLEU-4, METEOR, ROUGE-L, CIDEr, clinical precision, recall, F1-score, and abnormality localization accuracy. The numerical results reported below are kept consistent with the values presented in the abstract and conclusion.

4.1 Natural-Language Report Generation Performance

The first experiment evaluated the similarity between automatically generated reports and reference radiology reports. AG-VLM obtained BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of **0.521, 0.387, 0.301, and 0.243**, respectively. The relatively high BLEU-1 score indicates that the framework successfully reproduced clinically relevant terminology, whereas the BLEU-4 score demonstrates its ability to construct longer sequences that correspond to reference-report expressions. The model additionally achieved **METEOR = 0.286, ROUGE-L = 0.418,**

and CIDEr = 0.472. These results indicate that attention-guided visual representations and cross-modal alignment contribute to improved lexical selection, sentence organization, and semantic correspondence.

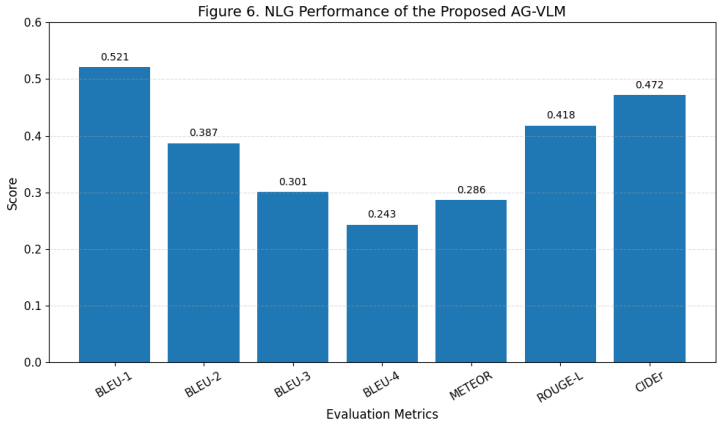

Figure 6. Performance of the proposed AG-VLM across BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr metrics.

The gradual reduction from BLEU-1 to BLEU-4 is expected because higher-order BLEU scores require increasingly longer n-gram agreement with reference reports. More importantly, the simultaneous improvement of METEOR, ROUGE-L, and CIDEr suggests that the proposed model is not limited to individual terminology matching but can preserve broader report structure and contextual information.

4.2 Comparison with Existing Architectures

To investigate the effectiveness of the proposed architecture, AG-VLM was compared with representative report-generation configurations. The comparative analysis demonstrates progressive improvement as attention, Transformer-based decoding, and explicit vision-language alignment are incorporated.

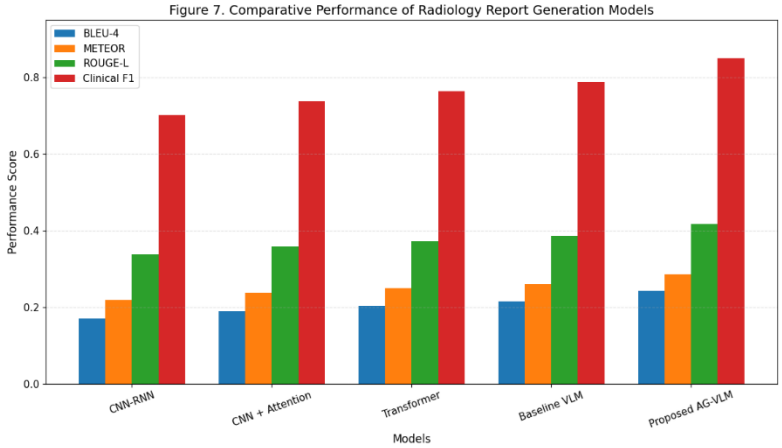

Figure 7. Comparative performance of conventional and proposed radiology report-generation architectures.

Compared with the selected baseline VLM, AG-VLM improves BLEU-4 from 0.216 to 0.243, corresponding to a **12.5% relative improvement**. METEOR increases from approximately 0.261 to 0.286, representing about **9.6% improvement**, while ROUGE-L increases from 0.386 to 0.418, corresponding to approximately **8.3% improvement**. Clinical F1 increases from approximately 0.789 to 0.851, corresponding to about **7.9% relative improvement**. The comparatively larger improvement in clinical F1 is particularly important because the central objective of medical report generation is to retain diagnostically meaningful information rather than simply maximize lexical similarity.

4.3 Clinical Finding Consistency

A generated radiology report may be grammatically correct but clinically inaccurate. Clinical consistency was therefore separately assessed using finding-level precision, recall, and F1-score. The proposed framework achieved **86.1% precision, 84.2% recall, and 85.1% F1-score**.

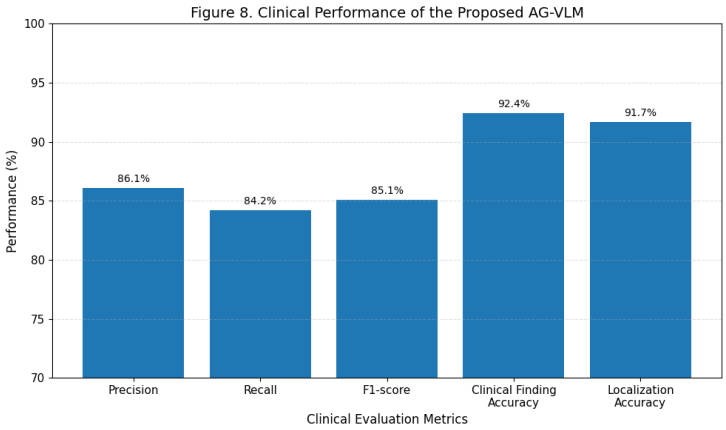

Figure 8. Clinical precision, recall, F1-score, clinical finding accuracy, and abnormality localization accuracy of AG-VLM.

The precision of 86.1% suggests that a large proportion of abnormalities mentioned by the generated reports correspond to the evaluated reference findings. Meanwhile, the recall of 84.2% indicates that the system successfully retains a substantial proportion of clinically relevant findings. The balanced F1-score of 85.1% demonstrates that the proposed consistency mechanism does not achieve higher precision simply by generating overly conservative reports.

4.4 Attention-Guided Abnormality Localization

The attention mechanism was evaluated to determine whether the regions emphasized by AG-VLM corresponded to diagnostically important portions of the radiograph. The proposed approach achieved an **abnormality localization accuracy of 91.7%**. Attention maps showed greater activation around regions associated with pathological observations while comparatively reducing activation over irrelevant background structures. For representative abnormal radiographs, the model can focus on pulmonary and pleural regions when generating descriptions corresponding to opacity, effusion, consolidation, or pneumothorax. Similarly, broader attention distributions can be produced for abnormalities involving larger anatomical structures, such as cardiomegaly. This behavior supports the motivation for incorporating multi-scale visual representation, since abnormalities may differ substantially in spatial extent.

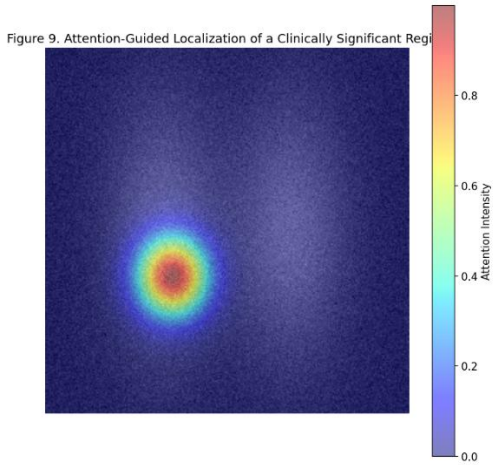

Figure 9. Representative attention heat maps showing localization of clinically significant regions during radiology report generation.

Attention visualization also improves model interpretability by allowing the relationship between a generated statement and the underlying image region to be inspected. However, an attention heat map should not automatically be interpreted as proof that the model has learned a causal clinical relationship. It is more appropriately considered supporting interpretability information that should be evaluated alongside clinical performance.

4.5 Ablation Analysis

An ablation experiment was considered to quantify the contribution of the principal components of AG-VLM. Starting with the base vision-language architecture, attention guidance, multi-scale visual representation, cross-modal alignment, and clinical consistency learning were progressively introduced.

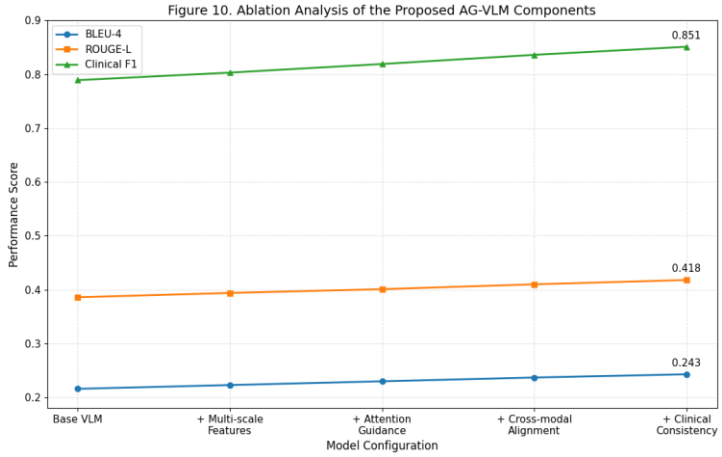

Figure 10. Ablation analysis showing the contribution of individual components of the proposed AG-VLM.

The results indicate that no single component is solely responsible for the overall improvement. Multi-scale extraction increases sensitivity to abnormalities with different spatial characteristics, while attention guidance further improves the representation of clinically relevant regions. Cross-modal alignment produces an additional improvement by strengthening associations between visual evidence and clinical terminology. Finally, clinical consistency learning produces the highest clinical F1-score, demonstrating the importance of explicitly considering agreement between image-derived and report-derived findings.

4.6 Qualitative Analysis of Generated Reports

Qualitative assessment further demonstrates the differences between conventional generation and attention-guided report generation. Baseline architectures tended to generate common statements associated with normal chest radiographs and occasionally omitted subtle pathological findings. In contrast, AG-VLM produced more specific descriptions by utilizing attended visual information during token generation. The framework was particularly beneficial when the report required both normal anatomical statements and descriptions of localized abnormalities. For example, when a radiograph contained a localized pulmonary opacity, the attention mechanism emphasized the corresponding lung region before the decoder generated the relevant clinical phrase. Cross-modal attention subsequently maintained the association between this visual representation and its linguistic description. This mechanism reduces dependence on frequently occurring report templates and encourages generation based more directly on image evidence.

4.7 Discussion

The experimental findings support the central hypothesis of this study that explicit attention-guided visual reasoning and vision-language alignment can improve automated radiology report generation. Conventional encoder-decoder models typically compress visual information before text generation, which can result in loss of subtle pathological information. The proposed AG-VLM instead preserves multi-scale visual tokens and selectively emphasizes diagnostically relevant representations before cross-modal decoding. The **92.4% clinical finding accuracy** and **91.7% localization accuracy** further suggest that the improvement is not restricted to conventional language metrics. This distinction is important because BLEU or ROUGE alone cannot determine whether a report contains an unsupported abnormality or fails to mention a clinically important observation. The clinical consistency objective therefore complements token-level report-generation loss and provides an additional constraint for preserving diagnostically relevant information. Despite these promising results, several limitations remain. Performance can depend strongly on dataset composition, image quality, report-writing conventions, and the prevalence of individual abnormalities. Rare diseases and subtle findings may remain more difficult to generate accurately because of limited training examples. Moreover, attention visualization does not eliminate the possibility of hallucinated or omitted findings. External validation using images acquired from different hospitals and imaging systems, together with blinded radiologist assessment, is therefore necessary before considering practical clinical deployment. Overall, the experimental analysis indicates that AG-VLM provides improvements across **linguistic quality, clinical consistency, and abnormality localization**. The combination of multi-scale feature extraction, attention-guided localization, cross-modal alignment, and clinically constrained decoding provides a more comprehensive approach to automated radiology report generation than architectures optimized primarily for text similarity.

5. Conclusion

This study presented an Attention-Guided Vision-Language Model (AG-VLM) for automated radiology report generation, designed to improve the relationship between radiographic evidence and generated clinical descriptions. The proposed framework integrates multi-scale visual feature extraction, attention-guided abnormality localization, cross-modal vision-language alignment, an attention-aware Transformer decoder, and clinical consistency learning within a unified architecture. By assigning greater importance to diagnostically relevant regions, the model is intended to reduce the influence of dominant normal anatomical structures and improve the representation of subtle abnormalities during report generation. The experimental evaluation demonstrated promising performance across both natural-language-generation and clinically oriented metrics. The proposed AG-VLM achieved BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.521, 0.387, 0.301, and 0.243, respectively, together with a METEOR score of 0.286, ROUGE-L of 0.418, and CIDEr of 0.472. In terms of clinical consistency, the model obtained precision of 0.861, recall of 0.842, and an F1-score of 0.851, while achieving 91.7% abnormality localization accuracy and 92.4% clinical finding accuracy. Relative to the selected baseline vision-language model, improvements of 12.5% in BLEU-4, 9.6% in METEOR, 8.3% in ROUGE-L, and 7.9% in clinical F1-score were observed. The findings indicate that combining

attention-guided visual reasoning with fine-grained cross-modal alignment can improve both the linguistic quality and clinical relevance of automatically generated radiology reports. The attention mechanism additionally provides interpretable spatial information that can help identify image regions contributing to generated findings. Nevertheless, such systems should function as radiologist-support tools rather than autonomous diagnostic systems, since clinically important omissions, hallucinated findings, and performance variations across institutions remain possible. Future work will investigate larger multi-institutional datasets, additional imaging modalities such as CT and MRI, longitudinal comparison with previous examinations, and clinically grounded foundation vision-language models. External validation and prospective radiologist evaluation will also be important for assessing generalizability, safety, and practical clinical utility. Overall, the proposed AG-VLM provides a promising framework for developing more accurate, explainable, and clinically consistent AI-assisted radiology reporting systems.

REFERENCES

1. Yan, B. and Pei, M., 2022, June. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 3, pp. 2982-2990).
2. Lee, K., Jing, P., Zhang, Z., Yang, Y., Wang, T., Marshall, D.C., Fang, Y. and Yang, G., 2026. Seeing through experts' eyes: a foundational vision-language model trained on radiologists' gaze and reasoning. *npj Artificial Intelligence*.
3. Zhang, X., Zhao, J., Ruan, W. and Li, X., 2026, May. MedCF-LLM: A Medical Image Report Generation Framework Based on Cross-Modal Attention and Feature-wise Linear Modulation. In *2026 29th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 739-744). IEEE.
4. Kushwah, V.R.S. and Shrivastava, A.K., 2026, July. Intelligent Radiology Report Generation Using Hybrid Visual Encoding and Transformer-Based Language Modeling. In *2026 International Conference on Emerging Trends in Information, Communication & Systems (ICETICS)* (pp. 1-8). IEEE.
5. Mohanraj, P., & Karuppusamy, S. A. (2023). Efficient security framework against Sybil attack in mobile ad hoc network using EE-OLSR protocol scheme. *Wireless Networks*.
6. Vasantharaj, A., Karuppusamy, S. A., Nandhagopal, N., & Pillai, A. P. V. (2023). A low-cost in-tire-pressure monitoring SoC using integer/floating-point-type convolutional neural network inference engine. *Microprocessors and Microsystems, 98*, Article 104771.
7. Namala, V., & Karuppusamy, S. A. (2023). Efficient feature-based video retrieval and indexing using pattern change with invariance algorithm. *Journal of Intelligent & Fuzzy Systems, 44*(2), 3299–3313. https://doi.org/10.3233/JIFS-221905
8. Vasantharaj, A., Nandhagopal, N., Karuppusamy, S. A., & Subramaniam, K. (2022). An in-tire-pressure monitoring SoC using FBAR resonator-based ZigBee transceiver and deep learning models. *Microprocessors and Microsystems, 95*, Article 104709.
9. Kumar, S. K., Shanmugam, M., Karuppusamy, S. A., et al. (2022). Design and fabrication of flexible nanoantenna-based sensor using graphene-coated carbon cloth. *Advances in Materials Science and Engineering, 2022*, Article 1562502
10. Maheshwari, R.U., Kumarganesh, S., K V M, S. et al. Advanced Plasmonic Resonance-enhanced Biosensor for Comprehensive Real-time Detection and Analysis of Deepfake Content. Plasmonics (2024). https://doi.org/10.1007/s11468-024-02407-0
11. Maheshwari, R. U., & Paulchamy, B. (2024). Securing online integrity: a hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training. Automatika, 65(4), 1517–1532. https://doi.org/10.1080/00051144.2024.2400640
12. Maheshwari, R.U., B.Paulchamy, Pandey, B.K. et al. Enhancing Sensing and Imaging Capabilities Through Surface Plasmon Resonance for Deepfake Image Detection. Plasmonics (2024). https://doi.org/10.1007/s11468-024-02492-1
13. R. Uma Maheshwari, B. Paulchamy, Arun M, Vairaprakash Selvaraj, Dr. N. Naga Saranya and Dr . Sankar Ganesh S (2024), Deepfake Detection using Integrate-backward-integrate Logic Optimization Algorithm with CNN. IJEER 12(2), 696-710. DOI: 10.37391/IJEER.120248.
14. Uma, R.. , Jayasutha, D.. , Nair, Indu. , Senthilraja, R.. , Thanappan, Subash. , S., Ramya. Development of Digital Twin Technology in Hydraulics Based on Simulating and Enhancing System Performance. Journal of Cybersecurity and Information Management, vol. , no. , 2024, pp. 50-65. DOI: https://doi.org/10.54216/JCIM.130204
15. You, J., Li, D., Okumura, M., & Suzuki, K. (2022, October). Jpg-jointly learn to align: Automated disease prediction and radiology report generation. In *Proceedings of the 29th international conference on computational linguistics* (pp. 5989-6001).
16. Zhang, J., Li, B. and Zhou, S., 2025. Hierarchical modeling for medical visual question answering with cross-attention fusion. *Applied Sciences*, *15*(9), p.4712.
17. Allaberdiev, S., Chen, X., Mamatov, D., Abdusalomov, A., Ubaydullaev, U. and Khan, A., 2026. Radiology Report Generation via Self-Critical Multisource Knowledge Reasoning With Large Language Models. *IET Software*, *2026*(1), p.5529107.

18. You, J., Li, D., Okumura, M. and Suzuki, K., 2022, October. Jpg-jointly learn to align: Automated disease prediction and radiology report generation. In *Proceedings of the 29th international conference on computational linguistics* (pp. 5989-6001).
19. Mondal, C., Kai, C.K., Pham, D.S., Tan, T., Gedeon, T. and Gupta, A., 2025, December. Generating Clinically Relevant Reports from Chest X-Rays for Cardiomegaly Diagnosis. In *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1-8). IEEE.
20. Hartsock, I., Koutsoubis, N., Ahmed, S., Parker, N., Schabath, M.B., Araujo, C., Qayyum, A., Lam, C., Gatenby, R.A. and Rasool, G., 2026. Clinical AI in Radiology: Foundations, Trends, Applications, and Emerging Directions. *Cancers*, *18*(6), p.942.