Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 12 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i12.56289

PAPER

Enhancing Automated Medical Report Generation: A Method Based on Semantic-Guidance and Dual Stage Alignment

Fatima Cheddi[1] (✉), **Ahmed Habbani**[1], **Hammadi Nait-Charif**[2]

[1]Smart Systems Laboratory ENSIAS, Mohammed V University in Rabat, Rabat, Morocco

[2]National Center for Computer Animation, Bournemouth University, Poole, UK

cheddi_fatima@um5.ac.ma

ABSTRACT

The increased availability of multimodal data in healthcare, particularly in clinical diagnosis, can improve diagnostic accuracy, patient outcomes, and support more effective clinical decision-making. However, previous methods face several challenges, including achieving effective cross-modal alignment between textual descriptions and visual data, missing small and rare lesions, imprecise diagnostic terminology, and difficulty in extracting and utilizing semantic knowledge. To address these issues, we propose a new framework named **s**emantic-guided **h**ierarchical feature **e**xtraction and cycle-**co**nsistent **f**usion **(SHECoF)** for automatic chest X-ray (CXR) report generation, based on supervised and unsupervised learning algorithms. Our model introduces a novel dual-alignment strategy to progressively bridge the modality gap. It first incorporates hierarchical feature extraction and semantic knowledge extraction (SKE) mechanisms from the report, guiding the model to focus on fine-grained lesion detection in the visual extraction process. Subsequently, a second, deep alignment is performed by our cycle-consistent cross-attention fusion (C3F) mechanism, which enforces a bidirectional, cycle-consistent loss, establishing a fine-grained correspondence between image regions and textual descriptions. Validation of our approach in comparisons with existing methods indicates a corresponding boost in report quality in terms of clinical accuracy of the description, localization of the lesion, and contextual consistency, positioning our framework as a robust tool for generating more accurate and reliable medical reports.

KEYWORDS

medical image, automated report generation, cross-modal, deep learning, contrastive learning, semantic knowledge

1 INTRODUCTION

Diagnostic imaging such as chest X-rays (CXR), computed tomography (CT), and magnetic resonance imaging (MRI) are essential in modern healthcare and provide crucial

Cheddi, F., Habbani, A., Nait-Charif, H. (2025). Enhancing Automated Medical Report Generation: A Method Based on Semantic-Guidance and Dual Stage Alignment. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(12), pp. 42–62. https://doi.org/10.3991/ijoe.v21i12.56289

Article submitted 2025-05-06. Revision uploaded 2025-08-11. Final acceptance 2025-08-11.

analysis for diagnosis and detection of diseases. The diagnostic report is a structured summary of the findings and analysis of the medical images and acts as a formal way to exchange information between radiologists and doctors [1]. However, analyzing these images and transforming them into detailed reports is a complex process, and radiologists have to conduct manual analysis of photographs and write detailed reports, which is labor-intensive, time-consuming, and prone to errors. Indeed, studies on clinical workflows have shown that replacing such manual systems leads to statistically significant improvements in report turnaround time and accuracy [15]. With the evolution of technology in the medical field and the increasing complexity of diagnostic imaging, the automated generation of diagnostic medical reports has become an attractive solution to these difficulties [2]. This solution enhances workflow efficiency, reduces the workload on radiologists, minimizes diagnostic errors, and improves the efficiency of patient treatment. Recently, deep learning has garnered increasing interest from researchers in generating reports from medical images, which has demonstrated remarkable capabilities.

Various existing methods primarily adopt the image captioning model [3]-[4]-[5] based on encoder-decoder frameworks to extract and describe disease features. Furthermore, enhancements such as transformers and attention mechanisms have improved the quality of generated reports [6–7] by capturing long-range dependencies and reasoning over the relevant region of the given image. However, they still face challenges. They often produce short phrases rather than cohesive reports and suffer from data bias, where common conditions are overrepresented while rare pathologies appear infrequently.

Recently, the increasing availability of the use of multimodal data in image analysis has created new opportunities for automating the generation of medical reports. Current models rely on multi-modal learning techniques that jointly process both imaging data and textual information [8]. These models perform a comprehensive analysis of medical imaging features alongside their associated clinical narratives, enabling a more thorough interpretation of radiological findings. Through the simultaneous processing of visual patterns and language representations. The combination of these modalities enables the models to effectively leverage both the nuanced features of the medical images and the diagnostic context from the associated text. However, while the multimodal fusion-based approaches have demonstrated some success [9] [10] [11] [12], they still have several limitations, including the failure to bridge the semantic and granularity gaps between modalities, resulting in imperfect alignment between textual descriptions and visual data, missing small or rare lesions, imprecise diagnostic terminology, and difficulty extracting and utilizing semantic knowledge.

To overcome these challenges, we introduce a new model based on semantic-guided hierarchical feature extraction and Cycle-Consistent Cross-Attention fusion (C3F)—SHeCoF—which directly uses important signals and features to advance automated medical reporting. Unlike traditional methods focusing on global visual and textual features, this model emphasizes sensitivity to crucial lesion locations, multi-scale characterization, and fine-grained alignment, thereby improving the quality and reliability of the generated report.

Our approach addresses the persistent challenges of aligning visual and textual data, detecting small and rare lesions, and ensuring the precision of diagnostic terminology. It uses structured semantic knowledge from medical reports. At the core of our framework lies the semantic-guided hierarchical text feature extraction (SHTFE) mechanism, which consists of two subparts: the hierarchical text feature extraction (HTFE), capturing multi-level text features at the report, sentence, and word levels. Semantic knowledge extraction (SKE) is a sub-part that identifies and encodes structured medical knowledge from historical reports, aiding the model in contextualizing lesion descriptions and refining diagnostic terminology. This mechanism leverages

structured semantic knowledge extracted from medical reports to guide the visual extractor to focus on fine-grained lesion details related to the text. Additionally, the C3F module establishes a multi-level alignment between visual and textual features, ensuring a precise correspondence between image regions and their diagnostic descriptions. To further refine the generated reports, we introduce a semantic self-correction mechanism in the decoder module, which iteratively evaluates and corrects semantic inconsistencies, ensuring both clinical accuracy and linguistic fluency. Experimental results demonstrate significant improvements in lesion localization, diagnostic precision, and contextual coherence, underscoring the effectiveness of our framework in generating high-quality medical reports. This positions our approach as a valuable tool for enhancing diagnostic workflows and improving patient outcomes. In summary, this paper presents three main contributions:

- Our first contribution is a trainable architecture for radiology report generation: we present a new paradigm that greatly enhances multimodal alignment and semantic consistency, addressing the problems with current approaches in producing correct and clinically relevant diagnostic descriptions.
- The second contribution is: enhancing lesion detection and diagnostic precision: by incorporating the SHTFE mechanism and C3F module, our framework achieves superior performance in detecting small and rare lesions while ensuring precise diagnostic terminology and overcoming the challenges of previous approaches.
- The last contribution: Demonstrating superior performance on benchmark datasets: experimental results indicate that our framework achieves superior performance on publicly available datasets, with significant improvements in lesion localization, diagnostic precision, and contextual coherence. For instance, our model achieves a BLEU-4 score of 0,214 and 0.430 in ROUGE-L on the IU X-RAY dataset, demonstrating a notable improvement compared to the baseline, delivering superior results, particularly in handling complex cases with multiple lesions.

2 RELATED WORK

The automatic generation of medical reports based on multimodal data has recently attracted much attention. In this field, diverse works are introduced to tackle key challenges including cross-modal alignment, reducing data bias, and improving clinical relevance. This section shows an overview of related work along four key areas, including (1) image captioning, (2) multi-modal fusion, and (3) semantic knowledge integration.

2.1 Image captioning based radiology report generation

Image captioning is one of the essential tasks that translate visual features into a textual description. Previous approaches in the automated report generation domain were inspired by natural image captioning methods [3]–[14], that exhibit RNN recurrent neural networks and attention-based methods to obtain descriptive captions. These techniques were later adapted to medical imaging, focusing on domain-specific challenges such as terminology consistency.

Recent papers introduce transformer-based models [6] and hierarchical LSTMs [16] to improve textual coherence in radiology reports. Transformers have transformed the practice of natural language processing, and as CNN-Transformer hybrid models, they are being widely used in this process. Chen et al. [17] developed a memory-driven

Transformer model, where relational memory modules store and continually refine important information while the report is being generated. This approach provides consistent medical terminology for basic clinical concepts and aids in improving the alignment of images and text. Wang et al. [18] applied common vision transformers (ViTs) for key encoders and decoders, specifically integrating graphical and textual matching targets into the transformer structure. It allows the model to learn feature dependencies and make feature-wise associations to discriminate similarities between the images. To enhance the weight of disease regions, Tanida et al. [33] introduced a new method named RGRG based on anatomical regions in medical images. This model extracts more precise features with the use of object detection.

2.2 Multimodal fusion-based method for report generation

Medical report generation methods are based on the extraction of visual features of the medical images. However, utilizing only one data modality does not provide enough information for some complex diseases. In order to overcome this limitation, integrating prior knowledge and multimodal fusion techniques has emerged as a promising approach to improve report accuracy and completeness.

In recent years, several studies have explored the use of multimodal data fusion to improve diagnostic accuracy and report coherence. Traditional fusion models relied on simple concatenation of features extracted from CNN-based encoders [21]. However, recent advancements focus on attention-based fusion mechanisms, which dynamically match textual and visual representations to improve interpretability [22]. Moreover, this work [19] demonstrated significant improvements in automated report generation by using multimodal data, including medical images with corresponding reports, showing the potential of multimodal approaches. Additionally, the authors in this work [20] further extended this concept to complex disease diagnosis, proving that leveraging multiple data modalities enhances the quality and depth of diagnostic insights. The method proposed by Tang et al. [23] addresses critical challenges in medical report generation by introducing a “locate then generate” pattern named CAT. This framework employs a multi-modality encoder and a dual-stream decoder to dynamically integrate retrieved terminologies and preceding sentences. Similarly, the approach introduced by Iqbal et al. [24] is an adaptive, multi-modal technique which leverages wisdom learning to extract medical insights from reports and employs cross-modal coherence to align semantic information across images, disease labels, and reports. Additionally, knowledge graphs have been incorporated to provide structured clinical information, bridging the semantic gap between different modalities [25]. In this context, contrastive learning frameworks such as Contrastive Attention Networks [26] align image embeddings with textual descriptions, enhancing the specificity and accuracy of generated reports. Further, multi-level contrastive learning strategies have been proposed to align global and local semantic characteristics among different modalities [12]. Cheddi et al. proposed RG-MASR [31], a framework that combines cross-modal alignment with self-refinement to improve report accuracy. While their approach integrates visual and textual features through hierarchical attention mechanisms. Furthermore, the context-enhanced framework presented by Li [32] et al. integrates multimodal contexts, including medical knowledge bases, diagnostic findings, and clinical text, to enrich the report generation process. However, the trend towards using general-purpose LLMs has shown mixed results. Elgayar et al. [13] found that in some cases, domain-specific models outperformed generalized models such as ChatGPT for radiology reporting. This underscores the need for specialized architectures, which integrate domain-specific semantic knowledge directly into the generation process.

While some frameworks also focus on refining multimodal representations, our model that integrates the Semantic-Guided Hierarchical Feature Extraction, Dual alignment and C3F modules introduce a key distinction. Unlike the multi-step 'Unify, Align and Refine' process in [12], our C3F module enforces a direct, bidirectional correspondence simultaneously at multiple hierarchical levels. Furthermore, our integration of a cycle-consistency loss within the fusion module itself provides a stronger training signal for preserving unimodal information compared to the post-hoc self-refinement decoder.

2.3 Semantic knowledge-based medical reports

Since medical reports contain domain-specific terminology, several studies have introduced a priori knowledge and template-based strategies ensuring both medical accuracy and structural consistency. Yuan et al. [27] proposed an approach for merging multiple images with enriching them with clinical concepts, integrating domain-specific knowledge to enhance report quality. The approach introduced by Alfarghaly et al. [28] proposed a work to represent the semantics of text by calculating the weighted product between predicted tags and pre-trained word embedding to improve feature integration by merging visual and semantic features within a multimodal decoder, thereby boosting the model's performance. Yang et al. [29] presented a network composite of three branch (TriNet) for automated report generation, integrating visual attention, report embeddings, and subject headings to eliminate the visual-to-semantic gap. The framework first combines these embeddings and decodes them into accurate medical reports. Furthermore, the framework was created by Liu et al. [30], It incorporates knowledge graphs to better match identified abnormalities with organized clinical knowledge. However, the caliber of the specified template database has a significant impact on how well retrieval-based models work. However, existing methods often fail to extract fine-grained semantic knowledge and align with visual features appropriately, resulting in imprecise pathologic diagnosis terms.

Despite significant advancements in radiology report generation, there are still several limitations in the mentioned methods. The attention-based fusion methods fail to address the hierarchical relationships between localized visual findings and their medical semantics, often relying on static feature concatenation or coarse-grained attention. Additionally, current knowledge-aware frameworks treat medical concepts as discrete entities, struggling with atypical presentations and failing to adapt to patient-specific contexts, such as comorbidities. Furthermore, static templates lead to semantic drift, with terminology not dynamically refined based on evolving visual evidence. Most existing methods also overlook the radiologist-such as reasoning process of iterative hypothesis refinement, treating report generation as a single-step task and resulting in inconsistent prioritization of critical findings. These limitations highlight the need for a more robust framework, motivating our proposed approach, which introduces innovations such as semantic and hierarchical text feature extraction dynamic multimodal fusion, context-aware semantic propagation, and **self-correction** mechanism. These advancements address gaps in alignment, knowledge integration, and clinical workflow, ultimately providing more clinically actionable and coherent descriptions.

3 PROPOSED METHOD

The overview of our proposed model for enhancing the generation of radiology reports SHECoF is illustrated in Figure 1 and comprises four major modules: Semantic

and SHTFE, which includes Hierarchical Text Feature HTFE and semantic-guided knowledge (SGK); the second module is the visual feature extraction module (VFE); the third module is the C3F module; and the end module is a dynamic hierarchical decoder (DHD).

First, the input medical report is processed by the Hierarchical Text Feature Extraction (HTFE) module to extract rich-grained features from the radiology reports through hierarchical attention networks. Subsequently, The SGK module is designed to extract relevant semantic knowledge from the fine-grained features obtained by HTFE, by categorizing similar information into groups that have the same characteristics. Moreover, the VFE serves to extract visual features from radiology images, guided by semantic output obtained by SGK, to ensure that only the important text-based features are captured. The features extracted by HTFE and VFE then feed into the C3F module, which focuses on aligning semantic association between the extracted fine-grained visual and corresponding multi-level textual features, ensuring cohesive multi-modal representations. Finally, the DHD module integrates the aligned multi-modal features to generate, in a hierarchical process, a coherent and contextualized radiology report.

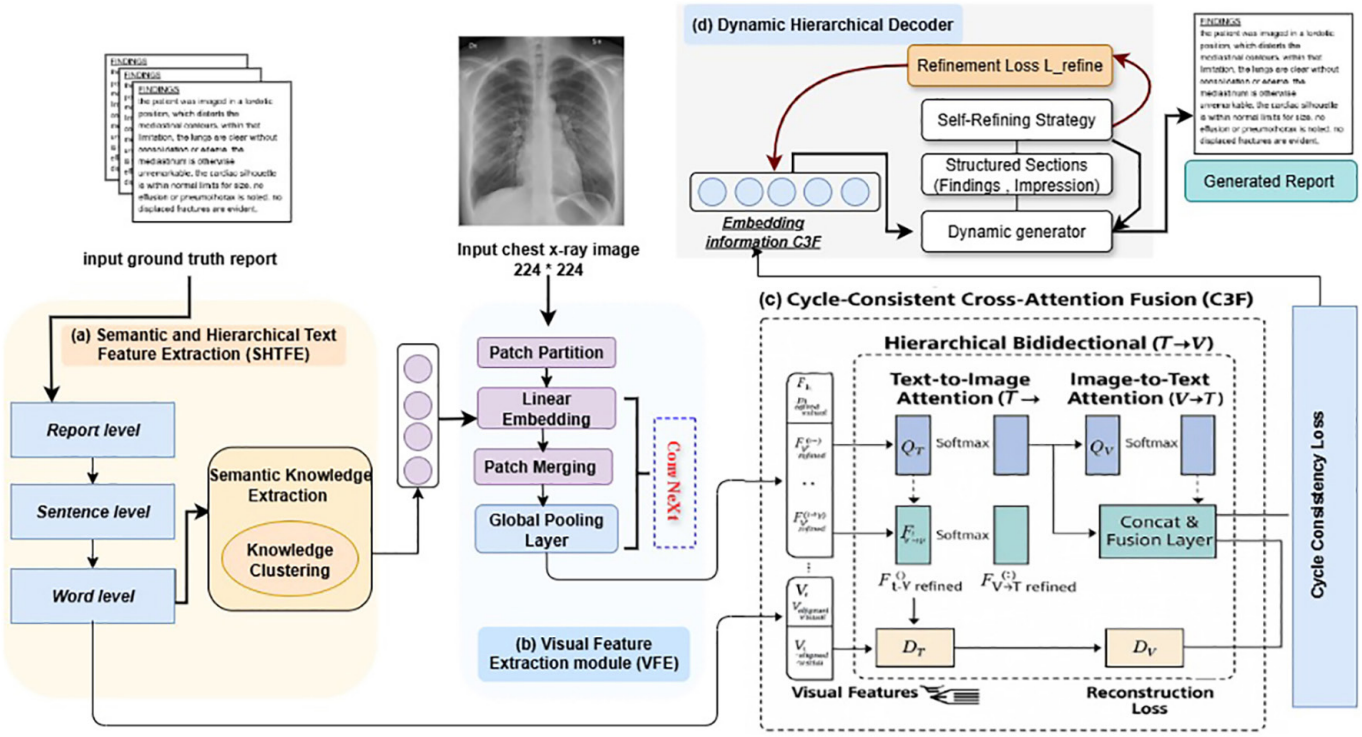


Fig. 1. Overview of the proposed SHECoF model

3.1 Semantic and hierarchical text feature extraction

Hierarchical text feature extraction. To extract rich, context-aware features from the medical reports, we use **ClinicalBERT**, a Transformer model pre-trained on clinical notations. Producing representations from a single, effective run through the model at the report, sentence, and word levels.

The process begins by tokenizing the entire medical report R. the sequence of tokens is fed into **ClinicalBERT model**. which outputs a contextualized embedding for each token. These fine-grained embeddings serve as the foundational **word-level representations $\mathbf{W}_{\mathbf{emb}}$**.

We build higher-level features from these fine-grained embeddings: a global report-level representation (R_{emb}) is derived from the output of the special [*CLS*] token, while sentence-level representations (S_{emb}) are derived by pooling the embeddings of their constituent tokens. By using a hierarchical approach and applying word- and sentence-level attention mechanisms, the model is able to link the textual findings to the corresponding medical image and concentrate on the most diagnostically important information, such as the words "nodule" or "opacity." This process can be summarized with the following equations. Let the input report R be tokenized into a sequence $T = \{t_{[CLS]}, t_1, t_2, ..., t_N\}$.

- **Word-level feature extraction: ClinicalBERT** processes the token sequence to generate a contextual embedding h_i for each token t_i. These embeddings are the word-level features.

$$H = \{h_{[CLS]}, h_1, h_2, ..., h_N\} = ClinicalBERT(T) \quad (1)$$

  where $h_i \in R^d$ and **d** is the dimensionality of the embedding (in this case d = 768).

- **Report-level representation:** The embedding for the entire report is represented by the hidden state of the [*CLS*] token.

$$R_{emb} = h_{[CLS]} \quad (2)$$

- **Sentence-level representation:** For a sentence s_i consisting of tokens $T\{t_k, ..., t_1\}$, its embedding is generated by pooling the corresponding token embeddings.

$$S_{emb_i} = Pool(\{h_j | t_j \in s_i\}) \quad (3)$$

Semantic guided knowledge extraction. The Semantic Guided Knowledge Extraction module is intended to automatically identify and classify reports that share semantic similarities after feature extraction. To do this, the set of report-level embeddings (R_{emb}) generated by the HTFE module is subjected to the K-Means clustering algorithm. By minimizing the within-cluster sum of squares, also referred to as inertia, K-Means seeks to divide a set of N report embeddings $\{R_{emb1}, R_{emb2}, ..., R_{embN}\}$ into K separate clusters, $C = \{C_1, C_2, ..., C_K\}$. Finding the group of clusters C that minimizes:

$$\underset{C}{\operatorname{argmin}} \sum_{j=1}^{K} \sum_{R_{embi} \in C_j} \| R_{embi} - \mu_j \|^2 \quad (4)$$

where μ_j is the centroid (mean vector) of all embeddings assigned to cluster C_j. The algorithm works iteratively: it begins by initializing K random centroids and then alternates between two steps until convergence. First, it assigns each report embedding to the cluster with the nearest centroid (based on Euclidean distance). Second, it recalculates each centroid as the means of all embeddings newly assigned to its cluster. The final output is K groups of reports, where each cluster contains reports with shared semantic features, such as similar diagnostic terminology (e.g., "pleural effusion," "cardiomegaly") or writing styles. This process effectively organizes the unstructured textual data into meaningful, data-driven categories.

3.2 Visual feature extraction module

To mitigate the problem of data bias and perform well on the spatial and semantic characteristics of the disease lesions, we use a multi-level feature extraction method guided by the prior knowledge (semantic clusters) obtained from the previous

module to align the visual regions and clinically relevant text-derived features, and also to guide the model's attention toward focal regions based on their semantic relevance through cluster-specific importance scores.

Given the input CXR images $X = \{x_1, x_2, x_3, ..., x_n\}$, where each image x is represented as a tensor in $R^{C \times H \times W}$, with C representing the number of channels and the H and W represent spatial height and width respectively. while H = W = 224 for CXR images. The feature extraction process begins by passing the images through a ConvNeXt [38] encoder as the backbone network. ConvNeXt is a modern, purely convolutional architecture designed with principles from VTs, enabling it to achieve state-of-the-art performance. It processes images through several hierarchical stages, each containing a series of ConvNeXt blocks that utilize key components such as depthwise convolutions and layer normalization. This structure allows the model to progressively downsample the input and build a rich, multi-scale feature representation.

This structure allows the model to progressively downsample the input and build a rich, multi-scale feature representation. From this, we extract three key feature maps, $L1$, $L2$, and $L3$, at resolutions of 56×56, 28×28, and 14×14, respectively. This process yields both localized, fine-grained features from the higher-resolution maps and global, contextual information from the lower-resolution map.

To achieve an initial semantic alignment between the visual and textual modalities, we employ a guided attention mechanism that uses the previously derived text clusters. For each cluster Tk, we first generate a **semantic query** q_k by projecting its centroid vector μ_k through a trainable linear layer W_{proj}, as defined by $q_k = W_{proj}(\mu_k)$. This query is then used to generate a spatial attention map $A_k^{(l)}$ for **each hierarchical level** $l \in \{1, 2, 3\}$, highlighting the image regions at that scale that are most relevant to the semantic concept:

$$A_k^{(l)} = \text{Softmax}\left(\frac{(q_k W_q)(L^{(l)} W_k)}{\sqrt{d}}\right)^{\top} \tag{5}$$

Each feature map $L^{(l)}$ is then refined by an element-wise multiplication with its corresponding attention map to produce the semantically guided features, which we will still refer to as $L^{(l)}$ for simplicity:

$$\boldsymbol{L}_{\boldsymbol{aligned}}^{(\boldsymbol{l})} = \boldsymbol{L}^{(\boldsymbol{l})} \odot \boldsymbol{A}_{\boldsymbol{k}}^{(\boldsymbol{l})} \tag{6}$$

3.3 Cycle-consistent cross attention fusion

The C3F module is intended to accomplish a more thorough and detailed integration of the multimodal features, building on the initial semantic guidance used in the VFE module. Using a potent, bidirectional cross-attention mechanism, the C3F module enhances the foundational alignment that the VFE module provides by concentrating the visual features on semantically relevant regions. It uses a cycle-consistency loss to enforce coherence and works across several hierarchical levels to create a fine-grained, contextually grounded correspondence between the two modalities.

Hierarchical bidirectional cross-attention. The core of this module is a bidirectional cross-attention mechanism applied at each level of the feature hierarchy. Let $\boldsymbol{F}_{\boldsymbol{aligned}_{\boldsymbol{visual}}}^{(\boldsymbol{l})} \in \boldsymbol{R}^{\boldsymbol{d}_l \times \boldsymbol{n}_v}$ denote the initially aligned visual features from the VFE module at level l, and let $F_{text} \in R^{d \times n_t}$ denote the textual features. Both are projected to a common dimension d and augmented with positional encodings. This second stage of deep fusion consists of two symmetrical attention branches:

1. Text-to-Image Attention (T → V): This branch uses text features as the query to further enrich the already-aligned visual features with explicit textual context.

Let $Q_T = \boldsymbol{W_Q F^{pos}_{text}}, \boldsymbol{K^{(1)}_V} = W_k \boldsymbol{F^{pos,(1)}_{aligned_{visual}}}, \boldsymbol{V^{(1)}_V} = W_v \boldsymbol{F^{pos,(1)}_{aligned_{visual}}}$. The refined aligned features are:

$$A^{(l)}_{V\to T} = \text{Softmax}\left(\frac{Q^l_V (K_T)^T}{\sqrt{d}}\right) \tag{7}$$

$$F^{\boldsymbol{refined},(1)}_{V\to T} = A^{(1)}_{V\to T}.(\boldsymbol{V})_T \tag{8}$$

The deeply aligned features from both branches at each level, $\boldsymbol{F^{refined,(1)}_{T\to V}}$ and $\boldsymbol{F^{refined,(1)}_{V\to T}}$ are then concatenated and passed through a final fusion layer to produce the level-specific fused representation, $\boldsymbol{F^{(1)}_{C3F}}$ The final comprehensive multimodal representation, F_{C3F}, is the aggregation of these features from all levels.

Cycle consistency loss. We introduce a cycle-consistency objective to guarantee that the fused representation F_{C3F} preserves the crucial information from both modalities. To reconstruct the original unimodal features from the fused representation, we use two lightweight decoders, D_T and D_V (implemented as Multi-Layer Perceptrons):

$$F'_{text} = D_T(F_{C3F}), F'_{visual} = D_V(F_{C3F}) \tag{9}$$

A reconstruction loss, L_{cycle}, is then computed as the $L2$ norm between the reconstructed features and the original features (prior to any alignment or positional encoding). This loss penalizes any deviation from the original semantic content.

$$L_{cycle} = \left\|F_{text} - F'_{text}\right\|^2_2 + \left\|F_{visual} - F'_{visual}\right\|^2_2 \tag{10}$$

By incorporating L_{cycle} into the main training objective, the model is encouraged to learn a shared embedding space where semantic concepts are symmetrically represented across both the image and text domains.

3.4 Dynamic hierarchical decoder

Report generation decoders face critical limitations, such as static and uniform templates, neglecting the hierarchical nature of the fused features, and data biases that focus on the global features and ignore the fine-grained details. These issues are worsened by the absence of dynamic mechanisms and feedback loops for refinement. For this, we propose a new DHD to build on multi-level features, dynamically adapt to the complexity of input data, and integrate feedback-driven refinement to generate accurate and professional reports.

To construct the report in a hierarchical process, we employ a transformer as the basic encoder-decoder module to process the visual sequence features hierarchically based on fused global, regional, and fine-grained details F_{CF3}. A key innovation is that instead of using a single flattened feature vector, the decoder's cross-attention mechanism **simultaneously attends to all levels of the fused features** $\left(F^{(1)}_{C3F}, F^{(2)}_{C3F}, \ldots\right)$. This allows the model to dynamically draw upon global, regional, or fine-grained information as needed to generate each word, seamlessly integrating high-level context with specific lesion details.

to balance the description of particular abnormalities with the generation of standard, template-based phrases (e.g., "The lungs are clear"). We introduce a dynamic

gating mechanism. The model uses the current decoder hidden state to predict a gating scalar at each decoding step t. This gate dynamically interpolates between two distinct vocabulary distributions: a findings vocabulary (P findings), which contains terms associated with pathologies, and a template vocabulary (P template), which contains common normal phrases. Controlled by the gating scalar $\alpha\, t$, the final word probability distribution is a weighted combination of the outputs from these two vocabularies. As a result, the model can seamlessly transition between outlining critical findings based on the fused multimodal evidence and describing normal anatomy.

To enhance the global semantic of the generated report, we incorporate **self-correction loop**. After an initial report Y' is generated, it is re-encoded using our original text encoder (ClinicalBERT) to produce a feature representation, $F_{Y'}$. We then compute a refinement loss, L_{refine}, which minimizes the distance between this representation and the original fused multimodal features F_{C3F}:

$$L_{refine} = \left\| F_{C3F} - F_{(Y')} \right\|_2^2 \tag{11}$$

The entire framework is trained end-to-end with a composite loss function, L_{total}, that combines the standard cross-entropy loss for token prediction (L_{XE}), the cycle-consistency loss from the fusion module (L_{cycle}), and the correction loss (L_{refine}):

$$L_{total} = \lambda_1 L_{XE} + \lambda_2 L_{cycle} + \lambda_3 L_{refine} \tag{12}$$

where λ_1, λ_2, and λ_3 are hyperparameters that balance the contribution of each objective.

4 EXPERIMENTS

IU X-ray [34]: a publicly available collection of CXR, includes 7,470 images paired with 3,955 corresponding medical reports, making it suitable for training models used in radiology report generation and computer-aided diagnosis. To ensure consistency with prior research, this collection is typically divided into 70% for training, 10% for validation, and 20% for testing.

4.1 Evaluation metrics

To evaluate the effectiveness of our proposed model, we use natural language generation (NLG) metrics. The BLEU [35] provides a score of n-gram overlap between the generated and ground truth texts, where a higher overlap indicates better text quality. METEOR [36] builds upon this approach and includes stemming and synonym matching based on WordNet, thereby better aligning it with human judgment. ROUGE-L [37] measures structural similarity by analyzing the length of the longest common subsequence between candidate and reference texts, highlighting fluency and coherence. Therefore, we used the CheXpert [42] tool to automatically extract the presence of 14 common radiological findings from both the generated and ground-truth reports in order to assess the accuracy of abnormality descriptions. Next, we treat this as a multi-label classification task, where each finding's labels are compared to the ground-truth labels in our generated reports. As a direct indicator of the model's clinical Efficacy metrics (CE metrics), we present the precision, recall, and F1-score from this comparison. A thorough evaluation of both linguistic quality and diagnostic accuracy is guaranteed by this dual evaluation method.

4.2 Implementation details

Our implementation follows a hierarchical multimodal architecture for medical report generation. pre-trained on ImageNet. It processes 224 × 224 resolution CXR through four hierarchical stages with depths of [3, 3, 9, 3] and feature dimensions of [96, 192, 384, 768], respectively. The text processing pipeline employs a hierarchical feature extraction by Clinical-BERT with separate hidden dimensions of 768 for the attention mechanism at the document, sentence, and word levels. The fusion module projects visual and textual features to a shared 128-dimensional space using linear transformations and applies cross-modal attention with 4 parallel heads. Our decoder consists of a 6-layer transformer model (512 hidden units, 8 attentions heads) with dynamic sentence-type regularization between the template and abnormal findings. Training employs a composite objective function with cross-entropy (weight = 1.0), contrastive (weight = 0.5, temperature = 0.07), and learned features refinement loss (weight = 0.5), optimized for Adam with a learning rate of 5×10^{-4}. We perform inference using beam search with width 3, using hyperparameters determined by ROUGE on the validation set from ranges $\tau \in [0.05, 0.2]$ for contrastive temperature and $\lambda \in [0.1, 0.9]$ for loss balancing. The system processes frontal and lateral chest radiographs when available. In the training stage the model was trained for 30 epochs, with each epoch representing a complete pass through the dataset.

5 EXPERIMENT RESULTS

5.1 Comparison experiment

We start by assessing our model's performance with the metrics mentioned below, as detailed in Table 1 and Figure 2, we compare our CHECoF model against several existing methods. including HRGR [39], R2GEN [17], PPKED [30], whereas multimodal-based models are ASGMD [41], context-enhanced framework [32], R2GenGPT [43] and Chen et al. [40]'s method. To assess clinical accuracy beyond linguistic fluency, we performed a clinical efficacy evaluation on the Indiana University CXR dataset. We employed the CheXpert labeler [42] to automatically extract 14 common thoracic findings from both the generated and ground-truth reports. By treating this as a multi-label classification task, we report the Precision, Recall, and F1-score to measure diagnostic accuracy. Table 2 illustrates the performance comparison of our model against several key baselines, including the methods from [3], [24], and R2GenGPT [43].

Table 1. Performance comparison of CHECoF model

Model	B-1	B-2	B-3	B-4	M.	R-L
HRGR	0.438	0.298	0.208	0.151	–	0.322
R2GEN	0.451	0.293	0.209	0.159	0.185	0.381
PPKED	0.483	0.315	0.224	0.168	–	0.376
ASGMD	0.489	0.326	0.232	0.173	0.206	0.397
C-Enhanced-F	0.491	0.359	0.263	0.209	0.212	0.408
R2GenGPT	0.405	–	–	0.093	0.123	0.325
Chen et al.	0.475	0.309	0.222	0.170	0.191	0.375
SHECoF(Ours)	**0.531**	**0.370**	**0.276**	**0.214**	**0.237**	**0.430**

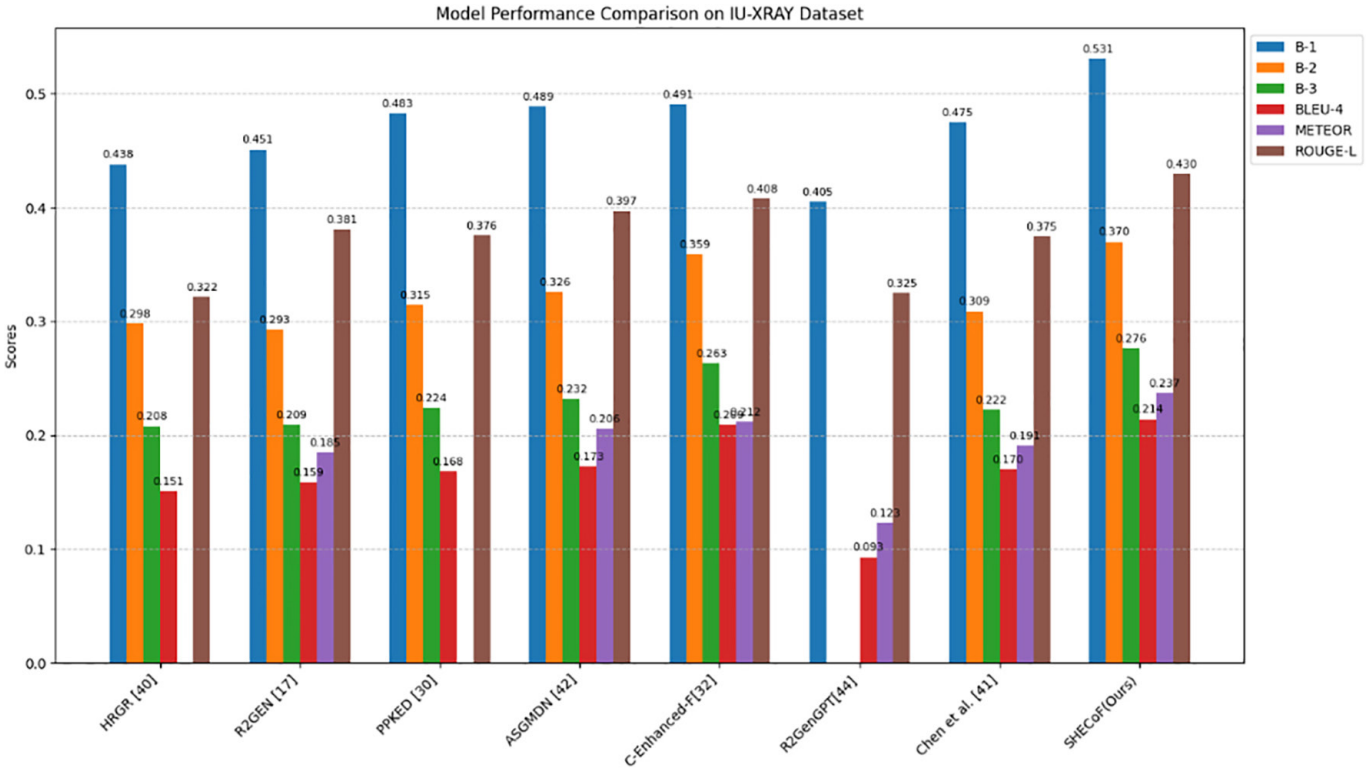


Fig. 2. Performance comparison of models on IU-XRAY dataset

Table 2. Clinical efficacy CE comparison on the IU X-ray dataset

Model	Precision	Recall	F1-Score
[3]	0.083	0.065	0.071
[24]	0.437	0.37	0.361
[43]	0.236	0.251	0.244
Ours	**0.453**	**0.363**	**0.382**

Our proposed model achieves state-of-the-art performance on the IU-Xray dataset, yielding significant improvement in BLEU-4, ROUGE-L, and METEOR scores. In particular, our model achieves BLEU-4 = 0.214, ROUGE-L = 0.430, and METEOR = 0.237, outperforming the runner-up **Context-enhanced Framework** by 0.5%, 2.5%, and 2.2%, respectively. These outcomes emphasize the effectiveness of context-sensitive fusing methodology that generates fluent and clinically accurate reports.

A notable observation is a marked improvement in BLEU-3 and BLEU-4 scores, implying that our method generates longer and more linguistically accurate n-grams. This improvement is derived from the better contextual alignment implemented in our framework to improve the semantic precision of the reports generated. This architecture demonstrates improved performance by utilizing hierarchical feature analysis and semantic-guided hierarchical feature fusion.

Overall, our findings highlight that integrating contextual knowledge and enhanced feature extraction significantly improves linguistic fluency, factual correctness, and clinical relevance, positioning our method as a leading approach for automated radiology report generation.

To validate our model's clinical efficacy, we benchmarked its performance on the Indiana University CXR dataset against several baseline methods, including the

models from [3] and [24]. As presented in the results in Table 2, our framework achieves a superior F1-score of 0.382, significantly outperforming both the weak baseline [3] (F1-score of 0.071) and the strong competitive method [24] (F1-score of 0.361). This enhanced performance is primarily driven by a notable increase in precision to 0.453 over model [24]. Although this is accompanied by a slight decrease in recall, the resulting higher F1-score demonstrates that our model achieves a more effective balance, indicating its ability to generate more reliable clinical findings.

5.2 Ablation experiment

In this section, we will conduct an ablation experiment to indicate the influence of each module on our model on overall performance. The experiments demonstrate the influence of the various components when they are integrated in the baseline model, to quantify the gain of each module (HTFE, VFE, C3F, and DHD) and the combination to final accuracy. Table 3 provides a detailed comparison over multiple evaluation metrics including B-1, B-2, B-3, B-4, METEOR, and ROUGE-L, for the IU University dataset.

Table 3. Ablation experiment

Dataset	Model	B-1	B-2	B-3	B-4	M.	R-L
IU University	Baseline Model	0.409	0.306	0.221	0.173	0.184	0.353
	Baseline + SHTFE	0.419	0.312	0.229	0.178	0.189	0.361
	Baseline + VFE	0.435	0.323	0.238	0.186	0.193	0.374
	Baseline + C3F	0.462	0.339	0.250	0.196	0.201	0.395
	Baseline + DHD	0.420	0.325	0.223	0.178	0.189	0.356
	Full model	0.531	0.370	0.276	0.214	0.237	0.430

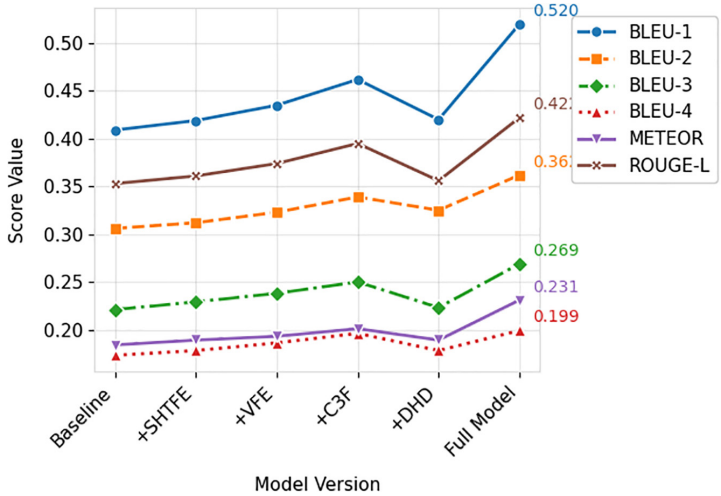


Fig. 3. Performance metrics of ablation study

As illustrated in Table 3 and Figure 3, the Full Model exhibits significant improvements over the Baseline Model, supporting the importance of every proposed module. Concretely, the Full Model surpasses the Baseline by 12.2% on B-1, 6.4% on B-2, 5.5% on B-3, 3.1% on B-4, 5.3% on METEOR, and 7.7% on ROUGE-L, suggesting the additional modules improved report generation to a significant degree. To systematically evaluate the effect of each component, we compared the following configurations.

BASELINE: The BASELINE model is a standard Transformer composite of three layers, eight attention heads, and 512 hidden units. This model serves as the starting point to evaluate the impact of additional modules.

+SHTFE (Hierarchical Text Feature Extraction): This configuration adds the Hierarchical Text Feature Extraction (HTFE) module to the baseline model. The HTFE module enhances the model's ability to capture hierarchical relations in the text, providing better context and structure for report generation. This module improves the overall feature extraction from text data. Adding HTFE to the BASE model makes all of the metrics better, especially BLEU-2, BLEU-3, and ROUGE. The model is better than the BASE model by 1.0% in BLEU-1, 0.6% in BLEU-2, and 0.8% in ROUGE-L.

+VFE (Semantic-Guided Visual Extractor): This configuration incorporates the Semantic-Guided Visual Extractor module to focus on fine-grained detection. This module guides VFE using semantic knowledge derived from medical reports. This ensures that the model focuses on clinically significant regions of the image, significantly improving the model's ability to generate precise and accurate medical reports. The performance boost is evident when comparing BASE+VFE to the BASE model, with significant improvements of 2.6% in B-1, 1.7% in B-2, 1.3% in B-4, 2.1% in ROUGE-L on the Indiana University dataset. This demonstrates that semantic-guided feature extraction significantly boosts the model's performance in producing precise and logically structured medical reports.

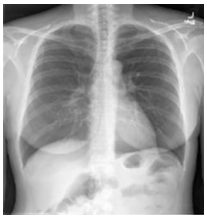
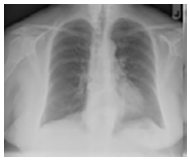
+C3F Cycle-Consistent Cross-Attention Fusion **module:** In this setup, the C3F module is integrated into the model. As shown in Table 3, this module improves visual-textual alignment by learning relationships between the two modalities. This leads to better cross-modal understanding, ensuring that both the visual and textual data are appropriately aligned, resulting in more accurate and coherent reports. +C3F outperforms the baseline by 5.3% in BLEU-1, 3.3% in BLEU-2, 2.9% in BLEU-3, 2.3% in BLEU-4, 1.7 in METEOR and 4.2% in ROUGE-L on the Indiana University dataset. This result underscores that the robust, bidirectional alignment of visual and textual features is the most critical factor in improving the quality of the generated report.

+Dynamic Hierarchical Decoder (+DHD): We add a mechanism that hierarchically processes the visual sequence features according to fused global, regional and fine-grained information and embeds the **Self-Correction loop** strategy module. This configuration substantially outperforms BASE, with margins 1.1% in BLEU-1, 1.9% BLEU-2, 0.2% BLEU-3, 0.5% BLEU-4, 0.3% in ROUGE-L applied on IU dataset, resulting in a minimal improvement. This suggests that while the advanced decoder is powerful, its full potential can only be realized when it is fed high-quality, deeply fused features from the preceding modules.

Full Model: Integrating HTFE, SGVE, C3F, and DHD consistently outperforms all other configurations, indicating the synergistic effect of these modules. This model not only produces accurate reports but also clinically coherent ones. As indicated in Table 3, the model achieves superior performance across all metrics. On the IU X-Ray dataset, the model surpasses the baseline with scores of 12.2%, 6.4%, 5.5%, 3.1%, 5.3%, and 7.7% in B-1, B-2, B-3, B-4, METEOR, and ROUGE-L, respectively. Moreover, the ablation study demonstrates a drop in performance when any module is removed, highlighting the necessity of each of the components in the overall model. Confirming the effectiveness of the combination of HTFE, SGVE, C3F, and DHD modules working synergistically to achieve the most accurate, clinically meaningful, and coherent reports. HTFE and VFE are both helpful for fine-grained lesion detection, C3F can guarantee accurate cross-modal alignment, and DHD refines semantic consistency, improving the context and quality of the generated reports.

5.3 Visual analysis

Table 4. The visualization of our SHECoF model

Image	Ground Truth	Generated Report
	The heart size and pulmonary vascularity appear within normal limits. There is evidence of pleural effusion in the left lung. Mild consolidation is seen in the right lower lobe, which could be indicative of a possible infection. No signs of pneumothorax are observed. The diaphragm appears normal, and there are no fractures visible in the ribs.	Heart size and pulmonary vascularity are normal. Pleural effusion noted in the left lung There is mild consolidation in the right lower lobe. No pneumothorax detected No fractures visible in the ribs, and the diaphragm is normal.
	Cardiac silhouette is normal, and pulmonary vascularity appears within normal limits. Mild bilateral pleural effusion is noted. Small nodule detected in the right upper lobe. No significant atelectasis is seen. The diaphragm is intact, and there is no evidence of pneumothorax or fractures.	Cardiac silhouette and pulmonary vascularity are normal. Bilateral pleural effusion observed. Right upper lobe nodule is noted (benign features). No atelectasis is seen. The diaphragm appears intact, with no pneumothorax or fractures.
	Diffuse consolidation is seen in both lungs, especially in the middle and lower lobes, consistent with bilateral pneumonia. Moderate pleural effusion is noted on both sides. The heart size is within normal limits. No pneumothorax is visible. The diaphragm is intact, with no evidence of abnormality.	Diffuse consolidation in the middle and lower lobes, suggestive of bilateral pneumonia. Moderate pleural effusion is visible on both sides (moderate, R>L). The heart size appears normal No signs of pneumothorax. The diaphragm is intact. (no elevation)
	Heart size and pulmonary vascularity appear normal. There is consolidation in the right lower lobe, which is indicative of pneumonia. No pleural effusion is seen, and the lung fields are otherwise clear. The diaphragm is intact with no abnormality detected. The left lung appears normal without any signs of atelectasis or consolidation.	Heart size and pulmonary vascularity are normal. Consolidation in the right lower lobe, consistent with pneumonia. No signs of pleural effusion (confirmed). The diaphragm appears normal, and the left lung is clear with no abnormalities.

To assess the performance of SHECoF's model (see Table 4), a comparative analysis is conducted between the ground truth medical report and the generated texts. The generated reports are analyzed using a color-coded system: green font indicates precise alignment with the ground truth, yellow color signifies rephrased content with identical medical meanings, and black color indicates a lack of similarity or missing details. For the first testing CXR image, the generated report shows a high degree of resemblance in terms of organs and lesion description, signifying the model's capability in encapsulating the intricate relationship between multi-level features in both image and reports. This enables its hierarchical attention mechanisms, which focus on both global and fine-grained details. Additionally, the output report of the case also contains detailed diagnostic details, such as specific characteristics of the referred lesions, which can be attributed to the

Semantic-Guided Visual Extractor during hierarchical VFE. This guidance enables the model to offer importance on critical areas, improving the clinical relevance of the generated reports.

Specifically, the **SHECoF** framework demonstrates improved performance on two different datasets. This means higher accuracy and more meaningful reports with better metrics. The closeness to ground truth and the inclusion of more clinically relevant details represent the high quality of the SHECoF's performance. With multi-level feature extraction and semantic-guided alignment mechanisms, the method achieves accuracy on par with clinical expectations while also providing important explanations that facilitate diagnostic reasoning.

6 ERROR ANALYSIS

In this section, we present a qualitative error analysis of our SHECoF framework. By examining specific cases where the model generated incorrect or incomplete descriptions, we aim to identify systematic weaknesses and provide insights into the remaining challenges in automated radiology report generation.

Table 5. Failure cases of the SHECoF model

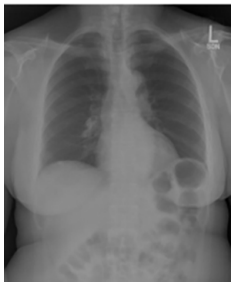
Image	Ground Truth	Generated Report	Error Type
	The heart is normal in size. The lungs are clear except for a **small nodule in the right upper lobe**. No pleural effusion	The lungs are clear	**False Negative** (Missed Finding)
	**Cardiomegaly and pulmonary vascular congestion** are noted. There is **no pleural effusion**	**Cardiomegaly and pulmonary vascular congestion** are present. **Small bilateral pleural effusions are also noted**	**False Positive** (Hallucination)

Table 5 presents qualitative error analysis led to the identification of two central failure modes: (1) the model is sensitive to missing findings that are too small or subtle in appearance. In the first case study, the system failed to report a small pulmonary nodule partially hidden by a blood vessel. We hypothesize that the lack of a strong visual signal led to this finding being missed in favor of a default "normal" class, in line with the previously described problem of perceiving low-contrast, fine-grained details. (2) The model produces "semantic hallucinations" of high-correlation findings. In the second case study, the model accurately predicted the presence of cardiomegaly but also included a description of the finding of pleural effusion, a frequent radiographic co-occurrence of cardiomegaly, in its report without any evidence in the visual input. This suggests a heavy reliance on statistical association from training data, rather than robust grounding of every individual claim to a visual feature.

7 DISCUSSION

In this work, we presented a novel framework for automated medical report generation that addresses key challenges in multi-modal alignment, fine-grained lesion detection, and semantic consistency. By integrating a semantic guided hierarchical feature extraction mechanism, a C3F, and a self-correction strategy in dynamic decoder module, our framework significantly improves the quality and accuracy of generated medical reports. Experimental results on Indiana University CXR dataset, show that our approach surpasses existing methods in terms of accuracy, lesion localization, and contextual coherence. For instance, our model achieves a B-4 score of 21.4%, particularly in handling complex cases with multiple lesions.

The Semantic and Hierarchical Text Feature Extraction mechanism plays a critical role in guiding the VFE process by leveraging structured semantic knowledge from medical reports. This allows the model to focus on fine-grained lesion details, addressing the challenge of detecting small or rare lesions that are often missed by traditional methods. The success of our framework is attributed to its innovative dual-alignment strategy, which progressively refines the correspondence between visual and textual features in two distinct stages. The first alignment occurs within the VFE module, where semantic priors derived from unsupervised text clustering are used to create a query that guides the visual encoder's attention. This initial, coarse alignment effectively "primes" the model, ensuring it focuses on clinically relevant image regions before any deep fusion occurs. This is followed by the second, deep alignment in the C3F module. Building on the semantically guided features, the C3F module employs a powerful bidirectional cross-attention mechanism and a cycle-consistency loss. This stage performs the intensive, fine-grained fusion, establishing a robust, contextually grounded link between specific visual findings and their descriptive text. Finally, the DHD module iteratively evaluates and refines the generated reports, ensuring both clinical accuracy and linguistic fluency.

Despite these advancements, our framework still has some limitations. The model's effectiveness largely depends on the quality and variety of the training data. While our approach performs well on publicly available datasets, its generalizability to other types of medical imaging such as MRI and CT scans or to datasets with different characteristics (e.g., rare diseases, underrepresented populations) remains to be validated. Second, the training process is computationally intensive, requiring significant resources. Future work could explore more efficient training strategies, such as parameter-sharing Siamese networks, to reduce computational overhead.

A direction for future work is to utilize pre-trained language models (e.g. BERT, GPT) as the decoder in order to enhance the fluency and cohesiveness of generated reports. Utilizing large language models that will be fine-tuned for generating medical reports may have substantial advantages as well. Moreover, the incorporation of knowledge graphs within the framework could better extract features through relationships between abnormal keywords, improving the quality and clinical significance of the produced reports.

7.1 Limitations and ethical considerations

A key limitation and ethical consideration of our work is the potential for bias propagation through the semantic guidance mechanism. Since the clusters are learned in an unsupervised manner from the training data, any demographic or disease prevalence biases present in the data could be encoded into the

cluster centroids. This could inadvertently cause the model to perform less effectively for underrepresented patient populations. Future work should include a thorough bias audit and explore fairness-aware clustering techniques to mitigate this risk.

Finally, our error analysis revealed instances of clinical hallucination. Although reduced, the generation of false-positive findings remains a critical safety concern that highlights the ongoing challenge of grounding every claim in direct visual evidence. Reducing this rate is paramount for clinical safety and remains a primary objective for future research.

8 CONCLUSION

In conclusion, our work represents a significant advance forward in the field of automating medical report generation by addressing key challenges. Through our contributions mentioned below, our approach generates reports that are both clinically accurate and contextually coherent, which makes it a potential tool for real-world clinical applications. Our focus in future research is to generalize the framework to increase its application potential, improve computational efficiency, and adopt advanced AI technologies.

9 REFERENCES

[1] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artificial Intelligence in Medicine*, vol. 106, p. 101878, 2020. https://doi.org/10.1016/j.artmed.2020.101878

[2] Y. Tang, D. Wang, L. Zhang, and Y. Yuan, "An efficient but effective writer: Diffusion-based semi-autoregressive transformer for automated radiology report generation," *Biomedical Signal Processing and Control*, vol. 88, Part B, p. 105651, 2024. https://doi.org/10.1016/j.bspc.2023.105651

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[4] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

[5] H. Zhang, C. Ma, Z. Jiang, and J. Lian, "Image caption generation using contextual information fusion with Bi-LSTM-s," *IEEE Access*, vol. 11, pp. 134–143, 2023. https://doi.org/10.1109/ACCESS.2022.3232508

[6] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[7] S. Wang, G. Fang, L. Liu, J. Wang, K. Zhu, and S. N. Melo, "Transformer based visual object tracking with global feature enhancement," *Appl. Sci.*, vol. 13, no. 23, p. 12712, 2023. https://doi.org/10.3390/app132312712

[8] X. Wang, G. Figueredo, R. Li, W. E. Zhang, W. Chen, and X. Chen, "A survey of deep learning-based radiology report generation using multimodal inputs," *Medical Image Analysis*, vol. 103, p. 103627, 2025. https://doi.org/10.1016/j.media.2025.103627

[9] G. V. Magalhães, R. L. De S Santos, L. H. S. Vogado, A. C. De Paiva, and P. De Alcântara Dos Santos Neto, "XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model," *Heliyon*, vol. 10, no. 7, p. e27516, 2024. https://doi.org/10.1016/j.heliyon.2024.e27516

[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, 2023. https://doi.org/10.1016/j.media.2023.102798

[11] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, 2022. https://doi.org/10.1109/JBHI.2022.3207502

[12] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, and Y. Zou, "Unify, align and refine: Multi-level semantic alignment for radiology report generation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 2851–2862. https://doi.org/10.1109/ICCV51070.2023.00268

[13] S. Elgayar, I. I. M. Manhrawy, A. M. Soliman, S. Hamad, and E.-S. M. Horbaty, "Exploring medical caption generation through OpenAI's ChatGPT-4 Model: A PRISMA review," *Int. J. Onl. Eng.*, vol. 21, no. 5, pp. 18–30, 2025. https://doi.org/10.3991/ijoe.v21i05.53529

[14] X. Kelvin *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML '15)*, vol. 37, 2015, pp. 2048–2057.

[15] P. Adikari *et al.*, "Improving report generation and delivery system of microbiological investigations at MRI – Sri Lanka with concern to turn-around-time, an intervention study," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 9, pp. 26–38, 2020. https://doi.org/10.3991/ijoe.v16i09.13799

[16] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, vol. 1, 2018, pp. 2577–2586. https://doi.org/10.18653/v1/P18-1240

[17] Z. Chen, Y. Song, T. H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," *arXiv preprint arXiv:2010.16056*, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.112

[18] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, "Automated radiographic report generation purely on transformer: A multicriteria supervised approach," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2803–2813, 2022. https://doi.org/10.1109/TMI.2022.3171661

[19] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, in Lecture Notes in Computer Science, M. de Bruijne et al., Eds., Springer, Cham, vol. 12903, 2021, pp. 72–82. https://doi.org/10.1007/978-3-030-87199-4_7

[20] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "RATCHET: Medical transformer for chest X-ray diagnosis and reporting," in *Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, in Lecture Notes in Computer Science, M. de Bruijne et al., Eds., Springer, Cham, vol. 12907, 2021, pp. 293–303. https://doi.org/10.1007/978-3-030-87234-2_28

[21] X. Huang, F. Yan, W. Xu, and M. Li, "Multi-attention and incorporating background information model for chest X-ray image report generation," *IEEE Access*, vol. 7, pp. 154808–154817, 2019. https://doi.org/10.1109/ACCESS.2019.2947134

[22] Z. Wang, L. Zhou, L. Wang, and X. Li, "A self-boosting framework for automated radiographic report generation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 2433–2442. https://doi.org/10.1109/CVPR46437.2021.00246

[23] Y. Tang, Y. Yuan, F. Tao, and M. Tang, "Cross-Modal augmented transformer for automated medical report generation," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 13, pp. 33–48, 2025. https://doi.org/10.1109/JTEHM.2025.3536441

[24] S. Iqbal, A. N. Qureshi, F. Khan, K. Aurangzeb, and M. Azeem Akbar, "From data to diagnosis: Enhancing radiology reporting with clinical features encoding and cross-modal coherence," *IEEE Access*, vol. 12, pp. 127341–127356, 2024. https://doi.org/10.1109/ACCESS.2024.3449929

[25] Y. Zhang *et al.*, "When radiology report generation meets knowledge graph," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12910–12917, 2020. https://doi.org/10.1609/aaai.v34i07.6989

[26] F. Liu *et al.*, "Contrastive attention for automatic chest X-ray report generation," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* pp. 269–280, 2021. https://doi.org/10.18653/v1/2021.findings-acl.23

[27] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Medical Image Computing and Computer Assisted Intervention, (MICCAI 2019)*, in Lecture Notes in Computer Science, D. Shen et al., Eds., Springer, Cham, vol. 11769, 2019, pp. 721–729. https://doi.org/10.1007/978-3-030-32226-7_80

[28] 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. https://doi.org/10.1016/j.imu.2021.100557

[29] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint embedding of deep visual and semantic features for medical image report generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 167–178, 2023. https://doi.org/10.1109/TMM.2021.3122542

[30] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13753–13762. https://doi.org/10.1109/CVPR46437.2021.01354

[31] F. Cheddi, A. Habbani, and H. Nait-Charif, "Improving the CXR reports generation with multi-modal feature alignment and self-refining strategy," in *2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)*, Fez, Morocco, 2024, pp. 1–7. https://doi.org/10.1109/ESAI62891.2024.10913509

[32] H. Li, H. Wang, X. Sun, H. He, and J. Feng, "Context-enhanced framework for medical image report generation using multimodal contexts," *Knowledge-Based Systems*, vol. 310, p. 112913, 2025. https://doi.org/10.1016/j.knosys.2024.112913

[33] 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, 2023, pp. 7433–7442. https://doi.org/10.1109/CVPR52729.2023.00718

[34] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association: (JAMIA)*, vol. 23, no. 2, pp. 304–310, 2016. https://doi.org/10.1093/jamia/ocv080

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, Association for Computational Linguistics, USA, 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135

[36] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 376–380. https://doi.org/10.3115/v1/W14-3348

[37] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain. 2004, pp. 74–81.

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, and Z. Zhang, "ConvNeXt: A ConvNet for the vision transformer era," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5578–5588. https://doi.org/10.1109/CVPR52688.2022.00320

[39] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[40] W. Chen *et al.*, "Cross-modal causal intervention for medical report generation," *arXiv preprint arXiv:2303.09117*, 2023.

[41] Y. Xue, Y. Tan, L. Tan, J. Qin, and X. Xiang, "Generating radiology reports via auxiliary signal guidance and a memory-driven network," *Expert Systems with Applications*, vol. 237, p. 121260, 2024. https://doi.org/10.1016/j.eswa.2023.121260

[42] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *arXiv preprint arXiv:1901.07031*, 2025.

[43] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2gengpt: Radiology report generation with frozen llms," *Meta-Radiology*, vol. 1, no. 3, p. 100033, 2023. https://doi.org/10.1016/j.metrad.2023.100033

10 AUTHORS

Fatima Cheddi is a teacher in computer science at the Higher Technician Certificate (BTS) in Alhoceima, Morocco. Fatima is also a PhD student at Ensias School, University Mohammed V, Rabat, Morocco, specializing in medical imaging, deep learning, software quality, and educational engineering. She holds a master's degree in quality software and a second master's in pedagogical engineering and multimedia. Her research focuses on deep learning applications in medical imaging. Throughout her academic journey, she has contributed to several projects and academic papers (E-mail: cheddi_fatima@um5.ac.ma).

Dr. Ahmed Habbani is a full Professor in the Department of Communication Networks. He holds a Ph.D. in Applied Sciences from the LEC (Laboratory of Electronics and Communications) at Mohammedia School of Engineers (EMI), affiliated with Mohammed V University, Rabat, Morocco, and from the LISIF (Laboratory of Instruments and Systems of Île-de-France) at Pierre and Marie Curie University, France. His research interests focus on artificial intelligence, security in intelligent mobile systems, ad hoc sensor networks, communication networks and systems, the Internet of Things (IoT), and smart homes, grids, and cities.

Dr. Hammadi Nait-Charif is a Senior Lecturer at the National Centre for Computer Animation (NCCA), Bournemouth University, United Kingdom. He earned his PhD in Information and Computer Sciences from Chiba University, Japan. From 1998 to 2001, he served as a lecturer at the Ecole Supérieure de Technologie, Mohamed I University, Oujda, Morocco. In 1999, he was a Fulbright Visiting Assistant Professor at Michigan State University, USA. From October 2001 to June 2005, he worked as a postdoctoral research fellow at the School of Computing, University of Dundee, Scotland. His research interests include computer animation, computer vision, pattern recognition, and machine learning, with a focus on applications in medical imaging.