

Explainable Approach Using Semantic-Guided Alignment for Radiology Imaging Diagnosis

Fatima Cheddi¹, Ahmed Habbani², Hammadi Nait-Charif³

Smart Systems Laboratory ENSIAS, Mohammed V University in Rabat, Rabat, Morocco^{1,2}

National Center for Computer Animation, Bournemouth University, United Kingdom³

Abstract—The increased success of deep learning in the radiology imaging domain has significantly advanced automated diagnosis and report generation, aiming to enhance diagnostic precision and clinical decision-making. However, existing methods often struggle to achieve detailed morphological description, resulting in reports that provide only general information without precise clinical specifics and thus fail to meet the stringent interpretability requirements of medical diagnosis. Also, the critical need for transparency in clinical automated systems has catalyzed the emergence of explainable artificial intelligence (XAI) as an essential research frontier. To address these limitations, we propose an explainable system for report generation that leverages semantic-guided alignment and interpretable multimodal deep learning. Our model combines hierarchical semantic feature extraction from medical reports with fine-grained features that guide the model to focus on lesion-relevant visual features and use Concept Activation Vectors (CAVs) to explain how radiological concepts affect report generation. A contrastive multimodal fusion module aligning textual and visual modalities through hierarchical attention and contrastive learning. Finally, an integrated concept activation system that provides transparent explanations by quantifying how radiological concepts influence generated reports. 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—Automated report generation; explainable AI; cross-modal fusion; contrastive learning; semantic-guided alignment

I. INTRODUCTION

The rapid success of deep learning in the medical domain has created transformative possibilities for improving diagnostic accuracy and clinical decision-making, particularly through AI-powered automated report generation. This solution enhances workflow efficiency, reduces the workload on radiologists, minimizes diagnostic errors, and improves the efficiency of patient treatment [1]. However, the integration of artificial intelligence in radiology has reached a critical inflection point, where diagnostic accuracy must now be matched by clinical interpretability to enable real-world adoption [2].

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 [3]. 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.

Transforming detailed medical images into descriptive text necessitates the accurate detection of anatomical structures, alongside precise identification of lesions based on their morphology and location. Although automated report generation systems have shown significant progress in analyzing chest radiographs [4], their widespread clinical adoption remains limited due to several challenges. Among these, a major barrier is the semantic and granularity gap between imaging modalities and textual descriptions, which leads to imperfect alignment between visual data and generated reports. Additionally, these systems suffer from data bias, as common conditions are often overrepresented in training datasets, whereas rare pathologies are underrepresented, hindering the robustness and generalizability of the models. Furthermore, interpretability continues to be a major challenge in medical applications, as AI models often operate as 'black boxes.' The doctors emphasize the need for transparent models that can provide clear, understandable explanations for their predictions, an essential requirement in medical practice.

Clinical automated systems require transparency, which has led to the creation of explainable artificial intelligence (XAI) becoming an essential research topic [5]. The ability to interpret AI outcomes remains essential for validating diagnostic accuracy while building clinician trust and meeting regulatory requirements [6]. The semantic gap between modalities, the "black-box" nature of deep learning models, and the need for high-fidelity descriptions—motivate the central research question of this work: How can we design a radiology report generation framework that not only achieves high clinical accuracy through fine-grained, semantically-guided alignment between images and text, but also provides transparent, concept-based explanations for its diagnostic outputs to foster clinical trust and adoption? This question guides our development of a novel architecture that simultaneously optimizes for descriptive precision and true model interpretability.

To answer this question, we introduce a new model based on semantic-guided alignment and interpretable multimodal deep learning named Explainable Report Generation Semantic-guided Alignment—Ex-RGSA—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 and fine-grained alignment, thereby improving the quality and reliability of generated reports.

Our approach addresses the persistent challenges of aligning visual and textual data, detecting small and rare pathologies, and ensuring the precision of diagnostic terminology. It uses structured semantic knowledge from medical reports. At the start of our framework lies the Explainable Semantic Feature Extractor (X-SFE) module, which consists of two subparts: the Multi-level Text Feature Extraction, capturing multi-level text features at the report, sentence, and word levels. Semantic Guided Knowledge extraction 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. At the core of our framework lies the Semantic-Aligned Visual Extractor (X-SAVE) module. 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 Contrastive Multimodal Fusion module (CMF) links visual-textual features anchored by concepts through three stages (global anatomy to regional findings to lesion-specific details). To further refine the generated reports, we introduce a semantic self-refining mechanism in the Decoder module that uses CAV confidence scores to correct terminology errors.

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:

1) *First, our primary contribution:* is a novel trainable architecture for radiology report generation, introducing a new paradigm that significantly enhances multimodal alignment and semantic consistency. A key aspect of this architecture is its ability to improve the detection of small and rare lesions, further ensuring the diagnostic precision achieved by the Explainable Semantic Feature Extractor (X-SFE), the Semantic-Aligned Visual Extractor (X-SAVE), and the Multimodal Fusion (CMF) modules.

2) *In the second contribution:* we introduce an explainable architecture that fundamentally transforms how multimodal systems process medical data, combining concept activation vectors with hierarchical semantic alignment to produce reports clinicians can trust—unlike current "black box" systems. The framework's unique integration of visual-textual grounding with CAV-based explanations yields not just better performance, but more importantly, delivers the transparent diagnostic reasoning demanded in clinical practice.

3) *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 0.189, 0.221 in METEOR and 0.412 in ROUGE-L on the IU X-RAY dataset, demonstrating a notable improvement compared to the baseline and delivering superior results on this dataset, particularly in handling complex cases with multiple lesions.

The remainder of this paper is organized as follows. A review of related work in explainable AI and multimodal report generation is given in Section II. The architecture of our proposed Ex-ReGSA model, including its main modules, is described in detail in Section III. The experimental setup, including the dataset, evaluation metrics, and implementation specifics, is covered in Section IV. The findings of our comparative and ablation studies are shown in Section V. Section VI, which discusses the results with suggestions for further research, addresses the limitations and wider ramifications of our work.

II. 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) multi-modal fusion and (2) explainable artificial intelligence-based report generation.

A. Multimodal Fusion-Based Method for Report Generation

In recent years, several studies have explored the use of multimodal data fusion to improve diagnostic accuracy and report coherence. The method proposed by Tang et al. [7] solves medical report generation challenges by proposing the "locate then generate" CAT pattern. The framework uses a multi-modality encoder combined with a dual-stream decoder to dynamically incorporate retrieved terminologies with preceding sentences. Similarly, the approach proposed by Iqbal et al. [8] improves radiology reporting by utilizing an adaptive multi-modal approach that encodes clinical data elements. The system uses "wisdom learning" to extract medical insights from radiology reports through textual embeddings while applying cross-modal coherence that employs clinical embeddings to direct visual feature learning and enhance semantic alignment between images, disease labels, and generated reports. Li et al. [9] continued to explore diverse data integration. The authors developed a context-enhanced framework to address the shortcomings of image-only dependent systems. The report generation process is improved through their method, which combines multiple contextual elements such as clinical texts, structured medical knowledge, and previous diagnostic outcomes with primary medical images. Zeng et al. [10] ensures that generated reports maintain internal coherence while preserving their contextual integrity. The model utilizes a relational memory module that updates itself with previously generated words to enhance word correlations, as well as a double LSTM network with an interaction module to preserve contextual information during text generation. Tsaniya et al. [11] conduct research on architectural advancements along with

input data enhancements. The authors developed a medical report generator that operates on a transformer-based model. The system integrates text feature embedding through BERT with visual feature extraction from a pre-trained CheXNet model enhanced by multi-head attention to evaluate how contrast-based image enhancement positively influences report quality.

B. XAI-Based Report Generation

Several recent approaches have focused on embedding explainability directly into the architecture of report generation models. Zhang et al. [12] proposed the Attribute Prototype-guided Iterative Scene Graph framework, which utilizes an autoregressive approach for structural edge reasoning to handle common attribute and region feature limitations and improve interpretability. Similarly, Tanida et al. [13] developed an interactive region-guided framework that begins by detecting anatomical regions and then describes key regions to generate medical reports. Chen et al. [14] introduced AdaMatch as a model to establish detailed connections between CXR image segments and medical report terminology; then they used their AdaMatch-Cyclic model to utilize these connections for producing understandable CXR reports, which showed excellent results on public datasets. The development of specialized datasets and models focused on aligning with clinical reasoning is also crucial. One such study [15] introduced the FG-CXR dataset, which provides fine-grained pairings between radiologists' generated captions and corresponding gaze attention heatmaps for specific anatomies. The same study proposed a Gen-XAI network, which simulates diagnostic activities by matching its outputs with radiologist gaze patterns and transcript data. The research by Taleb et al. [16] addresses broader topics beyond report generation. The researchers introduced ContIG as a self-supervised learning method that enables the alignment of medical images and genetic data through contrastive loss and utilizes gradient-based XAI for interpreting cross-modal connections to demonstrate XAI effectiveness on diverse medical data sources. XAI techniques are currently used to improve transparency across multiple diagnostic tasks that involve different data types. For example, Sangnark et al. [17] created an explainable deep learning model to diagnose dysynergic defecation from abdominal X-rays and questionnaires utilizing cross-modal attention mechanisms with Grad-CAM and DeepSHAP for image and symptom data interpretation. Transparent COVID-19 interpretation research [18] implemented deep neural networks like EfficientNet and DenseNet together with XAI techniques including LIME and Grad-CAM as well as a new "Modified Grad-CAM++."

While existing methods have achieved remarkable performance on multimodal fusion and started to integrate explainability, from our literature review, we identify the following gaps in the existing methods: 1) Many methods use global feature alignment, which neglects the fine-grained, lesion-specific evidence that doctors often need to pay attention to for a detailed and accurate clinical description. 2)

Explainability in these methods is often restricted to post-hoc visualizations, such as heatmaps that indicate where the model is looking. While they reveal the localization of model attention, these heatmaps do not offer explanations that are interpretable in terms of well-defined and clinically meaningful concepts, resulting in a "black box" problem that can undermine clinical trust. To address these problems, we employ hierarchical feature extraction and semantic-guided alignment to attend to fine-grained features. Moreover, we introduce Concept Activation Vectors (CAVs) to our model to enable quantifiable and interpretable concept-based explanations of the model prediction.

III. PROPOSED METHOD

The overview of our proposed model for enhancing the generation of radiology reports Ex-ReGSA is illustrated in Fig. 1 and comprises four major modules: Explainable Semantic Feature Extractor (X-SFE), which includes Multi-Level Text Feature Extraction and Semantic Guided Knowledge extraction. The second module is the Semantic-Aligned Visual Extractor (X-SAVE), the third module is the Contrastive Multimodal Fusion module (CMF); and the end module is a Dynamic Decoder (DD).

First, the input medical report is processed by the Explainable Semantic Feature Extractor module to extract rich-grained features from the radiology reports through hierarchical attention networks. Subsequently, the Semantic Guided Knowledge Extraction Module is designed to extract relevant semantic knowledge from the fine-grained features obtained by Multi-Level Text Feature Extraction by categorizing similar information into groups that have the same characteristics. Moreover, the Semantic-Aligned Visual Extractor serves to extract visual features from radiology images, guided by semantic output obtained by X-SFE, to ensure that only the important text-based features are captured. The features extracted by X-SFE and X-SAVE then feed into the Contrastive Multimodal Fusion module (CMF), 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 Dynamic Decoder (DD) module integrates the aligned multi-modal features to generate, in a hierarchical process, a coherent and contextualized radiology report.

A. Explainable Semantic Feature Extraction (X-SFE)

1) *Multi-level text feature Extraction:* To extract multi-level features from medical report data, hierarchical attention mechanisms based on RadBERT[19], a Bidirectional Encoder Representations from Transformers (BERT) model specifically pre-trained on a large corpus of clinical radiology text, operate at three levels: report, sentence, and word. This approach captures both global diagnostic context and fine-grained clinical details while leveraging RadBERT's improved pretraining for better representation learning.

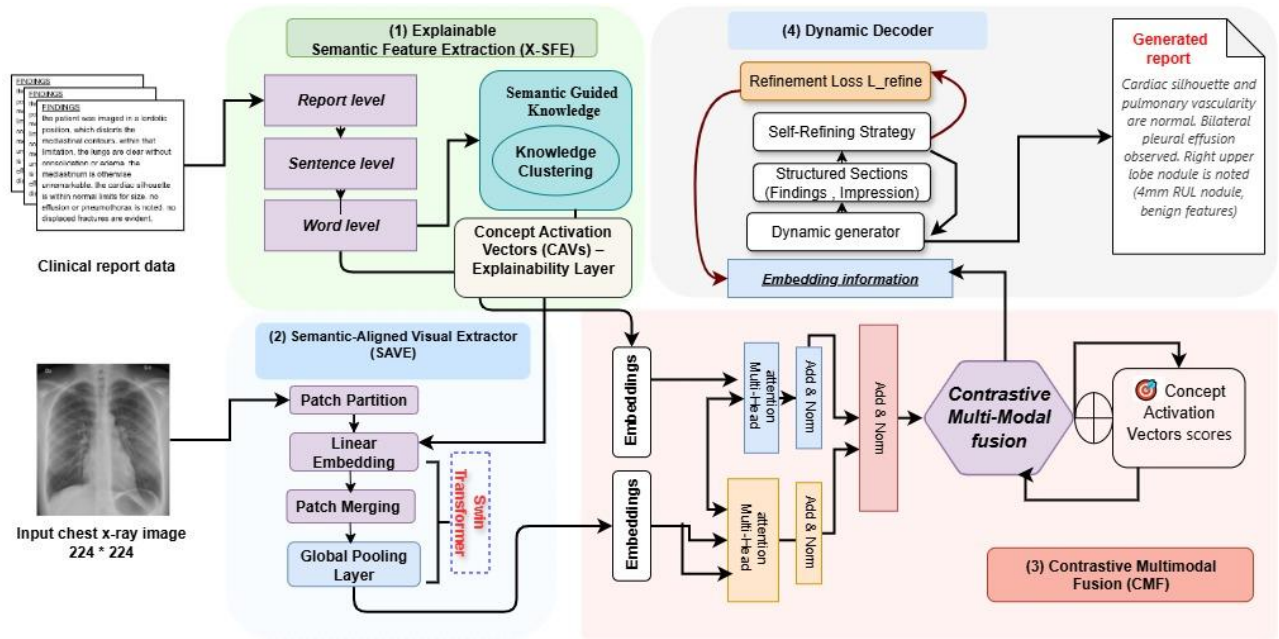


Fig. 1. Overview of the proposed Ex-ReGSA model.

Before robust hierarchical feature extraction can be applied to radiology reports data. They must first pass through an essential preprocessing stage. The process begins with extracting diagnostic sections such as "Findings" and "Impression," followed by text cleaning steps that include lowercasing text while removing unnecessary characters and artifacts. Crucially, the cleaned text, which we can denote as R_{clean_text} .

The prepared text undergoes processing by RadBERT, which conducts sub-word tokenization and creates foundational embeddings with deep context for all text tokens. The foundational token embeddings enable systematic creation of specialized representations covering all parts of the report down to the individual words and sentences. Converting the text into a sequence of M tokens, $T = \{t_1, t_2, \dots, t_M\}$. This token sequence T is then comprehensively processed by RadBERT to generate foundational contextual embeddings.

$$H = RadBERT(T) \quad (1)$$

where $H = \{h_1, h_2, \dots, h_M\}$ is the sequence of final hidden layer states from RadBERT. Each $h_j \in \mathbb{R}^{dh}$ represents a deeply contextualized embedding for token t_j , where dh is the dimensionality of RadBERT's hidden states. This output H serves as the rich, contextual basis from which all hierarchical features are subsequently derived. For Sentence-Level Feature Representation (Semb), Individual sentences within a clinical report often convey distinct findings or observations. To capture these, sentence-level embeddings, Semb, are derived for each of the N sentences, $S = \{S_1, S_2, \dots, S_N\}$, in the preprocessed report. For a given sentence S_i , its representation $Semb(S_i) \in \mathbb{R}^{dh}$ is constructed by aggregating the contextual token embeddings $\{h_j \in H\}$ that correspond to the Q_i tokens within that specific sentence (denoted $H_{S_i} = \{h_{S_i,1}, \dots, h_{S_i,Q_i}\}$). A common aggregation method is mean pooling:

$$Semb(S_i) = \frac{1}{Q_i} = \sum_{q=1}^{Q_i} h_{S_i,q} \quad (2)$$

Word embeddings for individual words make up the finest textual feature level Wemb. Word embeddings serve as essential tools for recognizing particular clinical entities as well as their complex meanings in medical reports. The word embedding $W_{emb}(W_k)$ for the k -th word W_k is derived from the RadBERT output H . Since RadBERT's tokenizer may split W_k into L_k sub-word tokens ($T(W_k) = \{t_{k,1}, \dots, t_{k,L_k}\}$), the corresponding contextual token embeddings ($h_{k,1}, \dots, h_{k,L_k}$ from H) are typically aggregated. The most common method is averaging:

$$W_{emb}(W_k) = \frac{1}{L_k} = \sum_{l=1}^{L_k} h_{k,l} \quad (3)$$

2) *Semantic guided knowledge extraction:* The Semantic Knowledge Module is designed to extract relevant knowledge from chest X-ray reports by categorizing similar information into groups that have the same characteristics. To achieve this, we employ a clustering algorithm on report-embedded representation vectors to group similar information together into several clusters. To quantify the presence and strength of these representations within each report and to guide the clustering process towards forming semantically interpretable groups, we leverage Concept Activation Vectors (CAV) [21]. In particular, given a set of embedding vectors $V = \{v_1, v_2, \dots, v_n\}$, the K-Means algorithm [20] is used to group these vectors into groups (clusters), denoted T_k , based on their similarities. This process is defined by minimizing the Euclidean distance between each report's embedding V_i and the centroid C_j of its corresponding cluster. Initialize K centroids randomly, where each centroid m_j (for $j=1, 2, \dots, K$) represents the initial mean of a cluster. This can be expressed as Eq. (4).

$$T_k = \arg \min_j \sum_{d=1}^D (v_{i,d} - m_{j,d})^2 + \alpha \cdot CAVsim(v_i, c_j) \quad (4)$$

Where T_k is the index of the cluster assigned to report v_i and m_j is the centroid, D is the dimensionality of each embedding vector y_i and $y_{i,d}$ and $m_{j,d}$ represent the d -th component of vectors v_i and m_j , respectively. After this step, the centroids continue to get refined and updated as the meaning of the embeddings associated to each cluster. This iterative process executes until labeled reports do not change anymore, resulting in k groups of reports, where each cluster encapsulates specific features, like similar writing style or diagnostic terminology. And the function $CAVsim(v_i, c_j)$ measuring the semantic or conceptual similarity between report v_i and cluster centroid m_j , based on their alignment with predefined concepts quantified by CAVs. While the α is a weighting parameter that balances the influence of the Euclidean distance against the CAV-guided conceptual similarity term.

B. Explainable Semantic-Aligned Visual Extractor (X-SAVE)

To reduce the problem of data bias and extract precise spatial and semantic features of disease lesions, we employ a multi-level feature extraction technique that relies heavily on prior knowledge obtained from semantic clusters found within clinical text reports. This guidance aims to align visual regions with relevant textual features and to direct the model's attention towards focal regions based on their semantic relevance, determined through cluster-specific importance signals. The X-SAVE module receives a collection of chest X-ray (CXR) images $X = \{x_1, x_2, \dots, x_n\}$, where each image x_i is a tensor in $RC \times H \times W$, with C representing the number of channels and H and W representing spatial height and width, respectively. While $H = W = 224$ for CXR images. The feature extraction process starts by passing the input images through a SwinTransformer [22], which serves as the visual backbone. The Swin Transformer is selected for its effectiveness in computer vision tasks and its inherent hierarchical architecture, which captures features at multiple scales by progressively downsampling the input image. Enabling both localized fine-grained features and global contextual information. The process is illustrated in Eq. (5) to Eq. (7).

1) *Macro-level features (F1)*: Capturing coarse, global anatomical structures, the image is downsampled by a factor of 4:

$$F1 = \text{SwinTransformer}(X) \quad (5)$$

where, $F1 \in \mathbf{R}^{(C \times H/4 \times W/4)} = \mathbf{R}^{(C_1 \times 56 \times 56)}$, the resolution becomes 56×56 with C_1 channels.

2) *Intermediate level (F2)*: At the second level, highlighting regional characteristics and intermediate structures, the CXR-image is downsampled by another factor of 2:

$$F2 = \text{SwinTransformer}(F1) \quad (6)$$

Where $F2 \in \mathbf{R}^{(C \times H/8 \times W/8)} = \mathbf{R}^{(C_2 \times 28 \times 28)}$ and the resolution is reduced to 28×28 , with C_2 channels.

3) *Micro-level features (F3)*: Detailing fine-grained visual patterns crucial for localizing subtle lesions., The image is downsampled by a factor of 2:

$$F3 = \text{SwinTransformer}(F2) \quad (7)$$

Here $F3 \in \mathbf{R}^{(C \times H/16 \times W/16)} = \mathbf{R}^{(C_3 \times 14 \times 14)}$ and the resolution becomes 14×14 , and C_3 channels are used.

The key innovation lies in our semantic alignment integrated with the Concept Activation Vectors (CAVs) mechanism, where text-derived clinical concepts from radiology reports are clustered (T_1, \dots, T_k) using RadBERT embeddings, and each cluster's importance weight w_k is computed based on feature dispersion from its centroid m_k to reflect semantic coherence, computed as Eq. (8):

$$w_k = \frac{1}{|T_k|} \sum_{i \in T_k} ||v_i - m_k||_2 \quad (8)$$

Where $|T_k|$ is the number of reports in cluster T_k .

The visual features extracted for each input image x are then aligned with the clusters as follows in Eq. (9):

$$F_{\text{Aligned}} = \sum_{k=1}^K (w_k \cdot \text{TCAV}_k) \cdot Fk \quad (9)$$

where $Fk = Wk \cdot [F2 \oplus F3]$ represents the visual features associated with cluster T_k , and Wk is a learnable projection matrix.

C. Contrastive Multimodal Fusion (CMF)

To apply a hierarchical alignment and fusion, we use a transformer-based model. Current multi-modal fusion approaches rely on global feature fusion. We design multimodal fusion based on self-supervised attention, which comprises two key stages: cross-modal attention and self-supervised contrastive learning. The CMF module learns aligned representations by combining detailed feature interactions with a comprehensive semantic understanding from visual and textual data sources. The system accomplishes this through a unified process that uses hierarchical cross-modal attention mechanisms to map visual and textual features at various scales into detailed correspondences. The features are attentively fused and aligned and then projected into a shared embedding space, which undergoes refinement with a self-supervised contrastive learning objective to ensure robust end-to-end semantic alignment. The system achieves explainability by leveraging Concept Activation Vectors (CAVs) for understanding the semantic characteristics of the shared embedding space.

This initial step within CMF focuses on establishing detailed correspondence between the hierarchically extracted visual features and report features. As previously detailed, hierarchical visual features ($F1$ for global, $F2$ for regional, and $F3$ for fine-grained details) are obtained from the input CXR image. Similarly, hierarchical textual features (Remb for report-level, Semb for sentence-level, and $\text{wemb}(j)$ for word-level) are extracted from the medical report.

To bridge representational differences, both feature sets undergo modality-specific normalization and are augmented with positional embeddings. The cross-modal attention then aligns these features.

The normalized and enriched features are then used to compute the queries (Q_s), keys (K_s), and values (V_s) directly within the attention mechanism:

$$A_s = \text{Softmax} \left(\frac{Q_s K_s^T}{\sqrt{d}} + \log(\text{TCAV} + 1) \right) \quad (10)$$

Where $Q_s = F_{text}^{pos}$: The query matrix is derived from position-enriched textual features. And $K_s = F_{visual}^{pos}$ Is the key matrix derived from the position-enriched visual features. The d is the dimensionality of the key vectors, and serves as a scaling factor to ensure gradient stability in training. By using the attention scores A_s , the aligned features F_s aligned are computed by applying A_s to the value matrix (V_s), which is also derived from the position-enriched visual features. $F_{aligned}^s = A_s V_s$ and $V_s = F_{visual}^{pos}$

The aligned features from all scales are concatenated into a unified representation:

$$F_{fusion} = \text{Concat}(F_{1aligned}, F_{2aligned}, F_{3aligned}) \quad (11)$$

After generating a unified representation (F_{fusion}), visual features will be projected into a shared latent space through a learnable matrix (W_v): $F^{visual} = W_v F_{fusion}$ and textual features with another one (W_t): $F^{text} = W_t F_{fusion}$. These embeddings are refined using multimodal contrastive learning, which optimizes a contrastive loss ($L_{contrastive}$) to align positive pairs, such as a CXR image and its corresponding report, while distancing unrelated pairs. The loss is calculated in Eq. (12) as:

$$L_{contrastive} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(F^{visual}_i, F^{text}_i) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(F^{visual}_i, F^{text}_j) / \tau)} \quad (12)$$

Where $\text{sim}(x, y)$ is the cosine similarity, and τ is a temperature parameter controlling the sharpness of the similarity distribution. This end-to-end, self-supervised training paradigm leverages inherent data pairings without requiring manual annotations, ensuring scalability and robust semantic alignment for downstream tasks.

D. Dynamic Decoder (DD)

To overcome critical limitations in existing report generation decoders—such as static templates, neglect of hierarchical features, and lack of dynamic refinement. We propose a novel Dynamic Decoder (DD) that processes multi-level fused visual features (F_{fusion}) in a hierarchical structure. The decoder process uses corresponding report, sentence, and word-level textual embeddings together with fused visual features to construct final reports. A key innovation is its Dynamic Sentence-Type Adaptation mechanism, which uses learned, Concept Activation Vector (CAV)-influenced attention weights (α_T, α_A) where, for instance, α_T is determined by Eq. (13):

$$\alpha_T = \sigma(W_T F_{fusion} + \beta \cdot \text{CAVsim}(F_{fusion}, c_k)), \quad \alpha_A = 1 - \alpha_T \quad (13)$$

Where CAVsim : Measures alignment between fused features and concept c_k . These weights ensure critical abnormalities are prioritized in complex cases, while template sentences dominate in normal cases. The final representation is presented in Eq. (14):

$$R_t = \alpha_T T + \alpha_A A \quad (14)$$

where T represents normal features and A represents abnormal features.

Furthermore, a Feedback-Driven Refinement loop iteratively improves semantic consistency by re-encoding the generated report Y' and minimizing a refinement loss (L_{refine}) against the input features F_{fusion} , defined as $L_{refine} = \|F_{fusion} - \text{Encoder}(Y')\|_2^2$. The DD is trained end-to-end with a hybrid loss function, combining objectives for token accuracy (L_{acc}), feature alignment (L_{refine}), and overall semantic coherence (via $L_{contrastive}$), enabling it to generate clinically accurate, contextually relevant, and professionally structured medical reports:

$$L_{refine} : L_{total} = \lambda_1 L_{acc} + \lambda_2 L_{contrastive} + \lambda_3 L_{refine} \quad (15)$$

IV. EXPERIMENTAL SETTINGS

A. Dataset

IU X-ray [23]: is a publicly accessible repository of chest X-rays, featuring 7,470 images and their 3,955 associated radiological reports. The nature of this data makes it a valuable asset for developing and training models in tasks such as automated radiology report generation and computer-aided diagnostic systems. To maintain consistency and allow for comparison with prior research, this collection is conventionally divided into standard proportions: 70% of the data for model training, 10% for validation during development, and the remaining 20% for final model testing.

B. Evaluation Metrics

The quality of our model was evaluated using four NLG Natural Language Generation metrics. To measure precision in the generated report compared to ground truth clinical report, we used BLEU [24], which scores n-gram overlap. For an evaluation more aligned with human intuition, we used METEOR [25], as it considers variations like word stems and synonyms from WordNet. To specifically assess the fluency and coherence of the generated content, we employed ROUGE-L [26], which measures similarity based on the longest common subsequence of words found in both the generated and reference texts. Additionally, CIDEr [27] was utilized to assess how well the generated text captures salient concepts present across a set of reference texts, by weighing n-grams based on their Term Frequency-Inverse Document Frequency (TF-IDF) scores, thereby emphasizing consensus and information richness.

C. Implementation Details

Our medical report generation system is built on a hierarchical, multimodal architecture. For visual input, we process 224x224 resolution chest X-rays using a Swin Transformer backbone. This visual encoder uses a patch size of 4, an embedding dimension of 96, and four feature extraction stages with 3, 6, 12, and 24 attention heads, respectively. The text processing component utilizes RadBERT for hierarchical feature extraction at the word, sentence, and document levels, each with a 768-dimensional hidden state for attention. To integrate these, a multimodal fusion module projects both visual and textual features into a shared 128-dimensional space via linear transformations, followed by cross-modal attention with 4 parallel heads. Report generation is handled by a 6-layer transformer decoder (512 hidden units, 8 attention heads) that dynamically regularizes sentence types between templates and abnormal findings. The model was trained for 40 epochs using

a composite loss function—combining cross-entropy (weight 1.0), contrastive loss (weight 0.5, temperature 0.07), and a learned features refinement loss (weight 0.5)—optimized with Adam at a learning rate of 5×10^{-4} . For generating reports (inference), we employ a beam search with a width of 3, using hyperparameters (contrastive temperature τ between 0.05-0.2, loss balancing λ between (0.1-0.9) selected based on ROUGE scores on a validation set.

V. EXPERIMENTS RESULTS

A. Comparison Experiment

We evaluated our model using standard metrics: we employed BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr metrics to measure both the quality and relevance of the reports. As shown in Table I and Fig. 2, comparative analysis is performed against state-of-the-art models: The evaluation includes comparisons with advanced models HRGR-Agent[28], R2GEN[29], PPKED [30], AERMNet[10], C-Enhanced-F [9], and work by SAEED et al. [8]. These benchmarks provide thorough assessments of fluency quality along with informative content and clinical consistency.

Analyzing the BLEU scores, which measure n-gram precision against reference texts, our model demonstrates strong performance, particularly for lower-order n-grams. It achieves the highest BLEU-1 score of 0.517, outperforming all listed competitors, including notable models like SAEED et al. (0.499) and C-Enhanced-F (0.491). For higher-order n-grams, our model remains highly competitive: its BLEU-2 score is 0.351, slightly behind C-Enhanced-F, which scored 0.359. Similarly, its BLEU-3 score is 0.251, with C-Enhanced-F at 0.263, and its BLEU-4 score is 0.189, compared to C-Enhanced-F's 0.209. While C-Enhanced-F exhibits an edge in these higher-order BLEU metrics, indicating a closer match in longer contiguous word sequences, our model consistently surpasses other methods like R2GEN and PPKED across all BLEU categories. This suggests a strong grasp of lexical choice and phraseology.

In terms of the METEOR metric, which evaluates generated text quality by considering stemming, synonyms, and word order to better align with human judgment, our model scores 0.221. This is the highest METEOR score among all models for which this metric is reported in the table, surpassing AERMNet (0.219), C-Enhanced-F (0.212), and R2GEN (0.185). This leading performance in METEOR suggests that our model generates reports that are not only accurate in terms of word choice but also capture semantic similarity effectively, aligning well with human assessment criteria.

TABLE I. PERFORMANCE COMPARISON

Model	B-1	B-2	B-3	B-4	M.	R-L	CIDEr
[28]	0.438	0.298	0.208	0.151	--	0.322	0.343
[29]	0.451	0.293	0.209	0.159	0.185	0.381	0.406
[30]	0.483	0.315	0.224	0.168	--	0.376	0.365
[10]	0.486	0.321	0.236	0.183	0.219	0.398	0.560
[9]	0.491	0.359	0.263	0.209	0.212	0.408	0.396
[8]	0.499	0.349	0.229	0.170	--	0.401	0.411
(Ours)	0.517	0.351	0.251	0.189	0.221	0.412	0.463

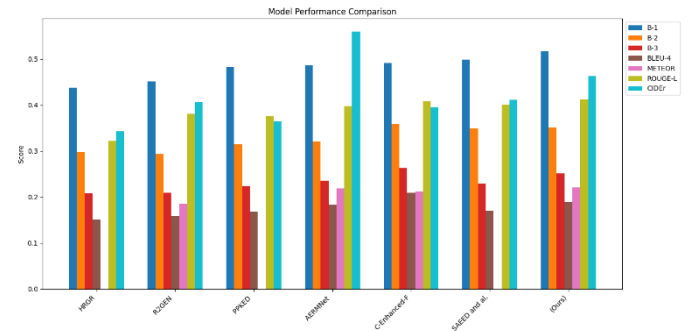


Fig. 2. Performance comparison of models on IU-CXR.

TABLE II. ABLATION EXPERIMENT

Model	B-1	B-2	B-3	B-4	M.	R-L	CIDEr
Baseline Model	0.405	0.301	0.219	0.168	0.181	0.350	0.392
+ X-SFE	0.413	0.314	0.222	0.172	0.188	0.358	0.409
+ X-SAVE	0.422	0.332	0.239	0.181	0.192	0.373	0.441
+CMF	0.451	0.339	0.243	0.182	0.202	0.398	0.453
+Self-Refining mechanism	0.470	0.349	0.249	0.189	0.214	0.408	0.461
Full Model	0.517	0.351	0.251	0.189	0.221	0.412	0.463

Our model particularly excels in the ROUGE-L score, achieving 0.412. This metric, which measures the longest common subsequence between generated and reference texts, reflects fluency and the recall of important content. Attaining the top score in ROUGE-L indicates our model's superior ability to capture the structural similarity and main ideas present in the ground truth reports when compared to all other listed models, including strong performers like C-Enhanced-F (0.408), SAEED et al. (0.401), and AERMNet (0.398). This suggests that the reports generated by our model are coherent and effectively convey the necessary information.

The CIDEr metric, designed to measure consensus and human-like quality in image descriptions by weighting n-grams using Term Frequency-Inverse Document Frequency (TF-IDF), offers further insights into the model's performance. Our model achieves a CIDEr score of 0.413. This score is competitive and robust, surpassing or closely aligning with several models such as SAEED et al. (0.411), R2GEN (0.406), and C-Enhanced-F (0.396). However, it is noteworthy that AERMNet demonstrates a significantly higher CIDEr score of 0.560. This indicates AERMNet's strong capability in generating text that effectively captures n-grams frequently found across multiple reference

reports, a key aspect of consensus. While our model performs well against many peers in CIDEr, AERMNet sets a high benchmark for this specific measure of quality.

In summary, this comparative analysis reveals that our proposed model exhibits a strong and highly competitive performance profile. It achieves leading results in BLEU-1, METEOR, and ROUGE-L, signifying its strengths in generating lexically precise, semantically rich, and structurally coherent radiology reports. While C-Enhanced-F shows an advantage in higher-order BLEU scores, and AERMNet leads substantially in CIDEr, our model consistently outperforms many established baselines across multiple critical metrics. The particularly strong ROUGE-L and METEOR scores suggest that the generated reports are fluent and capture clinical information effectively, which is paramount for clinical utility. The CIDEr score of 0.413, while not the highest, is solid and points towards good consensus capture. These results collectively affirm the Efficacy of our model as a significant contribution to the field of automated radiology report generation, though future work could explore strategies to further enhance performance on metrics like CIDEr to match the top specialist models.

B. Ablation Experiment

In this section, we will conduct an ablation experiment to indicate the influence of each module on our model's 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 (X-SFE, X-SAVE, CMF, and DD) and the combination on final accuracy. Table II provides a detailed comparison over multiple evaluation metrics, including B-1, B-2, B-3, B-4, METEOR, and ROUGE-L, for the Indiana University dataset.

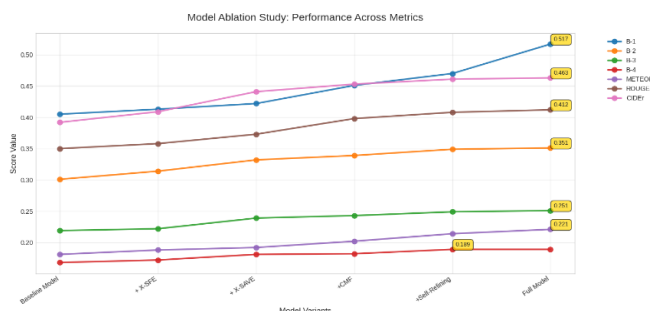


Fig. 3. Performance metrics of ablation study.

Table II and Fig. 3 illustrate that the full model yields substantial performance gains when compared to the baseline model, thereby underscoring the crucial role of every proposed module. Quantitatively, the Full Model surpassed the Baseline by 11.2% on BLEU-1, 5.0% on BLEU-2, 3.2% on BLEU-3, 2.1% on BLEU-4, 4.0% on METEOR, 6.2% on ROUGE-L, and 7.1% on CIDEr. These results indicate a marked improvement in report generation quality due to the additional components. A systematic evaluation of each component's individual effect was subsequently undertaken by comparing the various 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.

+X-SFE (Explainable Semantic Feature Extraction): This configuration adds the explainable text feature extraction module to the baseline model. This component leads to discernible gains across all metrics. For instance, BLEU-1 rises to 0.413, ROUGE-L to 0.358, and notably, CIDEr improves by 1.7%. This suggests that enhancing the model's ability to capture hierarchical relations in the text provides better context and structure for report generation.

+X-SAVE (Explainable Semantic-Aligned Visual Extractor): This module focuses on fine-grained lesion detection. The X-SAVE module guides visual feature extraction 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. This module yields more significant boosts, with BLEU-1 reaching 0.422, BLEU-2 improving from 0.314 to 0.332, ROUGE-L increasing to 0.373 (+0.015), and CIDEr showing a strong jump to 0.441 (+0.032). These results underscore the importance of semantic-aligned visual extractors in focusing the model on diagnostically relevant image regions, thereby enhancing the quality and relevance of the generated text, particularly reflected in the CIDEr score.

+CMF Multimodal Fusion module: In this setup, the CMF module is integrated into the model. As shown in Table II, 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. This addition results in BLEU-1 increasing to 0.451, METEOR to 0.202 (+0.010), and ROUGE-L to 0.398 (+0.025). The CIDEr score also sees a healthy increase to 0.453. These results highlight the importance of multi-level cross-modal alignment in improving the integration of visual and textual data, leading to more accurate and coherent medical report generation.

+Self-refining mechanism: We add a self-refining mechanism strategy in our decoder. As shown in Table II, this configuration substantially outperforms BASE; BLEU-1 reaches 0.470, METEOR improves to 0.214, ROUGE-L to 0.408, and CIDEr to 0.461. This indicates the importance of this configuration in improving the semantic quality of reports that are generated by maintaining consistent and context-dependent terminology.

The full model, which merges all modules into an optimized layout, demonstrates superior performance. The BLEU-1 score reached 0.517, which represents a considerable increase from the former performance at 0.470. The METEOR score achieves 0.221, while ROUGE-L reaches a level of 0.412 and CIDEr advances to 0.463. Transitioning from the "+Self-Refining mechanism" configuration to the "Full Model" generates substantial BLEU-1 improvements and consistent metric gains, which demonstrate that combined optimization of all modules creates a powerful synergy or enables the self-refining process to achieve its maximum potential in this final configuration. The BLEU-4 score maintains a value of 0.189 across the final two

stages because advancements have been concentrated on the text structure and semantic meaning along with lower-order n-gram precision.




C. Visual Analysis

We demonstrated the clinical insight and effectiveness of our Ex-ReGSA model by choosing illustrative cases for qualitative analysis, which appear in Table III. The table presents a side-by-side analysis of the original radiological report against Ex-ReGSA-produced output and the model-derived explanations for its results, which appear within the context of the input radiological image. This visual experiment surpasses quantitative metrics by providing richer insight into Ex-ReGSA's ability to interpret complex clinical narratives and demonstrate its decision-making transparency.

As shown in Table III, 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 images 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 X-SFE and X-SAVE during hierarchical feature extraction. This guidance enables the model to offer importance on critical areas, improving the clinical relevance of the generated reports. Furthermore, the "Interpretable Explanations" column provides important understanding of how the model arrives at its decisions and processes information internally. These examples illustrate how these explanations combine model-assigned confidence scores of specific findings (e.g., "cardiomegaly score 0.92," "pleural effusion score 0.93") with mentions of key visual features detected by internal components alongside potential semantic knowledge or quantifiers used by the model. Through detailed case-by-case explanations of its operational logic, Ex-ReGSA proves transparent because it shows both its pathologies and normality reports and explains the reasoning behind those conclusions, which builds confidence in its clinical usefulness and reliability.

TABLE III. THE VISUALIZATION OF OUR EX-RGSA MODEL

image	Ground truth	Generated report	Interpretable Explanations
	The heart is normal in size. The lungs are clear without evidence of pleural effusion, pneumothorax, or focal lesions. Bony structures are unremarkable.	Heart size is within normal limits. No pulmonary abnormalities detected. Lungs are clear bilaterally, with no signs of effusion, pneumothorax, or consolidation.	<ul style="list-style-type: none">- The diagnosis of "clear lung fields" achieved a high confidence level of 0.97 because GGO and consolidation features were not present.- The "Normal cardiac size" score of 0.95 reflects that the cardiothoracic ratio falls within the established normal range.- Diffuse attention across both lung parenchyma (avg. intensity :0.2 per zone), no focal hotspots.- The sharp and clear appearance of the costophrenic angles in visual features confirms the absence of pleural effusion with a high confidence level of 0.98, showing a contrast to typical fluid accumulation characteristics.
	Enlarged cardiac silhouette suggestive of cardiomegaly. Lungs are clear of consolidation, but mild pulmonary vascular congestion is noted, likely due to early congestive changes."	The cardiac silhouette is enlarged. Lungs appear clear. No acute infiltrates observed. Mild vascular congestion inferred from hilar vessel prominence."	<ul style="list-style-type: none">- High score for "cardiomegaly" (0.92) as calculated cardiothoracic ratio > 0.55.- "Vascular congestion" score (0.75) from subtle hilar vessel prominence.- Strong focus on lateral cardiac borders (left avg. attention=0.88, right avg. attention=0.85).- Term "mild" for "vascular congestion" selected from "severity quantifiers" cluster- "No acute infiltrates" (confidence 0.90) as lung fields lack dense opacification features, distinguishing from decompensated heart failure.
	Moderate right pleural effusion with associated compressive atelectasis. Left lung parenchyma is normal. No cardiomegaly.	Right-sided pleural effusion is noted (moderate volume) causing obscuration of the costophrenic angle. Left lung remains clear. Cardiac size unremarkable.	<ul style="list-style-type: none">- High score for "pleural effusion" (0.93) due to detection of costophrenic angle blunting & meniscus sign on the right.- Focus on right costophrenic angle (attention=0.94) and lateral right pleural lining.- Term "moderate" pulled from cluster "effusion quantifiers", correlated with estimated fluid volume- "Left lung clear" (confidence 0.96) confirmed by sharp left CPA and lack of opacity, ruling out bilateral effusion.

Collectively, Ex-ReGSA demonstrated through representative cases Table III its ability to produce accurate radiology reports that thoroughly include necessary clinical terms and findings from ground truth while maintaining high coherence. The "interpretable and explainable" evidence links generated statements with visual features and advanced analytical components to boost our model's clinical trustworthiness and usefulness.

The necessary degree of transparency supports radiologists by potentially decreasing their reporting workload and enhancing diagnostic speed while functioning as a dependable clinical tool. Ex-ReGSA's architecture demonstrates strong performance across various cases, which highlights its ability to manage different radiological presentations effectively.

VI. DISCUSSION

Our research shows that the Ex-ReGSA model demonstrates robust performance while achieving competitive results in automated radiology report generation, as confirmed by quantitative data and qualitative evaluations along with an insightful ablation study. Ex-ReGSA achieves state-of-the-art or similar scores when compared to existing methods across essential metrics with exceptional performance in all metrics, which indicates high lexical accuracy and semantic depth along with strong structural coherence. Furthermore, the ablation study systematically confirmed the individual and collective contributions of its core architectural modules, namely the X-Semantic Feature Extractor (X-SFE), X-Semantic-Aligned Visual Extractor (X-SAVE), Cross-Modal Fusion (CMF), and the Self-Refining mechanism—each incrementally enhancing report quality towards the Full Model's performance. Complementing these objective scores, our qualitative analysis, showcased in Table III, illustrates Ex-ReGSA's ability to generate clinically relevant and accurate narratives for diverse cases, from identifying chronic conditions to recognizing acute findings such as pleural effusions and even subtle incidentalomas like calcified granulomas.

Ex-ReGSA achieves its strength through a well-designed architecture specifically created to tackle multimodal medical report generation challenges. The collaboration between advanced visual feature extraction methods (X-SFE and X-SAVE) delivers precise identification and interpretation of clinically important image regions. The Cross-Modal Fusion (CMF) module demonstrates essential functionality by combining visual cues with available textual information to create a comprehensive understanding. The output labeled "Interpretable and Explainable" stands out as a major advancement for achieving transparent AI within radiology. The model's decision-making process becomes clear through an analysis of system-detected features along with attention focal points and confidence levels. Ex-ReGSA produces reports while delivering a foundation for interpreting its conclusions. The model's inherent interpretability remains essential for developing clinical confidence and assisting with debugging and refining the model.

When compared to the broader landscape of radiology report generation research, Ex-ReGSA positions itself favorably. Its leading ROUGE-L score, for instance, suggests a superior capability in capturing the essential content and narrative flow

of reference reports compared to many existing models. The METEOR score further supports its ability to generate semantically coherent and fluent text. While Ex-ReGSA achieves a solid CIDEr score (0.463 as per the ablation study's Full Model), some specialized models, such as AERMNet (which reported a CIDEr of 0.560 in comparative settings), demonstrate higher performance on this specific consensus-based metric. This suggests that while Ex-ReGSA demonstrates superior performance in report structure and semantic accuracy, it must improve n-gram overlap to maximize CIDEr rewards. While other models focus on single metrics and lack transparency, Ex-ReGSA demonstrates a distinct advantage through its comprehensive capabilities and built-in interpretability features. Therefore, our model, Ex-ReGSA, offers three key advantages over current methods. First, it achieves superior fine-grained accuracy by using a semantic-guided visual extractor (X-SAVE) to focus on specific, subtle lesions that other models miss. Second, it provides truly explainable and trustworthy AI by using Concept Activation Vectors (CAVs) to explain why it makes a diagnosis in quantifiable terms, rather than just showing a heatmap of where it looks. Finally, it ensures enhanced coherence and reliability through a dynamic, self-refining decoder that corrects its output, leading to more logical and professionally structured reports than static generators.

The study achieved promising results but remains limited by certain constraints. Although the IU X-ray dataset serves as a standard benchmark, its limited size and variety fail to capture the full spectrum of clinical scenarios found in worldwide medical practice. Further research is required to understand how the model handles fewer common conditions and complex cases that involve multiple subtle interacting findings. Our CIDEr score competes well but remains behind the field leaders, which demonstrates room for growth in producing reports that match multiple radiologists' preferred terminology. The existing interpretability methods provide valuable insights but represent only one phase in the continuous effort to achieve completely transparent and interactive artificial intelligence.

Future work should focus on several key directions. Addressing the CIDEr gap could involve exploring novel decoding strategies or incorporating more fine-grained common-sense and clinical knowledge into the generation process. Training and validating Ex-ReGSA on larger, more diverse, and multi-institutional datasets would be crucial for assessing its generalizability and robustness. Prospective clinical studies are also essential to evaluate Ex-ReGSA's real-world impact on radiologists' workflow, reporting efficiency, and diagnostic accuracy. Further development of the interpretability framework could involve more interactive explanation modalities, allowing clinicians to query the model's reasoning.

VII. CONCLUSION

Our research presents Ex-ReGSA as a novel method for automated radiology report generation which combines semantic-guided visual alignment, concept-based explainability via CAVs, and a dynamic self-refining decoder. Our model outperforms the state-of-the-art systems on major evaluation metrics such as ROUGE-L and METEOR, generating medically

accurate and fluent reports that are also interpretable. By explaining its diagnostic decisions in terms that can be quantified, Ex-ReGSA aims to promote transparency and trust in clinical AI, providing a powerful approach to improving diagnostic workflow, reducing radiologist workload, and helping interpretability in clinical decision-making. Although validation is successful, specific limitations are its testing on a single benchmark dataset and the potential to improve consensus-based metrics. Future work will focus on training with larger, multi-institutional datasets and conducting prospective clinical studies to affirm its real-world utility.

REFERENCES

- [1] James H. Thrall, Xiang Li, Quanzheng Li, Cinthia Cruz, Synho Do, Keith Dreyer, James Brink, Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success, Journal of the American College of Radiology, Volume 15, Issue 3, Part B, 2018, Pages 504-508.
- [2] Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 44–56 (2019).
- [3] Wang, Xinyi, Graziela Figueredo, Ruizhe Li, Wei E. Zhang, Weitong Chen, and Xin Chen. "A Survey of Deep Learning-based Radiology Report Generation Using Multimodal Data."
- [4] Tang, Yuhao & Wang, Dacheng & Zhang, Liyan & Yuan, Ye. (2024). An efficient but effective writer: Diffusion-based semi-autoregressive transformer for automated radiology report.
- [5] Tjoa, Erico & Guan, Cuntai. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems. PP. 10.1109/TNNLS.2020.3027314.
- [6] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q consortium (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making, 20(1), 310.
- [7] Y. Tang, Y. Yuan, F. Tao and M. Tang, "Cross-Modal Augmented Transformer for Automated Medical Report Generation," in *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 13, pp. 33-48, 2025.
- [8] 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," in *IEEE Access*, vol. 12, pp. 127341-127356, 2024.
- [9] Hongzhao Li, Hongyu Wang, Xia Sun, Hua He, Jun Feng, Context-enhanced framework for medical image report generation using multimodal contexts, Knowledge-Based Systems, Volume 310, 2025,112913.
- [10] ianhua Zeng, Tianxing Liao, Liming Xu, Zhiqiang Wang, AERMNet: Attention-enhanced relational memory network for medical image report generation, Computer Methods and Programs in Biomedicine, Volume 244, 2024, 107979
- [11] H. Tsaniya, C. Fatichah and N. Suciati, "Automatic Radiology Report Generator Using Transformer With Contrast-Based Image Enhancement," in *IEEE Access*, vol. 12, pp. 25429-25442, 2024.
- [12] K. Zhang *et al.*, "Attribute Prototype-Guided Iterative Scene Graph for Explainable Radiology Report Generation," in *IEEE Transactions on Medical Imaging*, vol. 43, no. 12, pp. 4470-4482, Dec. 2024.
- [13] T. Tanida, P. Müller, G. Kaissis and D. Rueckert, "Interactive and Explainable Region-guided Radiology Report Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 7433-7442.
- [14] Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9494–9509, Bangkok, Thailand. Association for Computational Linguistics.
- [15] Pham, T.T. et al. (2025). FG-CXR: A Radiologist-Aligned Gaze Dataset for Enhancing Interpretability in Chest X-Ray Report Generation. In: Cho, M., Laptev, I., Tran, D., Yao, A., Zha, H. (eds) Computer Vision – ACCV 2024. ACCV 2024. Lecture Notes in Computer Science, vol 15477. Springer, Singapore.
- [16] A. Taleb, M. Kirchler, R. Monti and C. Lippert, "ContIG: Self-supervised Multimodal Contrastive Learning for Medical Imaging with Genetics," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 20876-20889.
- [17] S. Sangnark, P. Rattanachaisit, T. Patcharatrakul and P. Vatekul, "Explainable Multi-Modal Deep Learning With Cross-Modal Attention for Diagnosis of Dyssynergic Defecation Using Abdominal X-Ray Images and Symptom Questionnaire," in *IEEE Access*, vol. 12, pp. 78132-78147, 2024.
- [18] Kinger, S., Kulkarni, V. Transparent and trustworthy interpretation of COVID-19 features in chest X-rays using explainable AI. Multimed Tools Appl 84, 19853–19881 (2025).
- [19] Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A. and Hsu, C.N., 2022. RadBERT: adapting transformer-based language models to radiology. Radiology: Artificial Intelligence, 4(4), p.e210258.
- [20] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020.
- [21] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J. and Viegas, F., 2018, July. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In International conference on machine learning (pp. 2668-2677). PMLR.
- [22] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
- [23] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).
- [24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- [25] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.
- [26] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- [27] Vedantam, R., Lawrence Zitnick, C. and Parikh, D., 2015. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).
- [28] Li, Y., Liang, X., Hu, Z. and Xing, E.P., 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems, 31.
- [29] Chen, Z., Song, Y., Chang, T.H. and Wan, X., 2020. Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056.
- [30] F. Liu, X. Wu, S. Ge, W. Fan and Y. Zou, "Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 13748-13757.