

# Medical Report Generation via Multimodal Spatio-Temporal Fusion

## ABSTRACT

Medical report generation aims at automating the synthesis of accurate and comprehensive diagnostic reports from radiological images. The task can significantly enhance clinical decision-making and alleviate the workload on radiologists. Existing works normally generate reports from single chest radiographs, although historical examination data also serve as crucial references for radiologists in real-world clinical settings. To address this constraint, we introduce a novel framework that mimics the workflow of radiologists. This framework compares past and present patient images to monitor disease progression and incorporates prior diagnostic reports as references for generating current personalized reports. We tackle the textual diversity challenge in cross-modal tasks by promoting style-agnostic discrete report representation learning and token generation. Furthermore, we propose a novel spatio-temporal fusion method with multi-granularities to fuse textual and visual features by disentangling the differences between current and historical data. We also tackle token generation biases, which arise from long-tail frequency distributions, proposing a novel feature normalization technique. This technique ensures unbiased generation for tokens, whether they are frequent or infrequent, enabling the robustness of report generation for rare diseases. Experimental results on the two public datasets demonstrate that our proposed model outperforms state-of-the-art baselines.

## KEYWORDS

Medical report generation, Multimodal Fusion, Cross-modal generation

### ACM Reference Format:

. 2018. Medical Report Generation via Multimodal Spatio-Temporal Fusion. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 INTRODUCTION

Radiological imaging is crucial for medical diagnosis, with the resultant reports being essential for clinical decision-making. However, the increasing demand for these services has significantly burdened radiologists, particularly affecting report quality and increasing the potential for errors [8, 18, 36]. This challenge necessitates the exploration of automatic radiology report generation systems. Current research efforts focus on automating the generation of reports, and enhancing the efficiency and quality of generated reports via

the utilization of multimodal processing methods rooted in the computer vision and natural language processing domains.

Recent works [4, 24–26, 29] for medical report generation leveraged medical tags [16, 46, 47], pretrained models [1, 32], and cross-modal memory networks [6, 35] to enhance the clinical accuracy and quality of the generated reports. Despite these advancements, existing methods mainly produce reports from individual chest radiographs, overlooking the complexity of disease progression and the valuable insights provided by historical reports for current report generation. In clinical practice, handling follow-up patients involves integrating data from past examinations into new reports. To address this, some researchers [3, 33] combined both previous and current medical images to formulate reports. However, approaches that solely rely on tracing historical visual features overlook the significance of patients' past diagnostic reports in textual form. In practice, radiologists analyze both previous and current images to assess disease progression, enhancing earlier reports with updated descriptions of the disease's evolution to compile comprehensive current reports. Referring to prior reports helps doctors compose coherent and consistent report content across different diagnostic instances.

In this paper, we introduce a novel report generation framework designed to align with the workflow of radiologists. This framework compares previous and current images of patients to identify the disease progression, incorporating previous reports and simulating their writing style to generate current ones. Given the challenges presented by textual diversity in cross-modal generation tasks, we propose an enhanced diagnostic report generation method via learning style-agnostic discrete representations of reports and predicting tokens accordingly. To achieve high-quality discrete representations of reports, we have developed the RadFusion module, which conducts multi-granular spatio-temporal fusion of textual and visual features within the patient's clinical context. Additionally, we have developed a novel feature normalization technique to address the challenges posed by the long-tail distribution of token frequencies in current report generation. This technique employs linear projection to adjust the initial semantic features of tokens, ensuring that their utilization is not biased by token frequencies, thus, mitigating the discrepancy in prediction likelihood between high-frequency and low-frequency tokens.

Our proposed method is evaluated on two public datasets, i.e., MIMIC-CXR [17] and MIMIC-ABN [31]. The experimental results demonstrate the improvements of our method in both language quality (ranking the highest across all six natural language generation evaluation metrics) and factual statement accuracy (+ 3% F1-RadGraph and + 1.8% in F1-Chexbert on MIMIC-CXR) over strong baselines, including large language models (LLMs). More importantly, our proposed feature normalization method achieves higher performance gains on infrequent disease types (+ 8.6% F1 on MIMIC-ABN) than on frequent ones (+ 4.7% F1 on MIMIC-ABN). The ethical implications of this improvement are substantial within the clinical domain. While traditional machine learning excels at identifying

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or professional use, not for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://www.acm.org/permissions).

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXXX.XXXXXXX>

2024-04-13 03:24. Page 1 of 1–10.

common patterns, its ability to generalize to less common patterns is relatively limited [28]. In clinical contexts, overlooking uncommon cases to achieve higher accuracy is not feasible. Our feature normalization method addresses this issue by refining wording preferences, thereby enhancing the accuracy and objectivity of factual descriptions of non-common diseases.

In summary, this paper makes the following contributions: (1) We propose a new framework that emulates radiologists' review processes, comparing historical and current images to detect disease progression and integrating previous reports to generate current reports. (2) We develop the RadFusion module to facilitate multi-granular spatio-temporal fusion of textual and visual features within a patient's clinical context. (3) We develop a feature normalization technique to tackle the long-tail distribution challenge in token frequency, improving factual statement accuracy across common and non-common diseases.

## 2 RELATED WORKS

In recent times, artificial intelligence has seen extensive utilization within the medical field [5, 9, 27, 43]. As a task that generates text from images, most medical report generation methods have adopted the encoder-decoder framework popularized in image captioning tasks. Initial efforts [37] utilized an encoder-decoder framework that combines Convolutional Neural Networks (CNN) for image encoding with Recurrent Neural Networks (RNN) for text decoding. Unlike concise image captions, medical reports entail elaborate long texts describing multiple organs and regions. To address this, Jing et al. [47] enhanced the CNN-RNN architecture by incorporating a co-attention module that merges visual and semantic features using disease tags, coupled with a hierarchical LSTM for crafting detailed report paragraphs. Additionally, recognizing the prevalence of normal samples over abnormal ones in medical reports, researchers have focused on mitigating data bias. CMAS-RL [15] refined the textual decoder through a multi-agent system, trying to balance descriptions of both normal and abnormal findings. Contrastive Attention [25] aimed to accentuate critical abnormalities by comparing the subject image against a normal image corpus. HRGR-Agent [22] merged retrieval-based and generative approaches for managing frequently normal and infrequently abnormal sentences.

Given the outstanding performance of pretrained models across various domains, recent works [1, 32] explored fine-tuning pretrained visual encoders and textual decoders for medical report generation. Several researchers leveraged auxiliary signals to guide the generation of medical reports. Li et al. [19] extracted normal and abnormal terms from the MIMIC-CXR dataset to serve as nodes, with edges defined by attention weights between them, thus constructing a knowledge graph. This knowledge graph has been utilized by other researchers [21, 49] as a form of prior domain knowledge to enhance report generation. Additionally, relevant research [24, 48] developed heterogeneous graphs by associating 8 organs with 20 findings, where findings linked to the same organ are interconnected. Liu et al. [24] leveraged global representations derived from pre-retrieved reports within the training corpus to encapsulate domain-specific knowledge. Li et al. [20] dynamically updated the pre-constructed graph to model domain knowledge. In contrast to methods focusing solely on abnormalities, Jain et al. [14]

employed natural language processing tools to extract clinical entity and relation annotations from reports, thereby establishing the comprehensive radiological knowledge graph, RadGraph. Yang et al. [45] introduced general domain knowledge by learning the universal representations of the pre-constructed RadGraph. Other studies have concentrated on enhancing report generation through cross-modal alignment. Chen et al. [6] introduced the Cross-Modal Memory Network (CMM), which employs a shared memory for aligning images and texts, thereby enriching the quality of generated reports. Qin et al. [35] further advanced CMM by integrating reinforcement learning, employing natural language generation metrics as rewards to refine the cross-modal mappings for better image-text alignment. However, these methods treat image-report pairs within datasets as isolated from each other, disregarding the fact that radiologists frequently refer to comparisons with patients' previous examinations in their reports. Addressing this gap, Ban-nur et al. [3] aim to capture historical relevance by comparing patients' previous and current images, facilitating improved cross-modal alignment. Similarly, CXRmate [33] generates reports by integrating patients' previous reports with current images. Recap [10] also contrasts patients' prior and current images to deduce disease progression, thereby enhancing report generation.

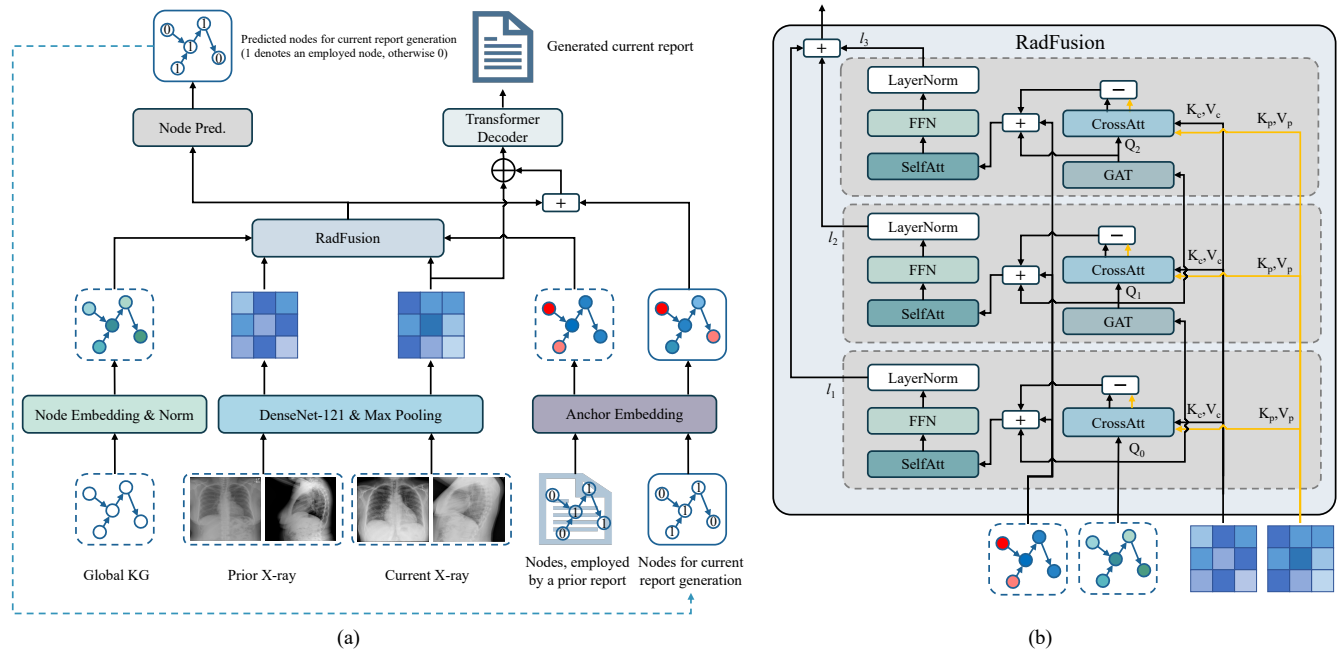
To sum up, despite the progress made in medical report generation, there are several limitations that need to be addressed. (1) Radiologists' varied writing styles pose challenges to cross-modal alignment and the generation of disease-relevant content. (2) Historical reports of follow-up patients, which provide critical references for current clinical assessments of disease progression, were not integrated into the current report generation process, despite their importance in real-world practice. (3) Severe data imbalances complicate the detection and description of non-common diseases, leading to diagnostic biases in the learning and inference processes of machine learning models.

## 3 METHODOLOGY

We propose a spatio-temporal multimodal fusion method (see Figure 1a) to learn two tasks: inclusive node prediction, and radiological report generation. Task 1 learns to identify nodes that should be included in a current report, where the nodes are entities related to different radiographic observations. Task 2 aims to generate reports from the identified nodes. For a follow-up patient, the original input includes the historical examination data (both radiographs and reports) and current radiographs. For those without historical records, the original input is the current radiographs<sup>1</sup>. To generate style-agnostic discrete reports, the input also includes a global radiology knowledge graph for both types of patients.

The global radiology knowledge graph is developed from the training reports. The knowledge graph has nodes representing anatomical or observational entities and directed edges indicating the relationship between nodes. The prior report of a follow-up patient will be also converted into the sub-graph of the global graph. Converting textual reports into graphs (style-agnostic discrete representations) offers the advantage of filtering out stylistic wording variations present in the original contexts from various radiologists.

<sup>1</sup>In Figure 1a, the input does not contain the prior X-ray, the prior report and its associated nodes for new patients.



**Figure 1: (a) The overall framework of our proposed model. (b) The proposed RadFusion module.  $\boxplus$  denotes matrix addition;  $\boxminus$  denotes subtraction;  $\oplus$  denotes concatenation. Colored rounded rectangles denote computational layers with learnable parameters, while the white ones do not have learnable parameters. The graphs and matrices are the input or output of a computational layer.**

These variations can otherwise disrupt the learning of modal alignment between factual information (such as entities) in the text and observations in radiographs.

To mitigate the biased impact of the uneven distribution of node frequencies, i.e., long-tail distribution, on the inclusive node prediction task (Task 1), we introduce a mathematics-explainable feature normalization method that operates on the node feature representations. This method projects the node features into a designated space (see Figure 2), rendering the prediction of nodes insensitive to their frequency attributes. Given input with spatial (e.g., the knowledge graphs) and temporal (e.g., the current and prior information) relationships, we also develop a RadFusion module (see Figure 1b) for the spatio-temporal fusion of the multimodal features (e.g., graphs and images). Ultimately, a comprehensive diagnostic report (Task 2) is generated by synthesizing the knowledge graph, enriched with current chest radiograph features, spatio-temporal and multimodal features, and the predicted discrete nodes.

### 3.1 Knowledge Graph Initialization

**3.1.1 Construction.** To extract diagnostic visual features that are prioritized by radiologists, we propose harnessing radiological knowledge graphs to guide the cross-modal feature alignment and facilitate multimodal feature fusion. Utilizing the tool developed by Wang et al. [14], we structure entities mentioned in the reports in the training set into corresponding knowledge graphs. The nodes of these graphs represent either anatomical entities (e.g., lung, mediastinum) or observational entities (e.g., pneumonia). Edges are

directed and heterogeneous, capturing three types of relationships among entities: modify, located at, and suggestive of. The graphs from individual reports are amalgamated to form a comprehensive global knowledge graph  $G$ . Given the multiplicity of potential relationships between the same pair of entities and the relatively low semantic differentiation among the three types of edge relationships, we opt to disregard edge-type attributes within the global knowledge graph.

**3.1.2 Node Embedding and Normalization.** We plan to identify diagnostic visual features associated with nodes from knowledge graph and predict their inclusion in the report. However, the nodes' frequency distribution reveals a long-tail curve, marked by prevalent common nodes (such as "lung", and "heart") versus infrequent abnormal findings. This distribution inherently biases node prediction towards frequently occurring nodes being identified in the report, while rarer findings are frequently overlooked. As illustrated in Figure 2, in predicting labels for the inclusion of nodes, we establish two anchor points, e.g.,  $S_0$  and  $S_1$ , representing the binary node labels of presence (1) and absence (0) within reports. High-frequency nodes tend to gravitate towards the vector space, associated with label 1, whereas low-frequency nodes are closer to label 0. This phenomenon imparts an inherent bias in the semantic features of nodes towards label prediction.

To mitigate this issue, we propose a method to normalize the node feature representations, eliminating innate label bias in node inclusion prediction while maintaining semantic differentiation. As



depicted in Figure 2, an equidistant hyperplane is initially determined based on anchor points  $S_0$  and  $S_1$ , such that every point on this plane is equidistant to both nodes  $S_0$  and  $S_1$ . Subsequently, a linear projection transformation is applied to map the initial node features onto this equidistant hyperplane. This process ensures that the updated node features  $E'^T$  are unbiased toward any label while preserving their semantic relationships in the vector space. The specific process is outlined as follows:

**Step1: Node and Anchor Embedding.** Node features  $E^G \in \mathcal{R}^{M \times d}$  ( $M$  denotes the number of nodes;  $d$  denotes the dimension of embeddings.) within the constructed knowledge graph were initially randomized. Similarly, the features of the two anchor points  $S_0$  and  $S_1$  are defined as  $s_0$  and  $s_1$ , respectively, both initialized randomly. Both node features and anchor point features are learnable.

**Step2: Equidistant hyperplane.** For any point  $X$  on the equidistant hyperplane with coordinates  $\mathbf{x}$ , the equation of the equidistant hyperplane can be derived as follows:

$$|\mathbf{x} - \mathbf{s}_0|^2 = |\mathbf{x} - \mathbf{s}_1|^2 \quad (1)$$

Expanding the squared distances, we have

$$(\mathbf{x} - \mathbf{s}_0) \cdot (\mathbf{x} - \mathbf{s}_0) = (\mathbf{x} - \mathbf{s}_1) \cdot (\mathbf{x} - \mathbf{s}_1). \quad (2)$$

Simplifying Eq. 2 by expanding and rearranging terms yields

$$2(\mathbf{b} - \mathbf{s}_0) \cdot \mathbf{x} = |\mathbf{s}_1|^2 - |\mathbf{s}_0|^2. \quad (3)$$

**Step3: Equation of the Perpendicular.** For an initial node  $E$ , represented by coordinates  $\mathbf{e}$ , the aim is to project it onto the equidistant hyperplane, ensuring orthogonality between  $E$  and its projection  $E'$ . Using the normal vector  $\mathbf{n} = \mathbf{s}_1 - \mathbf{s}_0$  of the hyperplane, defined by vectors  $\mathbf{s}_0$  and  $\mathbf{s}_1$  of anchor points  $S_0$  and  $S_1$ , we delineate the perpendicular from  $E$  as a linear trajectory guided by  $\mathbf{n}$ .

$$\mathbf{x} = \mathbf{e} + t\mathbf{n}, \quad (4)$$

where  $t$  symbolizes a scalar parameter indicative of the displacement along the direction vector  $\mathbf{n}$ .

**Step4: Target Mapping node  $E'$ .** To solve for the mapping point  $E'$  and its coordinates  $\mathbf{e}'$ , substitute the equation of the perpendicular into the equation defining the equidistant hyperplane, thereby solving for the scalar parameter  $t$ .

$$2(\mathbf{s}_1 - \mathbf{s}_0) \cdot (\mathbf{e} + t(\mathbf{s}_1 - \mathbf{s}_0)) = |\mathbf{s}_1|^2 - |\mathbf{s}_0|^2 \quad (5)$$

Next, expanding this and solving for  $t$  yields

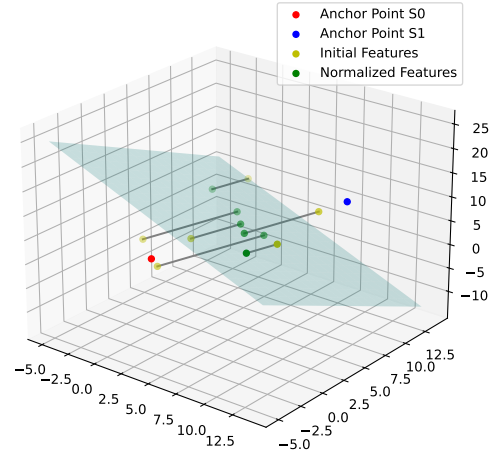
$$2(\mathbf{s}_1 - \mathbf{s}_0) \cdot \mathbf{e} + 2t(\mathbf{s}_1 - \mathbf{s}_0) \cdot (\mathbf{s}_1 - \mathbf{s}_0) = |\mathbf{s}_1|^2 - |\mathbf{s}_0|^2, \quad (6)$$

$$t = \frac{|\mathbf{s}_1|^2 - |\mathbf{s}_0|^2 - 2(\mathbf{s}_1 - \mathbf{s}_0) \cdot \mathbf{e}}{2(\mathbf{s}_1 - \mathbf{s}_0) \cdot (\mathbf{s}_1 - \mathbf{s}_0)}. \quad (7)$$

With  $t$  now determined, the coordinates of  $E'$  can be found by substituting  $t$  back into the equation  $\mathbf{x} = \mathbf{e} + t\mathbf{n}$ , yielding  $\mathbf{e}'$ , the coordinates of the mapping point  $E'$  on the equidistant hyperplane.

$$\mathbf{e}' = \mathbf{e} + \frac{|\mathbf{s}_1|^2 - |\mathbf{s}_0|^2 - 2(\mathbf{s}_1 - \mathbf{s}_0) \cdot \mathbf{e}}{2(\mathbf{s}_1 - \mathbf{s}_0) \cdot (\mathbf{s}_1 - \mathbf{s}_0)} \mathbf{n}, \quad (8)$$

where  $\mathbf{e} \in E^G$ . This formula effectively yields  $\mathbf{e}'$ , pinpointing the location of  $E'$  on the equidistant hyperplane where the line segment joining  $E$  to  $E'$  is perpendicular to the hyperplane, thus satisfying the geometric condition of equidistance from  $E$  to points  $S_0$  and  $S_1$ .



**Figure 2: Feature normalization visualization.** The initial features (light green dots) of the global knowledge graph nodes are projected onto a hyperplane using our proposed feature normalization method. The normalized features (dark green dots) of the nodes are equidistant from both anchor points, ensuring unbiased representations for predicting the inclusion of nodes in the current report generation.

## 3.2 RadGraph guided Spatio-temporal Multimodal Hierarchical Fusion

In alignment with radiologists' workflow, their diagnostic process initiates with a review of patients' previous examination data to comprehend the current state of the patient's condition. Then, they analyze the discrepancies between the current and prior examination results to discern critical information on disease progression, forming the basis for report generation. To emulate this procedural framework, we propose the integration of knowledge graphs to guide the fusion of clinical context and current chest X-rays, thereby capturing disease progression for precise diagnostic reporting.

**3.2.1 Visual Encoding.** A single chest radiological examination generates one or more chest radiographs  $I = \{I_1, I_2, \dots, I_m\}$  and a corresponding diagnostic report  $R$ . For a patient's current examination, we initially encode all produced images using DenseNet-121 [12], applying max pooling to ascertain the present chest radiograph features  $\mathbf{I}^C \in \mathcal{R}^{8 \times 8 \times 1024}$ . Likewise, we acquire prior radiograph features  $\mathbf{I}^{\text{Pr}} \in \mathcal{R}^{8 \times 8 \times 1024}$  from the patient's former examination.

**3.2.2 Comparison of Prior and Current X-rays.** Radiologists typically examine important disease-related regions in patients' successive chest radiographs to assess disease progression, essentially comparing visual features pertinent to report content. Accordingly, we utilize the constructed knowledge graphs to extract salient visual features correlated with report narratives from both previous and current X-rays, followed by a comparative analysis of the extracted features.

Utilizing node features as queries, and relative visual features as keys and values, cross-attention mechanisms are employed to

extract node-relevant significant features from visual features.

$$\mathbf{E}^C = \text{CrossAtt}(\mathbf{E}'^G, \mathbf{I}^C) \quad (9)$$

$$\mathbf{E}^{Pr} = \text{CrossAtt}(\mathbf{E}'^G, \mathbf{I}^{Pr}) \quad (10)$$

$$\text{CrossAtt}(\mathbf{X}, \mathbf{Y}) \begin{cases} \text{head}_i = \text{softmax}\left(\frac{\mathbf{X}\mathbf{Y}^T}{\sqrt{d}}\right)\mathbf{Y} \\ \mathbf{X}' = \text{Concat}(\text{head}_1, \dots, \text{head}_i)\mathbf{W} \end{cases} \quad (11)$$

where  $\mathbf{E}'^G$  represents the normalized node features;  $d$  is the dimension of features;  $\mathbf{W}$  is a learnable weight matrix;  $\mathbf{E}^{Pr}$  and  $\mathbf{E}^C$  represent the extracted report-related key visual features of prior X-ray and current X-ray, respectively. Thus, the disease progression features, i.e., the difference of key features between current and previous X-rays can be represented as  $\mathbf{E}^I = \mathbf{E}^C - \mathbf{E}^{Pr}$ .

**3.2.3 Spatio-temporal Multimodal Hierarchical Fusion.** To integrate the clinical context of follow-up patients, we merge knowledge graph embeddings with anchor embeddings to encode the information from prior reports, thus representing the current state features  $\mathbf{E}^S \in \mathbb{R}^{N \times d}$  of patients. Each node in the knowledge graph is allocated a unique status determined by its occurrence in the patient's antecedent examination report: nodes reported previously are encoded with the anchor embedding  $\mathbf{s}_1$  corresponding to label 1, while unreported nodes receive the anchor embedding  $\mathbf{s}_0$  of label 0. For new patients, the node features are designated as null. This approach allows the initial features of the knowledge graph to effectively reflect the patient's original health condition, facilitating the integration of information from prior reports while efficiently distinguishing between follow-up and new patients.

Based on the foundational state of patients, integrating disease progression features, e.g.,  $\mathbf{E}^I$ ,  $\mathbf{E}'^G$ , and  $\mathbf{E}^S$ , enables the current condition feature  $\mathbf{E}$  of the patient

$$\mathbf{E} = \text{LayerNorm}(\mathbf{E}^I + \mathbf{E}'^G + \mathbf{E}^S), \quad (12)$$

where  $\text{LayerNorm}(\cdot)$  is from the work of [41].

To enhance the integration of inter-node dependencies, we use a self-attention mechanism to merge global contextual semantic information, thereby refining node representations.

$$\mathbf{E}' = \text{LayerNorm}(\text{FFN}(\text{SelfAtt}(\mathbf{E}))) \quad (13)$$

$$\text{SelfAtt}(\mathbf{E}) = \text{CrossAtt}(\mathbf{E}, \mathbf{E}) \quad (14)$$

$$\text{FFN}(\mathbf{E}) = \max(0, \mathbf{E}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (15)$$

However, considering the initial node features are too granular in semantic detail to compose complete textual semantics independently, we aggregate neighboring node features through Graph Attention Networks (GAT) to incorporate spatial context information within the knowledge graph. For node  $i$ , the updated embedding is defined as:

$$\mathbf{e}_i^{(l)} = \sigma\left(\sum_{j \in N_i} \alpha_{ij} \mathbf{W}_j^{(l-1)}\right) \quad (16)$$

$$\begin{aligned} \alpha_{ij} &= \text{softmax}(\text{att}(\mathbf{W}_{e_i}, \mathbf{W}_{e_j})) \\ &= \frac{\exp(\text{LeakyReLU}(\mathbf{a}[\mathbf{W}_{e_i} \oplus \mathbf{W}_{e_j}]))}{\sum_{k \in N_i} \exp(\text{LeakyReLU}(\mathbf{a}[\mathbf{W}_{e_i} \oplus \mathbf{W}_{e_k}]))}, \end{aligned} \quad (17)$$

where  $\mathbf{e}_i \in \mathbf{E}$ ;  $\sigma$  denotes an activation function;  $N_i$  represents neighbor nodes of node  $i$ ;  $\mathbf{W} \in \mathbb{R}^{d' \times d}$  is a learnable weight matrix;  $\oplus$  is

the concatenation operation;  $\text{att}$  is a feedforward neural network, parameterized by a weight vector  $\mathbf{a} \in \mathbb{R}^{2d'}$ .

As illustrated, starting from the second layer, each layer is composed of a GAT, a cross-attention mechanism, and a self-attention mechanism. The second layer utilizes GAT to update the node features with one-hop neighborhood features, subsequently extracting relevant visual features under the guidance of updated node features. The third layer's update involves feature extraction utilizing two-hops neighbors, and so forth, facilitating the acquisition of visual features extracted according to varying granularities of textual semantics in different layers.

$$\text{RradFusion} \begin{cases} \mathbf{E}'^{G(l)} = \text{GAT}(\mathbf{E}'^{G(l-1)}) \\ \mathbf{E}^{c(l)} = \text{CrossAtt}(\mathbf{E}'^{G(l)}, \mathbf{I}^C) \\ \mathbf{E}^{p(l)} = \text{CrossAtt}(\mathbf{E}'^{G(l)}, \mathbf{I}^{Pr}) \\ \mathbf{E}^{I(l)} = \mathbf{E}^{C(l)} - \mathbf{E}^{Pr(l)} \\ \mathbf{E}^{(l)} = \text{LayerNorm}(\mathbf{E}^{I(l)} + \mathbf{E}'^{G(l)} + \mathbf{E}^S) \\ \mathbf{E}'^{(l)} = \text{LayerNorm}(\text{FFN}(\text{SelfAtt}(\mathbf{E}^{(l)}))) \end{cases} \quad (18)$$

where  $l \in \{2, 3, \dots, L\}$ . The output features from each layer are aggregated and subjected to layer normalization, yielding an updated knowledge graph fused spatio-temporal multimodal features.

### 3.3 Inclusive Node Prediction

The presence of each node within the report is predicted based on its distance to designated anchor nodes, thereby enabling the identification of relevant tokens within the diagnostic report.

$$\mathbf{E}^0 = \text{LayerNorm}\left(\sum_{L=1}^L \mathbf{E}'^{(L)}\right) \quad (19)$$

We define the training loss according to the distances between updated node embeddings and two anchor points.

$$\mathbf{D} = [|\mathbf{e}' - \mathbf{s}_1|, |\mathbf{e}' - \mathbf{s}_0|] \quad (20)$$

$$\mathbf{P} = \text{Softmax}(\mathbf{D}) \quad (21)$$

where  $\mathbf{s}_0$  and  $\mathbf{s}_1$  represent embeddings of the anchor points,  $\mathbf{e}' \in \mathbf{E}^0$ . The loss function is the cross-entropy loss

$$L_c = -\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^2 y_{ij}^n \log(p_{ij}^n), \quad (22)$$

where  $y_{ij}^n \in \{0, 1\}$  and  $p_{ij}^n \in [0, 1]$  are the ground-truth label and predicted label of the  $i$ -th node, respectively. Updated node features are also used to predict their presence in the current report. If  $p_{i1}^n$  exceeds a predefined threshold  $\text{Thred}$ , the  $i$ -th node is classified as label 1; otherwise, it is classified as label 0.

### 3.4 Report Generation

We employ a constructed global knowledge graph to guide the spatio-temporal multimodal fusion of patients' previous examinations, including chest radiographs and diagnostic reports, as well as current chest radiographs. After the multimodal fusion, we predict the nodes contained in the current report based on the integrated node features. Subsequently, these predicted nodes are utilized to generate diagnostic reports. To ensure that no critical details in the current chest X-ray images are overlooked, we integrate the

features of the X-rays and the nodes during the report generation process. The input of the decoder is defined by

$$\mathbf{H} = [\mathbf{I}^c \oplus (\mathbf{E}^o + \mathbf{E}^S)] \quad (23)$$

If the node label is 1,  $\mathbf{E}^S = \mathbf{s}_1$ ; otherwise,  $\mathbf{E}^S = \mathbf{s}_0$ . Finally, we use a Transformer decoder to generate diagnostic reports.

$$R = TF - Decoder(\mathbf{H}) \quad (24)$$

The decoder is optimized with cross-entropy loss to maximize the conditional log-likelihood.

$$L_g = -\frac{1}{l} \sum_{i=1}^l \sum_{j=1}^v y_{ij} \log(p_{ij}), \quad (25)$$

where  $l$  is the length of generated report;  $v$  represents vocabulary size;  $p_{ij}$  is the probability of the  $i$ -th word of the report is the  $j$ -th word in the vocabulary;  $y_{ij}$  is the corresponding ground truth. The overall loss function is defined as:

$$L = L_c + L_g \quad (26)$$

## 4 EXPERIMENT

### 4.1 Datasets

**MIMIC-CXR** [17] is a large dataset of 227,835 imaging studies involving 65,379 patients who presented to the emergency department of Beth Israel Deaconess Medical Center between 2011 and 2016. We use the official split and exclude studies without X-ray images or missing findings. **MIMIC-ABN** [31] is a subset of MIMIC-CXR proposed. Reports in MIMIC-ABN only contain abnormal sentences. We partitioned our dataset into training, validation, and testing sets following the strongest baseline, Recap [10].

### 4.2 Evaluation Metrics

To thoroughly assess the generated reports' quality, we utilize both natural language generation (NLG) metrics and factual correctness (FC) metrics. The NLG metrics we adopt include BLUE [34], ROUGE-L [23] and METEOR [2]. These metrics are designed to evaluate the descriptive accuracy of the reports by comparing them with the ground truth reports. On the other hand, FC metrics, specifically the factual-oriented metric F1-RadGraph and the clinical efficacy metric F1-Score (which leverages 14 observations from CheXbert [38]), are implemented to assess the reports' accuracy in depicting clinical abnormalities.

### 4.3 Implementation details

For our study, the PyTorch framework facilitated the model's development, which was trained on an NVIDIA Tesla V100 GPU with 32GB of memory. Input images were resized to 256x256 and processed via a pre-trained DenseNet121 to extract features, yielding a 1024x8x8 feature map. Both self-attention and cross-attention are 8-head multihead attention. A 12-layer Transformer with four attention heads was utilized for decoding. We used an Adam optimizer with a learning rate of  $3e-4$ , 0.01 weight decay, 0.1 dropout, and a batch size of 8. Embedding dimensions were set at 256. Node prediction counts were 255 for MIMIC-ABN and 769 for MIMIC-CXR, with a node prediction threshold of 0.17. Hyperparameters were refined based on validation set performance.

### 4.4 Baselines

We conducted a comparative analysis against a wide range of state-of-the-art baselines in medical report generation. This comparison included models with a focus on cross-modal alignment such as R2Gen [7], CMN [6], and Aligntransformer [46]. We also assessed models that utilize reinforcement learning to optimize fact-related rewards, notably  $M^2\text{fact}_{ENTNL}$  [30], and others like CvT-212DistilGPT2 [32] that employ fine-tuning strategies with pretrained models. Additionally, models integrating domain knowledge such as PPKED [24], M2KG [44], KiUT [13], and ORGAN [11], as well as those incorporating historical examination data like CXRmate [33] and Recap [10], were examined. Our comparative analysis also encompassed medical report generation techniques based on LLMs, specifically XrayGPT [39] and MedPaLM [40], to provide a holistic understanding of our model's standing within the current technological landscape. The metrics reported in the original papers of these models serve as reference benchmarks for our comparative analysis.

### 4.5 Main results

**4.5.1 Language quality.** As presented in Table 1, our model achieves superior performance compared to the baselines on both MIMIC-ABN, primarily involving first-visit patients, and MIMIC-CXR, which includes numerous follow-up cases. This underscores our model's capability to generate accurate reports for diverse patient types. Compared to large-scale model-based methods e.g., XrayGPT [39] and Med-PaLM [40], our model shows significant superiority in NLG metrics, surpassing Med-PaLM(562B) by 11.9% in BLEU-1 and 3% in Rouge-L on MIMIC-CXR. Relative to CXRmate [33], which also incorporates historical patient data, our approach shows enhancements of 5% in BLEU-4 and 4.3% in Rouge-L. Although Recap [10] performs best among all the baselines, its overall performance still lags behind our model, particularly in Rouge-L, where we exceed it by 1.7% and 2.5% on MIMIC-CXR and MIMIC-ABN, respectively. Despite Organ [11] and Recap [10] showing significant advantages over other baselines by integrating complex graph-building processes, this approach causes their models to be highly dependent on the quality of these constructions, reducing their generalizability.

**4.5.2 Factual correctness.** We evaluate the factual accuracy of the generated reports from different models in Table 1. Our model reaches the highest scores in both factual-oriented metric, such as F1-RadGraph(ER) and clinical efficacy metric, such as F1-Chexbert across the two datasets, cementing its preeminence in generating factually correct reports. When juxtaposed with the leading baseline  $M^2\text{fact}_{ENTNL}$  [30] in terms of F1-RadGraph performance, our model exhibits a 3% improvement. While  $M^2\text{fact}_{ENTNL}$  excels in F1-RadGraph, its performance is notably lower in F1-Chexbert. This discrepancy suggests that the optimization of  $M^2\text{fact}_{ENTNL}$  using F1-RadGraph-related rewards may inflate its performance on this metric, potentially masking its true clinical accuracy. Compared with the best-performing Recap on F1-Chexbert, we improved 1.8% and 0.8% on MIMIC-CXR and MIMIC-ABN respectively. This advantage is particularly significant on MIMIC-CXR, which includes many follow-up patients, reflecting our model's efficacy in integrating historical patient data.



**Table 1: Comparison of NLG and FC metrics on MIMIC-ABN and MIMIC-CXR testing sets. B denotes BLEU scores; MTR denotes METEOR; R-L denotes ROUGE-L. F1-Rad denotes 1-RadGraph; F1-CE denotes F1-CheXbert; \* denotes that the improvements of our model over state-of-the-art baselines on major metrics are statistically significant, based on two-tailed t-tests ( $p < 0.001$ ).**

Datasets	Method	NLG Metrics							FC Metrics	
		B-1	B-2	B-3	B-4	MTR	R-L	AVG	F1-Rad	F1-CE
MIMIC-ABN	R2Gen [7]	0.290	0.157	0.093	0.061	0.105	0.208	0.152	-	0.272
	CMN [6]	0.264	0.140	0.085	0.056	0.098	0.212	0.142	-	0.280
	ORGAN [11]	0.314	0.180	0.114	0.078	0.120	0.234	0.173	-	0.293
	Recap [10]	0.321	0.182	0.116	0.080	0.120	0.223	0.174	-	0.305
	<b>Ours</b>	<b>0.322</b>	<b>0.192</b>	<b>0.125</b>	<b>0.085</b>	<b>0.128</b>	<b>0.248</b>	<b>0.183*</b>	<b>0.227</b>	<b>0.313*</b>
MIMIC-CXR	R2Gen [7]	0.353	0.218	0.145	0.103	0.142	0.277	0.206	0.196	0.276
	CMN [6]	0.353	0.218	0.148	0.106	0.142	0.278	0.207	0.218	0.278
	Aligntransformer [46]	0.378	0.235	0.156	0.112	0.158	0.283	0.220	-	-
	CvT-212DistilGPT2 [32]	0.394	0.249	0.172	0.127	0.155	0.287	0.230	0.219	0.258
	$M^2\text{fact}_{ENTNL}$ [30]	-	-	-	0.083	-	0.269	0.218	0.320	0.311
	XrayGPT(7B) [39]	0.128	0.045	0.014	0.004	0.079	0.111	0.064	-	-
	Med-PaLM(12B) [40]	0.309	-	-	0.104	-	0.262	0.225	0.252	0.373
	Med-PaLM(562B) [40]	0.317	-	-	0.115	-	0.275	0.235	0.261	0.378
	PPKED [24]	0.360	0.224	0.149	0.106	0.149	0.284	0.212	-	-
	M2KG [44]	0.386	0.237	0.157	0.111	-	0.274	0.233	-	0.352
	KiUT [13]	0.393	0.243	0.159	0.113	0.160	0.285	0.225	-	0.321
	ORGAN [11]	0.407	0.256	0.172	0.123	0.162	0.293	0.235	-	0.385
	CXRmate	-	-	-	0.079	-	0.262	0.170	0.272	0.357
Recap [10]	0.429	0.267	0.177	0.125	0.168	0.288	0.242	-	0.393	
<b>Ours</b>	<b>0.436</b>	<b>0.275</b>	<b>0.184</b>	<b>0.129</b>	<b>0.177</b>	<b>0.305</b>	<b>0.251*</b>	<b>0.350*</b>	<b>0.411*</b>	

**Table 2: Ablation study, evaluated on validation sets.**

Datasets	Variant	AvgNLG	F1-Rad	F1-CE
MIMIC-ABN	Full model	0.181	0.220	0.317
	w/o hist.info.	0.180	0.217	0.314
	w/o RadFusion	0.174	0.207	0.294
	w/o FeatureNorm	0.173	0.197	0.276
MIMIC-CXR	Full model	0.253	0.351	0.413
	w/o hist.info.	0.248	0.333	0.401
	w/o RadFusion	0.244	0.328	0.389
	w/o FeatureNorm	0.238	0.314	0.376

**Table 3: The effectiveness analysis of our proposed feature normalization (FN) method on diseases with different label frequencies, evaluated on the MIMIC-ABN validation set.**

Variant	High-freq. diseases			Low-freq. diseases		
	P-CE	R-CE	F1-CE	P-CE	R-CE	F1-CE
Full model	0.506	0.549	0.526	0.320	0.281	0.264
w/o FN	0.480	0.502	0.479	0.256	0.193	0.178
Gains	+0.026	+0.047	+0.047	+0.064	+0.088	+0.086

## 4.6 Ablation study

Our ablation study used validation sets to avoid overfitting the model to the testing set during hyperparameter tuning and model architecture decisions, ensuring the final evaluation is unbiased.

**4.6.1 Effect of historical information.** To examine the impact of integrating patients' historical examination data on the performance of the model, we conducted an experiment by omitting the historical information of follow-up patients. As indicated in Table 2, on MIMIC-ABN, the exclusion of historical data did not significantly

affect the model's performance, which can be attributed to the fact that less than 10% of the patients on MIMIC-ABN are follow-up cases. However, on the MIMIC-CXR dataset, which contains a higher proportion of follow-up patients, the model's performance declined upon the removal of historical data, underscoring the significance of historical information for follow-up patients.

**4.6.2 Effect of RadFusion.** To validate the effectiveness of the proposed RadFusion module, we substituted it with the transformer-based Fusion module from MedKLiP [42] for integrating node and visual features. Given that the Fusion module does not account for historical information, we applied the same method of incorporating historical features as with RadFusion. As illustrated in Table 2, the model's performance deteriorated upon replacing the RadFusion module, underscoring the efficacy of the RadFusion module.

**4.6.3 Effect of Feature Normalization.** Table 2 shows that our proposed feature normalization method is the most significant among our technical innovations. To assess its effectiveness on diseases with varying frequencies, we evaluated the clinical accuracy of the generated reports in describing 14 abnormal observations of MIMIC-ABN. The diseases were categorized into two groups based on the frequency of occurrence of these observations. Those above the average frequency were considered common diseases with high frequency; Those below were classified as non-common diseases with low frequency. In Table 3, the performance improvements in low-frequency diseases are about twice as pronounced as those in high-frequency ones. This finding underscores the significance of our feature normalization method in enhancing clinical accuracy, particularly for less common diseases. This is because we mitigate the challenge posed by the long-tail distribution of tokens by addressing the frequency biases in the node feature representations.

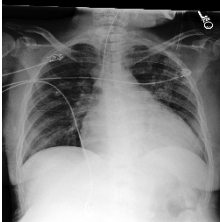
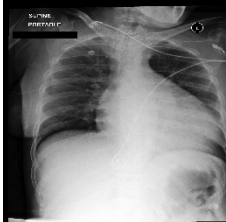

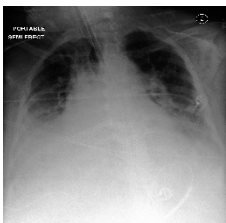
Chest X-ray	Prior X-ray	Prior report	Ground truth and generated reports
		<p>portable supine radiograph of the chest. there is <b>diffuse indistinctness</b> of the <b>pulmonary vasculature</b>, suggestive of <b>mild interstitial pulmonary edema</b>. although the <b>heart size</b> is likely exaggerated by the technique, there is <b>moderate cardiomegaly</b> which is <b>stable</b> from &lt;unk&gt; but not present on &lt;unk&gt;. the <b>lungs</b> are <b>clear</b>, the there is no <b>pneumothorax</b> or <b>pleural effusion</b>. <b>chronic rightward</b> tracheal <b>deviation</b> secondary to thyromegaly</p>	<p>Ground truth as compared to prior chest radiograph from earlier today, there has been interval placement of an <b>endotracheal tube</b>, terminating &lt;num&gt; cm above the <b>carina</b>, the <b>cardiac silhouette</b> is <b>enlarged</b>, as before, there is <b>mild pulmonary edema</b>, lungs are <b>otherwise clear</b>, there is no <b>local consolidation</b>, <b>pneumothorax</b> or <b>pleural effusion</b>.</p> <p>Generated reports the <b>endotracheal tube</b> terminates &lt;num&gt; cm above the <b>carina</b>, the <b>lungs</b> are well - expanded, there is <b>mild pulmonary vascular congestion and interstitial edema</b>, there is no <b>local consolidation</b>, <b>pleural effusion</b> or <b>pneumothorax</b>, the <b>cardiac silhouette</b> is <b>enlarged</b> but <b>stable</b>, the <b>mediastinal silhouette</b> is <b>unchanged</b>, the <b>left</b> upper extremity <b>tube</b> is in the <b>stomach</b>.</p>
		<p>as compared to the previous radiograph, the patient has received a <b>new dobhoff</b> tube. the tip of the <b>tube</b> projects over the <b>middle parts</b> of the <b>stomach</b>, the <b>course</b> of the <b>tube</b> is <b>unremarkable</b>, there is no evidence of complications, notably no <b>pneumothorax</b>, otherwise, the radiographic appearance of the <b>thoracic</b> organs is similar to the previous examination.</p>	<p>Ground truth as compared to the previous radiograph, there is no relevant change. the <b>tip</b> of the <b>endotracheal tube</b> projects &lt;num&gt; cm above the <b>carina</b>, the <b>tube</b> could be advanced by &lt;num&gt; cm, <b>unchanged moderate-to-severe cardiomegaly</b> with <b>signs of mild-to-moderate pulmonary edema</b> and a moderate right-sided <b>atelectasis</b>, <b>bilateral areas of atelectasis</b> at the <b>lung bases</b>, no <b>pneumothorax</b>, <b>right</b> <b>neck line</b> in <b>unchanged position</b>.</p> <p>Generated reports as compared to the previous radiograph, the monitoring and support devices, including the <b>endotracheal tube</b> and the <b>tip</b> of the <b>tube</b> are in <b>unchanged position</b>, the <b>tip</b> of the <b>tube</b> projects &lt;num&gt; cm above the <b>carina</b>, <b>unchanged moderate cardiomegaly with mild pulmonary edema</b> and a moderate <b>atelectasis</b>, <b>mild atelectasis</b> in the left lower lung, no pneumothorax.</p>

Figure 3: Case study of two follow-up-visit samples. Predicted token nodes in reports are highlighted in green, and colored text in reports represents the abnormalities.

### 4.7 Case Study

A qualitative analysis was conducted on two follow-up-visit samples from the MIMIC-CXR dataset in Figure 3. Our model performs a comparative analysis of patients’ previous and current chest radiographs while integrating historical report information by discretizing prior reports into a series of tokens, thereby synthesizing a comprehensive diagnostic report. Within this illustration, crucial tokens predicted by our model are accentuated with green highlighting, effectively encompassing the principal semantic content of the report. Moreover, there is a significant overlap between the tokens within the generated report and the ground truth, exemplified by terms such as “endotracheal”, “tube”, “pneumothorax”, “effusion”, “edema”, and “enlarged”. Abnormal descriptions within the report are marked in colored, with uniform coloring denoting similar disease types, underscoring our model’s accuracy in delineating multiple abnormalities. For instance, in the context of support devices, the report specifies: “the endotracheal tube terminates <num> cm above the carina”; regarding cardiomegaly, it notes: “the cardiac silhouette is enlarged but stable”; and in the case of edema, it mentions: “there is mild pulmonary vascular congestion and interstitial edema.” The generated reports also covers long, complex sentences describing multiple abnormalities, such as “unchanged moderate cardiomegaly with mild pulmonary edema and a moderate right pleural effusion.”

### 5 CONCLUSION

In this paper, we present a medical report generation framework that emulates radiologists’ workflows by integrating both historical and current patient data, enabling disease progression tracking and personalized report creation. Our method tackles textual diversity

in cross-modal tasks through style-agnostic representations and advanced token prediction. We introduce a novel spatio-temporal fusion method for integrating textual and visual data across multiple granularities and apply a feature normalization technique to address biases from long-tail frequency distributions, enhancing accuracy in rare disease reporting. Experimental results on two public datasets demonstrate our model’s superiority over state-of-the-art baselines.

**Limitation and Future Work.** Our enhancements include discrete node prediction to enrich reports. However, the typical scarcity of nodes in reports leads to significant data imbalances, impacting prediction accuracy. We address this by excluding some low-frequency nodes, which risks omitting critical semantic information. Future work will aim to balance node prediction tasks more effectively, ensuring the comprehensive capture of crucial diagnostic details to improve the reliability of automated medical reporting.

### 6 ETHICS STATEMENT

The datasets used in this work, MIMIC-ABN [31] and MIMIC-CXR [17], are publicly available and have been automatically de-identified to protect patient privacy. Our review confirms that the usage of these datasets poses no substantial ethical risks. However, despite the model’s ability to enhance the factual accuracy of medical reports, it has not yet reached a level suitable for clinical application. The generated reports may occasionally include inaccurate or biased observations and diagnostic suggestions. We strongly recommend that healthcare professionals critically evaluate and validate the model’s outputs before any clinical application to ensure patient safety and care quality.



## REFERENCES

- [1] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked* 24 (2021), 100557.
- [2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (Eds.). Association for Computational Linguistics, 65–72. <https://aclanthology.org/W05-0909/>
- [3] Shruthi Bannur, Stephanie L. Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya V. Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 15016–15027. <https://doi.org/10.1109/CVPR52729.2023.01442>
- [4] Adrian Brady, Risteárd Ó Laoide, Peter McCarthy, and Ronan McDermott. 2012. Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster medical journal* 81, 1 (2012), 3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3609674/>
- [5] Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven Pillars for the Future of Artificial Intelligence. *IEEE Intelligent Systems* 38, 6 (2023), 62–69. <https://doi.org/10.1109/MIS.2023.3329745>
- [6] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 5904–5914. <https://doi.org/10.18653/v1/2021.acl-long.459>
- [7] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1439–1449. <https://doi.org/10.18653/v1/2020.emnlp-main.112>
- [8] Louke Delrue, Robert Gosselin, Bart Ilsen, An Van Landeghem, Johan de Mey, and Philippe Duyck. 2011. Difficulties in the interpretation of chest radiography. *Comparative interpretation of CT and standard radiography of the chest* (2011), 27–49. [https://doi.org/10.1007/978-3-540-79942-9\\_2](https://doi.org/10.1007/978-3-540-79942-9_2)
- [9] Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 94–104. <https://aclanthology.org/2022.coling-1.9>
- [10] Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023. RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 2134–2147. <https://aclanthology.org/2023.findings-emnlp.140>
- [11] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 8108–8122. <https://doi.org/10.18653/v1/2023.acl-long.451>
- [12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.
- [13] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. KIUT: Knowledge-injected U-Transformer for Radiology Report Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 19809–19818. <https://doi.org/10.1109/CVPR52729.2023.01897>
- [14] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre J. Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c8ffe9a587b126f152ed3d89a146b445-Abstract-round1.html>
- [15] Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6570–6580. <https://doi.org/10.18653/v1/p19-1657>
- [16] Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*. Association for Computational Linguistics, 2577–2586. <https://doi.org/10.18653/v1/P18-1240>
- [17] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR abs/1901.07042* (2019). <http://arxiv.org/abs/1901.07042>
- [18] Jeffrey P Kanne, Nisa Thoongsuwan, and Eric J Stern. 2005. Common errors and pitfalls in interpretation of the adult chest radiograph. *Clinical Pulmonary Medicine* 12, 2 (2005), 97–114. <https://doi.org/10.1097/01.cpm.0000156704.33941.e2>
- [19] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*. AAAI Press, 6666–6673. <https://doi.org/10.1609/aaai.v33i01.33016666>
- [20] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3334–3343.
- [21] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2023. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web (WWW)* 26, 1 (2023), 253–270. <https://doi.org/10.1007/S11280-022-01013-6>
- [22] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*. 1537–1547. <https://proceedings.neurips.cc/paper/2018/hash/e07413354875be01a996dc560274708e-Abstract.html>
- [23] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://aclanthology.org/W04-1013>
- [24] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13753–13762. <https://doi.org/10.1109/CVPR46437.2021.01354>
- [25] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive Attention for Automatic Chest X-ray Report Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 269–280. <https://doi.org/10.18653/v1/2021.findings-acl.23>
- [26] Jenny X. Liu, Yevgeniy Goryakin, Akiko Maeda, Tim Bruckner, and Richard Scheffler. 2017. Global health workforce labor market projections for 2030. *Human resources for health* 15, 1 (2017), 1–12. <https://doi.org/10.1186/s12960-017-0187-2>
- [27] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTeval: A Survey on Assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. Torino, Italia.
- [28] Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word Embedding and WordNet Based Metaphor Identification and Interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1. 1222–1231.
- [29] Robert J McDonald, Kara M Schwartz, Laurence J Eckel, Felix E Diehn, Christopher H Hunt, Brian J Bartholmai, Bradley J Erickson, and David F Kallmes. 2015. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic radiology* 22, 9 (2015), 1191–1198. <https://doi.org/10.1016/j.acra.2015.05.007>
- [30] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5288–5304. <https://doi.org/10.18653/v1/2021.naacl-main.416>
- [31] Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian J. McAuley. 2020. Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1954–1960. <https://doi.org/10.18653/v1/2020.FINDINGS-EMNLP.176>
- [32] Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2022. Improving Chest X-Ray Report Generation by Leveraging Warm-Starting. *CoRR abs/2201.09405* (2022). <https://arxiv.org/abs/2201.09405>
- [33] Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Longitudinal Data and a Semantic Similarity Reward for Chest X-Ray Report Generation. *CoRR abs/2307.09758* (2023). <https://doi.org/10.48550/arXiv.2307.09758>
- [34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of*

- 1045 the 40th Annual Meeting of the Association for Computational Linguistics. ACL, 311–318. <https://doi.org/10.3115/1073083.1073135>
- 1046 [35] Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 448–458.
- 1047 [36] I Satia, S Bashagha, A Bibi, R Ahmed, S Mellor, and F Zaman. 2013. Assessing the accuracy and certainty in interpreting chest X-rays in the medical division. *Clinical medicine* 13, 4 (2013), 349. <https://doi.org/10.7861/clinmedicine.13-4-349>
- 1048 [37] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2497–2506. <https://doi.org/10.1109/CVPR.2016.274>
- 1049 [38] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1500–1519. <https://doi.org/10.18653/V1/2020.EMNLP-MAIN.117>
- 1050 [39] Omkar Thawakar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models. *CoRR abs/2306.07971* (2023). <https://doi.org/10.48550/ARXIV.2306.07971> arXiv:2306.07971
- 1051 [40] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David J. Fleet, Philip Andrew Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Nataraajan. 2023. Towards Generalist Biomedical AI. *CoRR abs/2307.14334* (2023). <https://doi.org/10.48550/ARXIV.2307.14334>
- 1052 [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
- 1053 [42] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023*. IEEE, 21315–21326. <https://doi.org/10.1109/ICCV51070.2023.01954>
- 1054 [43] Jialun Wu, Kai He, Rui Mao, Chen Li, and Erik Cambria. 2023. MEGACare: Knowledge-guided Multi-view Hypergraph Predictive Framework for Healthcare. *Information Fusion* 100 (2023), 101939. <https://doi.org/doi.org/10.1016/j.inffus.2023.101939>
- 1055 [44] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Anal.* 86 (2023), 102798. <https://doi.org/10.1016/j.media.2023.102798>
- 1056 [45] Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Anal.* 80 (2022), 102510. <https://doi.org/10.1016/j.media.2022.102510>
- 1057 [46] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI (Lecture Notes in Computer Science, Vol. 12903)*. Springer, 72–82. [https://doi.org/10.1007/978-3-030-87199-4\\_7](https://doi.org/10.1007/978-3-030-87199-4_7)
- 1058 [47] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment. In *Medical Image Computing and Computer Assisted Intervention - MICCAI (Lecture Notes in Computer Science, Vol. 11769)*. Springer, 721–729. [https://doi.org/10.1007/978-3-030-32226-7\\_80](https://doi.org/10.1007/978-3-030-32226-7_80)
- 1059 [48] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. 2020. When Radiology Report Generation Meets Knowledge Graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, 12910–12917. <https://ojs.aaai.org/index.php/AAAI/article/view/6989>
- 1060 [49] Haifeng Zhao, Jie Chen, Lili Huang, Tingting Yang, Wanhai Ding, and Chuanfu Li. 2021. Automatic Generation of Medical Report with Knowledge Graph. In *ICCP 21: 10th International Conference on Computing and Pattern Recognition, Shanghai, China, October 15 - 17, 2021*. ACM, 1. <https://doi.org/10.1145/3497623.3497658>
- 1061
- 1062
- 1063
- 1064
- 1065
- 1066
- 1067
- 1068
- 1069
- 1070
- 1071
- 1072
- 1073
- 1074
- 1075
- 1076
- 1077
- 1078
- 1079
- 1080
- 1081
- 1082
- 1083
- 1084
- 1085
- 1086
- 1087
- 1088
- 1089
- 1090
- 1091
- 1092
- 1093
- 1094
- 1095
- 1096
- 1097
- 1098
- 1099
- 1100
- 1101
- 1102
- 1103
- 1104
- 1105
- 1106
- 1107
- 1108
- 1109
- 1110
- 1111
- 1112
- 1113
- 1114
- 1115
- 1116
- 1117
- 1118
- 1119
- 1120
- 1121
- 1122
- 1123
- 1124
- 1125
- 1126
- 1127
- 1128
- 1129
- 1130
- 1131
- 1132
- 1133
- 1134
- 1135
- 1136
- 1137
- 1138
- 1139
- 1140
- 1141
- 1142
- 1143
- 1144
- 1145
- 1146
- 1147
- 1148
- 1149
- 1150
- 1151
- 1152
- 1153
- 1154
- 1155
- 1156
- 1157
- 1158
- 1159
- 1160