LLM-RG4: Flexible and Factual Radiology Report Generation Across Diverse Input Contexts

Zhuhao Wang[1], Yihua Sun[1], Zihan Li[1], Xuan Yang[1], Fang Chen[2], Hongen Liao[1,2*]

[1]School of Biomedical Engineering, Tsinghua University, Beijing, China
[2]School of Biomedical Engineering, and Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
wangzhuh23@mails.tsinghua.edu.cn, liao@tsinghua.edu.cn

Abstract

Drafting radiology reports is a complex task requiring flexibility, where radiologists tail content to available information and particular clinical demands. However, most current radiology report generation (RRG) models are constrained to a fixed task paradigm, such as predicting the full "finding" section from a single image, inherently involving a mismatch between inputs and outputs. The trained models lack the flexibility for diverse inputs and could generate harmful, input-agnostic hallucinations. To bridge the gap between current RRG models and the clinical demands in practice, we first develop a data generation pipeline to create a new MIMIC-RG4 dataset, which considers four common radiology report drafting scenarios and has perfectly corresponded input and output. Secondly, we propose a novel large language model (LLM) based RRG framework, namely LLM-RG4, which utilizes LLM's flexible instruction-following capabilities and extensive general knowledge. We further develop an adaptive token fusion module that offers flexibility to handle diverse scenarios with different input combinations, while minimizing the additional computational burden associated with increased input volumes. Besides, we propose a token-level loss weighting strategy to direct the model's attention towards positive and uncertain descriptions. Experimental results demonstrate that LLM-RG4 achieves state-of-the-art performance in both clinical efficiency and natural language generation on the MIMIC-RG4 and MIMIC-CXR datasets. We quantitatively demonstrate that our model has minimal input-agnostic hallucinations, whereas current open-source models commonly suffer from this problem.

Code — https://github.com/zh-Wang-Med/LLM-RG4
Extended version — https://arxiv.org/abs/2412.12001

Introduction

The automatic generation of textual descriptions for radiographs has the potential to reduce clinicians' workload, enhance the efficiency of image interpretation, and support informed treatment decisions. Numerous works have concentrated on generating the comprehensive findings section of the report from a single radiology image (Li et al. 2018; Chen et al. 2020, 2021; Wang et al. 2022, 2023a; Yan et al. 2024). However, certain information in the report is uninferable within a single image, resulting in a mismatch between the input and the output, as illustrated in Figure 1(a). Concretely, Nguyen et al. (2023) classify the information in a report into several key components: positive mentions, negative mentions, prior comparisons, prior procedures, image views, doctor communications, and medical recommendations. Notably, elements such as comparisons, procedures, communication and views are uninferable within a single image. Current paradigm amplifies model hallucination, reduces model performance, and lowers clinical acceptance.

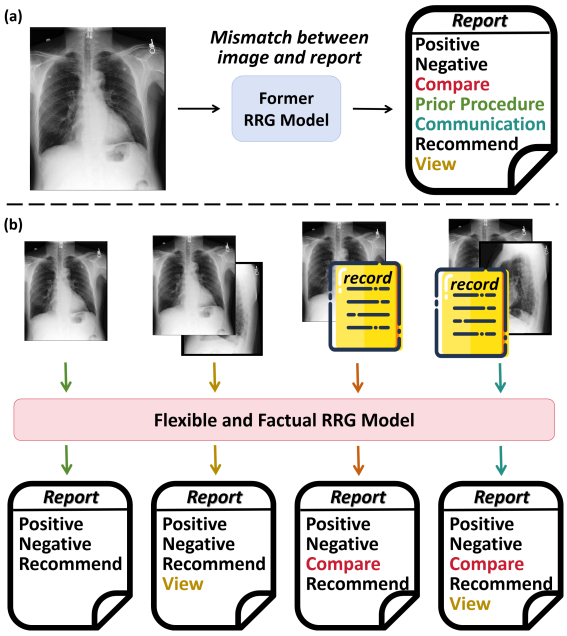


Figure 1: (a) Mismatch between image and report in typical RRG model. Comparisons, procedures, communication and views are uninferable. (b) A flexible and factual RRG paradigm, which emphasizes the flexibility of input and the alignment between input and output.

In response to this phenomenon, several studies have sought to clean and reconstruct the report content. For instance, Ramesh, Chi, and Rajpurkar (2022) introduced the GILBERT model, which utilizes token-by-token classifica-

*Corresponding author

tion to eliminate comparative descriptions, Thawakar et al. (2024) and Nguyen et al. (2023) leveraged the large language model to delete uninferable descriptions with a single image. However, such approaches drastically reduce the information included in the report, potentially undermining its effectiveness in fulfilling the intended function of radiology reports (Hartung et al. 2020). Furthermore, several studies have explored multi-view modeling (Yuan et al. 2019; Miura et al. 2021; Lee et al. 2023) and longitudinal historical information modeling (Dalla Serra et al. 2023; Sanjeev et al. 2024) to incorporate addition valid information. However, their performance deteriorates in the absence of additional information, and continues to produce hallucinations

In clinical practice, doctors adaptively draft reports based on available information and clinical requirements (Johnson et al. 2019). The radiologist compare the current findings with previous examinations when historical records are accessible. They integrate information from multiple views if provided, and concentrate on findings within a single frontal view if only one is available. Therefore, a more flexible and factual model for RRG should adapt to diverse input scenarios and produce reports inferred from the input within a unified framework, as illustrated in Figure 1(b). Meanwhile, it is crucial to identify definitive or potential lesions across various input scenarios to ensure timely intervention.

Inspired by clinical practices, this work introduces a flexible and factual framework consisting of a new data generation pipeline and a novel model architecture for RRG. The data generation pipeline produces the MIMIC-RG4 dataset from MIMIC-CXR, taking into account four common input scenarios that involve the integration of multi-view and longitudinal data. The pipeline comprises a BERT-based discriminator, namely DiscBERT, and a generator, Llama3-70B (AI@Meta 2024). It utilizes a cyclic generation approach to ensure that the reconstructed reports closely correspond to the input while minimizing information loss. Additionally, DiscBERT, as a byproduct of the pipeline, allows for quantitative input-agnostic information evaluation. In terms of the model architecture, we introduce LLM-RG4, which utilizes LLM's flexible instruction-following capabilities and extensive general knowledge. (Li et al. 2023a,b; Guo et al. 2023). To avoid increasing the computational burden with additional input types, we design an adaptive token fusion module that accommodates various inputs. The underlying intuition is that an efficient and high-fidelity information encoding for multimodal large language model (MLLM) can be achieved within specialized medical tasks. Furthermore, to improve clinical accuracy, we propose a token-level loss weighting strategy which enhances clinical efficacy directly at the loss layer, without depending on reinforcement learning (Miura et al. 2021) or classifier-assisted techniques (Jin et al. 2024). We validate our framework through experiments on MIMIC-CXR and MIMIC-RG4. Our contributions are summarized as follows.

- We present a novel paradigm MIMIC-RG4 for pragmatic RRG, introduce a new pipeline for data generation and a product DiscBERT to quantitatively evaluate input-agnostic hallucinations.
- We develop LLM-RG4, a LLM-based RRG model that incorporates an adaptive token fusion module to efficiently accommodate different inputs and a token-level loss weighting strategy to enhance diagnostic accuracy.
- We conduct extensive experiments demonstrating that LLM-RG4 achieves state-of-the-art performance in CE and NLG dimensions on both the MIMIC-CXR and MIMIC-RG4 datasets, while minimizing input-agnostic hallucinations, thus bridging the gap to clinical practice.

MIMIC-RG4 Paradigm

Problem Formulation

In contrast to the typical report generation paradigm, MIMIC-RG4 exhibits two notable differences. Firstly, it requires the model to be able to handle different input scenarios. Secondly, regardless of the input case, the model generates reports corresponding to the inputs.

We define four common clinical input scenarios: single view no longitudinal, multi-view no longitudinal, single view with longitudinal, and multi-view with longitudinal. Longitudinal refers to previous X-ray examinations and here we only include previous reports T_p. Single view refers to frontal image I_f. Multi-view refers to frontal and lateral images I_l. Meanwhile, the indication T_i or history T_h section is also important for report generation (Miura et al. 2021; Hyland et al. 2023), thus we incorporate them as inputs if available. Denote the model as L_g and the current report as T_c, the entire task is formalized as:

$$T_c = L_g(I_f, I_l, T_p, T_h, T_i) \tag{1}$$

where I_l, T_p, T_i, T_h allow for absence and the corresponding T_c will vary accordingly. With respect to the contents of reports, we follow the definition of Nguyen et al. (2023) and reconstruct the report to maximize the retention of effective information. Specifically, for different input scenarios, the communication and prior procedure are removed. Positive mentions, negative mentions, and medical recommendations are retained. View and prior comparison are retained or rewritten depending on the inputs.

Dataset Generation Pipeline

The overall pipeline is shown in Figure 2. Based on problem formulation, the prior comparison, prior procedure, view, and communication sections are of greater importance. We utilize Llama3-70B model as the generator to reconstruct reports. However, some challenging cases require multiple modifications to meet the requirements. Therefore, we adopt a cyclic process of judgment, rewriting, and re-evaluation until the report is satisfactory. In iterative modifications, the previous process is also used to form a dynamic prompt that fully leverages the model's chain-of-thought (COT) capability. This is a highly time-consuming process. To expedite it, we train a BERT-based model, DiscBERT, as the discriminator to perform judgment tasks. DiscBERT is trained with the Llama3-70B's judgment results, and exhibits judgement capabilities comparable to Llama3-70B. To preserve the diagnostic information of the report, we employ CheXbert to compute the disease labels for the impression section before

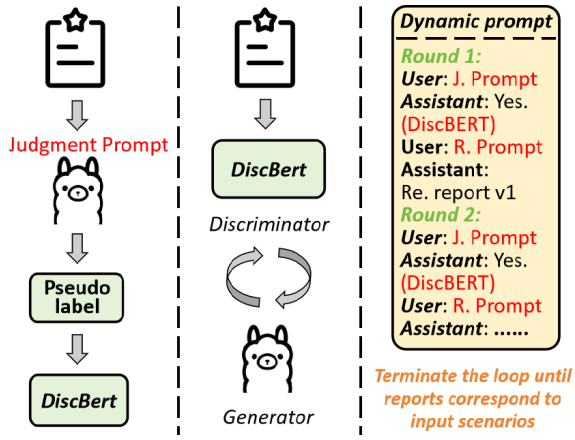


Figure 2: The pipeline employs an iterative approach that integrates a BERT-based discriminator and a LLM-based generator, ensuring minimal input-agnostic information and effective information loss. J.: Judgment, R.: Rewrite.

and after processing. If there is any change in labels, the corresponding example is discarded. We set the maximum iterations to 3 rounds.

Our integrated approach enables us to leverage the COT of LLM while accelerating the pipeline by reducing the reliance on LLM. Moreover, DiscBERT allows for convenient dataset analysis that distinguishes information categories within the generated reports, offering a tool for evaluating input-agnostic hallucinations. Doctors manually label 200 reports to evaluate the discriminatory performance of DiscBERT and assess the effectiveness of the pipeline. The details about DiscBERT, instructions and evaluation are presented in the supplementary material.

Dataset Statistics

We utilize the MIMIC-CXR dataset (Johnson et al. 2019), which is the only publicly available dataset that encompasses both multi-view and longitudinal information, to generate the MIMIC-RG4 dataset. We evaluate the proportion of reports containing input-agnostic information in the MIMIC-CXR and MIMIC-RG4 datasets under the single view no longitudinal setting, as assessed by DiscBERT. As shown in Table 1, a few cases in MIMIC-CXR meet the criterion of single view no longitudinal. However even in these cases, the reports still include a number of descriptions of prior comparisons and procedures. This indicates that obtaining the dataset for single view no longitudinal data directly from MIMIC-CXR is inappropriate. In contrast, MIMIC-RG4 minimizes the presence of input-agnostic information and features a larger dataset scale.

Method

The overall architecture, depicted in Figure 3, consists of three primary components: modality encoder, adaptive token fusion module (ATF), and token-level loss weighting strategy (TLW). The modality encoder utilizes pretrained encoders to extract features from a range of visual and textual inputs. The adaptive token fusion module subsequently compresses and integrates these features into a fixed-length fusion token. This token is then supplied to the LLM alongside instructions and indications/histories for decoding. The token-level loss weighting strategy identifies critical diagnostic tokens and assigns them higher loss weights, thereby directing the model to emphasize positive or ambiguous descriptions regardless of the input conditions.

Dataset	split	PC	PP	View	Comm
MIMIC CXR	Tr/16.9K	39.90	15.99	1.17	4.56
	Val/0.1K	38.36	11.28	1.50	3.01
	Ts/0.1K	65.63	25.78	3.13	10.16
MIMIC RG4	Tr/172.6K	0.30	0.30	0.12	0.00
	Val/1.4K	0.07	0.07	0.14	0.00
	Ts/2.4K	0.42	0.42	0.04	0.04

Table 1: Percentage (%) of reports with single image no longitudinal setting, that encompass various categories of information. PC: Prior Comparison; PP: Prior Procedure; Comm: Communication; Tr: train; Ts: test.

Modality Encoder

Under different input scenarios, the model can access frontal image I_f, lateral images I_l, and previous examination reports T_p. We utilize frozen image encoder E_v and text encoder E_t to obtain corresponding features v_f, v_l, v_t.

$$v_f = E_v(I_f) \tag{2}$$
$$v_l = E_v(I_l) \tag{3}$$
$$v_t = E_t(T_p) \tag{4}$$

where $v_f, v_l \in \mathbb{R}^{N_I \times D'}$, $v_t \in \mathbb{R}^{N_T \times D'}$, D' is the dimension of feature obtained from modality encoder, N_I is the number of visual tokens, N_T is the number of text tokens. N_I and N_T are generally distinct, with N_I frequently being larger when higher resolution images are employed.

Adaptive Token Fusion Module

Our objective is to maintain a consistent number of feature tokens across different inputs. For instance, with I_f, I_l and T_p present, we hope to produce a fused feature with dimensions equivalent to those of I_f alone. We first employ perceiver p_f, p_l, p_t (Jaegle et al. 2021) and linear layers to further extract and compress modality feature to a consistent dimension h_f, h_l, $h_t \in \mathbb{R}^{N \times D}$ as follows:

$$h_f = Linear(p_f(v_f, V')) \tag{5}$$
$$h_l = Linear(p_l(v_l, p_f(v_f, V'))) \tag{6}$$
$$h_t = Linear(p_t(v_t, p_f(v_f, V'))) \tag{7}$$

where N is the compressed number of feature tokens and considerably smaller than N_I, D is the feature dimension in LLM. Query tokens in p_f are learnable variables $V' \in \mathbb{R}^{N \times D'}$, whereas the query tokens in p_l, p_t are $p_f(v_f, V')$ derived from frontal image. This approach leverages the frontal image as the primary feature for modality integration, underscoring its critical role across different scenarios.

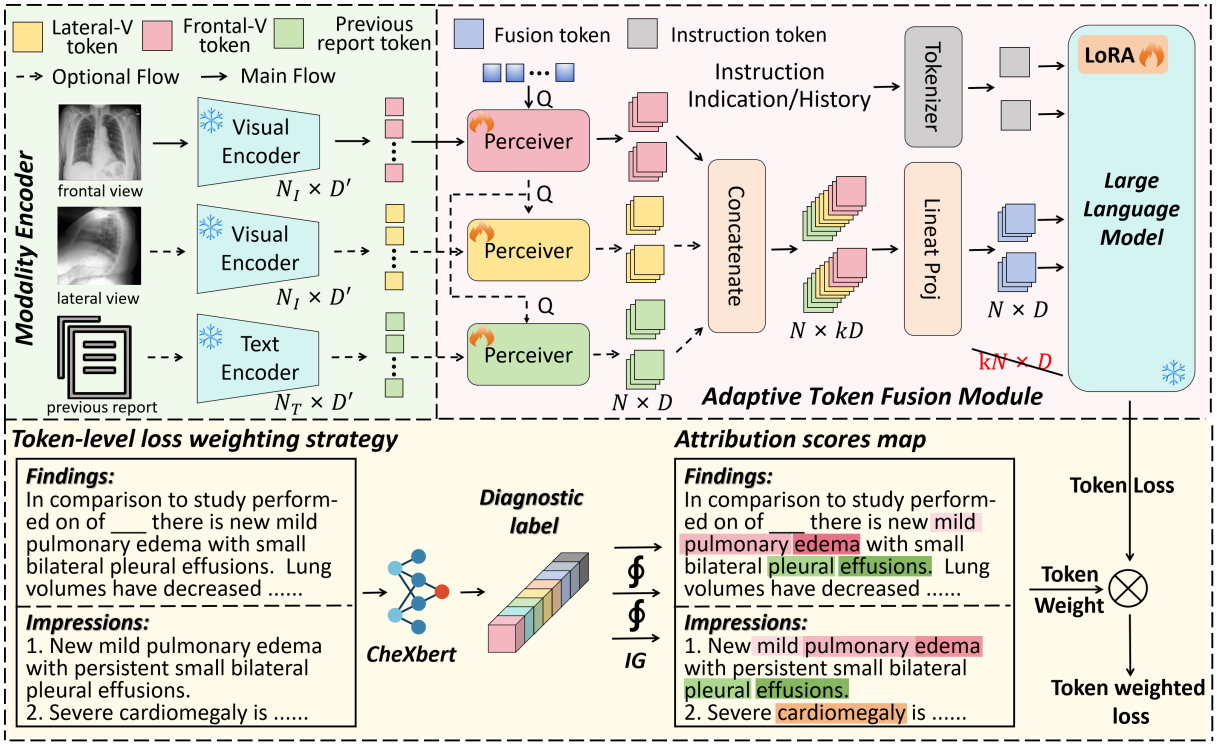


Figure 3: The LLM-RG4 architecture consists of a modality encoder, an adaptive token fusion module, and a token-level loss weighting strategy. The modality encoder extracts features from various modalities. The adaptive token fusion module combines different feature tokens into a fixed length, minimizing computational burden. The token-level loss weighting strategy identifies key diagnoses and adjusts token loss weights, enhancing the model's clinical efficacy across diverse input scenarios.

Typical visual instruction tuning can be viewed as stacking modality features along the token dimension, and escalates computational complexity with increasing input volumes. Given that instruction tuning enables a LLM to understand unseen visual tokens, it is also feasible to train it to comprehend mixed visual-linguistic tokens similarly. Therefore, we consider compressing information along the feature dimensions of each token. Specifically, we first concatenate h_f, h_l, $h_t \in \mathbb{R}^{N\times D}$ along the feature dimensions and get $h_o \in \mathbb{R}^{N\times 3D}$, formulated as:

$$h_0 = concat(h_f, h_l, h_t, dim = 1) \tag{8}$$

If h_l or h_t are not available, we replace it with zeros. To avoid confusion between different modalities, the concatenation order is fixed. A linear projection layer is utilized to maintain the feature dimensions, resulting in the final features h'_o, which are then input into the LLM. This approach guarantees that the number of feature tokens remains constant while efficiently encoding modality information across varying input scenarios.

Token-Level Loss Weighting Strategy

We denote the logit computation function as $f(\theta)$, where θ is the parameter of LLM, denote the current report with a length L as $T = [t^1, t^2, t^3, \ldots, t^L]$, and denote the instruction as P. The predicted logit o is depicted as:

$$o^j = f_\theta(P, h_o, T^{<j}) \tag{9}$$

where $T^{<j}$ represents previous tokens before position j in the current report. The loss at token t^j is depicted as:

$$L_{\text{MLE}}^{(t^j)} = -c_j \log_{\text{softmax}}(o^j) \tag{10}$$

where $c_j = 1$, for $j = 1, 2, \ldots, L$, if all tokens are treated equally. To identify key diagnostic tokens in each report and subsequently adjust the coefficient c_j for each token loss, we utilize CheXbert (Smit et al. 2020) and Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017). CheXbert conducts multi-label categorization to identify diagnostic labels $Y = [y_1, y_2, y_3, \ldots, y_{14}]$ for 14 distinct diseases in reports. $y_i \in \{-1, 0, 1, 2\}$, -1 indicates uncertainty, 0 indicates negative, 1 indicates positive, 2 indicates not mentioned.

Given that the 14th category in CheXbert is "NO FINDING", we focus solely on the attribution maps of the first 13 categories. The detailed algorithm is depicted in Algorithm 1. We use CheXbert to identify uncertain or positive diagnostic labels in the report and apply Integrated Gradients to generate attribution maps for each label. We then maximize these maps and smooth them with a Gaussian kernel to mitigate token anomaly. If a token's attribution score exceeds a threshold, the weight of each token in the whole sentence is increased to λ. Otherwise, it remains unchanged. Through this approach, we emphasize entire sentences with positive or uncertain diagnoses, minimizing the impact of minor discrepancies between tokens and attribution scores resulting from different tokenizers.

Model	CE Metrics			Clean NLG			Original NLG			*hall.*
	P	R	F1	B@1	B@4	R-L	B@1	B@4	R-L	
R2Gen(Chen et al. 2020)	0.456	0.306	0.366	0.363	0.090	0.269	0.356	0.097	0.267	0.779
R2GenCMN(Chen et al. 2021)	0.486	0.400	0.439	0.385	0.102	0.278	0.349	0.094	0.270	0.695
CVT2Dis.(Nicolson, Dowling, and Koopman 2023a)	0.498	0.414	0.452	0.374	0.103	0.272	0.390	0.123	0.282	0.875
KiUT†(Huang, Zhang, and Zhang 2023)	0.371	0.318	0.321	-	-	-	0.391	0.113	0.285	-
RGRG†(Tanida et al. 2023)	0.461	0.475	0.447	-	-	-	0.373	0.126	0.264	-
EKAGen†(Bu et al. 2024)	0.517	0.483	0.499	-	-	-	0.419	0.117	0.287	-
Promptmrg(Jin et al. 2024)	**0.618**	0.491	0.548	0.326	0.080	0.261	0.381	0.096	0.258	0.896
R2GenGPT(7B)(Wang et al. 2023b)	0.506	0.414	0.456	0.401	0.118	0.277	0.396	0.113	0.273	0.917
CheXagent(7B)(Chen et al. 2024)	0.506	0.306	0.381	0.265	0.058	0.239	0.189	0.040	0.208	0.549
MAIRA-1(7B)†(Hyland et al. 2023)	-	-	0.553	-	-	-	0.392	0.142	0.289	-
Med-PaLM(562B)†(Tu et al. 2024)	-	-	0.516	-	-	-	0.317	0.115	0.275	-
R2-LLM(14.2B)†(Liu et al. 2024a)	0.465	0.482	0.473	-	-	-	0.402	0.128	0.291	-
InVERGe(7B)†(Deria et al. 2024)	-	-	-	-	-	-	**0.425**	0.100	0.309	-
Ours	0.583	**0.593**	**0.588**	**0.498**	**0.203**	**0.387**	0.377	**0.144**	**0.318**	**0.015**

Table 2: Comparison with SOTA methods for the setting of ***sn***. † indicates the results are quoted from the published literature. **Clean NLG** refers to using the cleaned reports from MIMIC-RG4 as ground truth, while **Original NLG** denotes using the original reports from MIMIC-CXR as ground truth. ***hall.*** means the percentage of reports containing input-agnostic information. The best results are in **bold**.

Experimental Setup

Datasets and Metrics

We train LLM-RG4 on MIMIC-RG4 consisting of four input scenarios: single view no longitudinal, multi-view no longitudinal, single view with longitudinal, and multi-view with longitudinal (denoted as ***sn***, ***sw***, ***mn***, ***mw*** below). We include the indication/history section as input when available.

Model performance is assessed across three dimensions: natural language generation (NLG), clinical efficacy (CE), and hallucination (***hall.***). For NLG, we use BLEU (B@n) (Papineni et al. 2002), METEOR (MTR) (Banerjee and Lavie 2005) and ROUGE-L (R-L) (Lin 2004). For CE, we adopt CheXbert to extract category labels and calculate micro-averaged precision (P), recall (R), and F1-score (F1), following established settings (Jin et al. 2024). For hallucinations, we emphasize input-agnostic hallucinations (***hall.***) and employ DiscBERT to measure the proportion of generated reports containing input-agnostic information. A lower value of ***hall.*** reflects a diminished occurrence of input-agnostic hallucination. Additionally, we use the Wilcoxon Signed-Rank Test (Woolson 2005) to assess performance improvements over baselines. All reports are kept untruncated during testing.

Algorithm 1: Detailed Procedure of Token Weight C

Input: Report $T = [t^1, t^2, t^3, \ldots, t^L]$, CheXbert f_c
Output: $C = [c_1, c_2, c_3, \ldots, c_L]$

Initialize $c_i = -1$
Get $Y = [y_1, y_2, y_3, \ldots, y_{13}] = f_c(T)$
For y_j **in** G:
  if $y_j = -1$ or 1: **then**
    $c_i' = \text{IG}_i(x)$
    $c_i = \max(c_i, c_i')$
Define $g_k = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{k^2}{2\sigma^2}}$
Split C into M sentences $C^s = [c^1, c^2, c^3, \ldots, c^M]$, c^n is the nth sentence's weights with length Ln, $c^n = [c_1^n, c_2^n, c_3^n, \ldots, c_{L_n}^n]$.
if $c_i^n > threshold$ **then**
  $c^n = \lambda$ and $\lambda > 1$
else
  $c^n = 1$
end if
return C

Baselines

Comparative experiments are performed on the traditional ***sn*** task and the multi-task MIMIC-RG4. For traditional ***sn***, we adopt frontal images and focus solely on the findings section in MIMIC-CXR following (Chen et al. 2020). For MIMIC-RG4, both finding and impression are evaluated. We retrain the encoder-decoder model CXRMate (Nicolson, Dowling, and Koopman 2023b), which handles four input scenarios and is trained with reinforcement learning. Additionally, we compare LLM-RG4 with RadFM (Wu et al. 2023), a 14B radiology multimodal large language model capable of processing interleaved text and image inputs.

Implementation Details

We adopt RAD-DINO (Pérez-García et al. 2024) as the image encoder and BiomedVLP-CXRBERT (Boecking et al. 2022) as the text encoder, with Vicuna 7B v1.5 (Chiang et al. 2023) as the text decoder. The number of learnable variable tokens in the perceiver is set to 128, threshold is set to 0.4 and λ is set to 1.75. Following LLAVA (Liu et al. 2024b), we employ a two-stage training strategy. Initially, we only train the ATF with ***sn*** data to achieve modality alignment. Subsequently, we conduct instruction tuning on the MIMIC-RG4 dataset, training the ATF, and applying LoRA (Hu et al. 2021) for fine-tuning Vicuna.

Dataset	Model	CE Metrics			NLG Metrics						*hall.*
		P	R	F1	B@1	B@2	B@3	B@4	R-L	MTR	
RG4 ***sn***	cxrmate*	0.572‡	0.560‡	0.566‡	0.421‡	0.271‡	0.179‡	0.122‡	0.311‡	0.174‡	**0.010**
	RadFM	0.413‡	0.303‡	0.350‡	0.188‡	0.090‡	0.048‡	0.028‡	0.190‡	0.094‡	0.737‡
	Ours	**0.588**	**0.632**	**0.609**	**0.479**	**0.343**	**0.255**	**0.196**	**0.384**	**0.209**	0.014
RG4 ***sw***	cxrmate*	0.573‡	0.549‡	0.561‡	0.361‡	0.220‡	0.139‡	0.093‡	0.284‡	0.153‡	**0.009**
	RadFM	0.508‡	0.365‡	0.425‡	0.211‡	0.103‡	0.056‡	0.033‡	0.183‡	0.105‡	0.092‡
	Ours	**0.599**	**0.622**	**0.610**	**0.455**	**0.321**	**0.239**	**0.186**	**0.382**	**0.199**	0.021
RG4 ***mn***	cxrmate*	**0.544**	0.522‡	0.533‡	0.437‡	0.289‡	0.199‡	0.141‡	0.332‡	0.179‡	0.009
	RadFM	0.323‡	0.187‡	0.237‡	0.246‡	0.113‡	0.060‡	0.034‡	0.194‡	0.104‡	0.140‡
	Ours	0.541	**0.578**	**0.559**	**0.491**	**0.359**	**0.274**	**0.216**	**0.405**	**0.214**	**0.008**
RG4 ***mw***	cxrmate*	0.548‡	0.499‡	0.523‡	0.379‡	0.241‡	0.158‡	0.110‡	0.305‡	0.159‡	0.007
	RadFM	0.456‡	0.297‡	0.360‡	0.191‡	0.095‡	0.054‡	0.034‡	0.178‡	0.095‡	0.052‡
	Ours	**0.560**	**0.565**	**0.563**	**0.461**	**0.331**	**0.250**	**0.197**	**0.401**	**0.204**	**0.002**

Table 3: Comparison with SOTA methods supporting MIMIC-RG4 across four settings. $*$ indicates the model is retrained on MIMIC-RG4. ‡ denotes statistical significance in paired comparisons with LLM-RG4 based on the Wilcoxon signed-rank test. ***hall.*** means the percentage of reports containing input-agnostic information. The best results are in **bold**.

Results and Analysis

Overall Results

Table 1 presents the experimental results on the conventional ***sn*** task. Benefiting from the token-level loss weighting strategy, LLM-RG4 shows a 4.0% and a 3.5% absolute improvement respectively in F1 score over best classifier-assisted model Promptmrg and MAIRA-1 of the same 7B size using 1369 visual tokens. In clean NLG, where ground truth reports get 2.3% ***hall.*** score, LLM-RG4 has a 24.2% relative improvement in BLEU-4 and a 39.2% relative improvement in ROUGE-L, attributed to training on a cleaner dataset. Regarding hallucination, we find most open-source models suffer from hallucination problems. Even a multi-tasking model CheXagent also gets a 54.9% ***hall.*** score. However, LLM-RG4 achieves only a 1.5% ***hall.*** score, indicating that it exhibits minimal input-agnostic hallucinations. Finally, we compute the original-NLG metrics with MIMIC-CXR reports for comparsion with powerful closed-source models. Surprisingly, although the ground truth reports attain a 88.5% ***hall.*** score, our model remains competitive with SOTA models. Both BLEU-4 score and ROUGE-L score of LLM-RG4 are comparable to or slightly exceed the SOTA specialised large multimodal model MAIRA-1. The minimal ***hall.*** suggests a substantial number of valid descriptions matching the ground truth. This result likely arises from our model being trained on a mixture of four kinds of clean reports, which enhance its learning of valid descriptions.

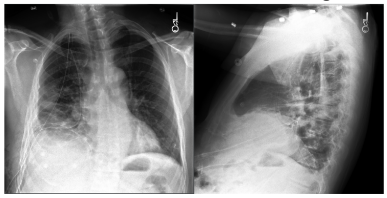
Under the MIMIC-RG4 setting, LLM-RG4 outperforms all existing models supporting MIMIC-RG4. Although cxrmate utilized reinforcement learning to enhance clinical accuracy, we still achieve an average absolute F1 improvement of 3.8% across four tasks. LLM-RG4's superior performance on NLG metrics further demonstrates that the MLLM architecture is well-suited for multi-tasks featuring flexible language generaion. While RadFM can support interleaved text and image inputs, it performs unsatisfactorily and exhibits a significant input-agnostic hallucination issue on the ***sn*** setting likely due to overfitting.

Ablation Analysis

We conduct ablation experiments of alternative model designs, as shown in Table 4. The adaptive token fusion module helps reduce feature tokens by approximately 60% when both multi-view and longitudinal information are available, offering similar or slightly better performance than original interleaved inputs. This supports our hypothesis that, efficient and high-fidelity encoding for MLLM can be achieved within specialized medical tasks. TLW benefits CE results on both ATF architecture and interleaved inputs architecture, suggesting that assigning higher loss weights to positive and uncertain descriptions during training can effectively enhance the model's focus on underlying lesions. This phenomenon is evident in both the first and second stages. Further analyses about the ATF module and TLW module are provided in the supplementary materials.

stage	ATF	TLW	N.	F1 Score		NLG	
				F-14	F-5	B@1	B@4
1	-	✘	128	0.519	0.563	0.388	0.126
1	-	✔	128	**0.580**	**0.619**	**0.400**	**0.129**
2	✘	✘	502	0.551	0.582	0.468	0.201
2	✘	✔	502	0.577	0.603	0.469	0.197
2	✔	✘	**204**	0.556	0.580	0.467	**0.205**
2	✔	✔	**204**	**0.585**	**0.610**	**0.472**	0.199

Table 4: Ablation study of each module on MIMIC-RG4. Stage2 results are the average scores across four settings. N. means the max input token numbers.



last report: *finding : the patient has undergone vats decortication ....*

Gold Standard: *finding : in comparison with the study of there is some decrease in the opacification at the right base . chest tubes remain in place and there is no evidence of pneumothorax . some residual atelectasis and effusion are noted . the possibility of supervening pneumonia at the right base could not be excluded . the left lung is essentially clear with mild atelectatic changes at the base . subcutaneous emphysema persists along the right lateral upper abdominal wall .*

sn (+TLW): *findings : two chest tubes are in place and there is no evidence of pneumothorax . opacification at the right base is prominent consistent with residual pleural fluid and atelectatic changes . subcutaneous emphysema is present along the right lateral chest wall . left lung is essentially clear except for mild atelectatic changes at the base .*

sw (+TLW): *finding : in comparison with the study of there is little overall change in the appearance of the three right chest tubes with no evidence of pneumothorax . subcutaneous emphysema along the right lateral chest wall appears to be slightly less than on the previous study . opacification at the right base again could reflect post-surgical changes with some atelectatic changes and small effusion . left lung is essentially clear .*

mn (+TLW): *finding : there is a small apical pneumothorax on the right with chest tube in place . opacification at the right base is consistent with residual atelectasis and effusion . subcutaneous emphysema is present along the right lateral chest wall extending into the neck . mild atelectatic changes are seen at the left base .*

mw (+TLW): *finding : in comparison with the study of there is little overall change in the appearance of the three right chest tubes with no evidence of pneumothorax . subcutaneous emphysema is again seen along the right lateral chest wall . opacification at the right base again could reflect atelectasis and effusion though in the appropriate clinical setting superimposed pneumonia would have to be considered . left lung remains clear .*

mw (-TLW): *finding : in comparison with the study of there is little change in the appearance of the right hemithorax with three chest tubes in place and no definite pneumothorax . opacification at the right base again is consistent with some combination of atelectasis and re-expansion edema . in the appropriate clinical setting superimposed pneumonia would have to be considered .*

Figure 4: An illustration of a challenging case featuring five positive or uncertain diagnoses across four different settings, where the gold standard is also presented for reference. Diagnosis shared by the gold standard and model outputs are highlighted in the same color. LLM-RG4 identifies nearly all diagnoses, whereas the absence of TLW leads to the missing of two diagnoses.

Influence of Mixed Training

To further investigate whether mixed training across the four settings yields improvement, we train each setting individually for the same number of epochs, shown in Table 5. We find that mixed training lead to varying degrees of improvement across different scenarios, particularly in challenging settings (***mn***, ***mw***). We hypothesize that such mixed training can be regarded as a form of data augmentation, effectively increasing the diversity of the training data. It explicitly enables the model to learn the varying demands across different settings during training.

Setting	T.S.	CE Metrics			NLG Metrics		
		P	R	F1	B@1	B@4	R-L
sn	**S**	**0.589**	0.609	0.599	0.456	0.186	0.384
	M	0.588	**0.632**	**0.609**	**0.479**	**0.196**	**0.384**
sw	**S**	0.596	0.588	0.592	0.427	0.173	0.376
	M	**0.599**	**0.622**	**0.610**	**0.455**	**0.186**	**0.382**
mn	**S**	0.526	0.531	0.528	0.464	0.202	0.384
	M	**0.541**	**0.578**	**0.559**	**0.491**	**0.216**	**0.405**
mw	**S**	0.519	0.544	0.531	0.446	0.179	0.384
	M	**0.560**	**0.565**	**0.563**	**0.461**	**0.197**	**0.401**

Table 5: Influence of mixed training across four settings. T.S. represents the training strategy, **M** represents mixed training, **S** represents training on the specific setting.

Case study

We present a qualitative example to illustrate LLM-RG4's flexibility of diverse inputs and investigate the impact of the TLW module on the model's capabilities, as shown in Figure 3. We select a challenging case involving a patient with confirmed or suspected diagnoses of five different diseases. In the context of multi-view X-ray inputs and longitudinal data, the reports generated by LLM-RG4 accurately cover all five diseases, demonstrating its clinical diagnostic accuracy. In the absence of the token-level loss weighting strategy (TLW), the generated reports lack two diagnostic information. Additionally, the comparative description appears only when longitudinal data is included (***sw***, ***mw***), indicating consistency between model inputs and outputs. For the limitations of LLM-RG4, while it provides four types of diagnostic descriptions under scenarios such as ***sn***, ***sw***, and ***mn***, it does not mention the need to consider pneumonia. In the ***mn*** scenario, it suspects the presence of a small pneumothorax. Future work should further constrain LLM-RG4 to ensure the consistency across different input scenarios. Such enhancement would be beneficial for clinical practice.

Conclusion

In this work, we introduce MIMIC-RG4, a novel paradigm for radiology report generation that adapts to varying input scenarios, aligning more closely with clinical report writing practices. We further propose ATF and TLW to enhance the flexibility and accuracy of large language models in handling diverse inputs, with a consistent emphasis on identifying pathological findings across different settings. Experiments conducted on two datasets illustrate the effectiveness of our method, highlighting that MLLM can achieve more compact information encoding for specific medical tasks. Additionally, the emphasis on key semantic tokens at the loss layer is crucial for enhancing clinical efficacy. We hope these efforts will provide new insights into radiology report generation and the application of large language models in biomedical domains.

Acknowledgments

The authors acknowledge supports from National Key Research and Development Program of China (2022YFC2405200), National Natural Science Foundation of China (U22A2051, 82027807), Tsinghua-Foshan Innovation Special Fund (2021THFS0104), and Institute for Intelligent Healthcare, Tsinghua University (2022ZLB001).

References

AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3. Accessed: 2024-04-18.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, 1–21. Springer.

Bu, S.; Li, T.; Yang, Y.; and Dai, Z. 2024. Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14204.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 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 (Volume 1: Long Papers)*, 5904–5914.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1439–1449.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Dalla Serra, F.; Wang, C.; Deligianni, F.; Dalton, J.; and O'Neil, A. 2023. Controllable Chest X-Ray Report Generation from Longitudinal Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4891–4904.

Deria, A.; Kumar, K.; Chakraborty, S.; Mahapatra, D.; and Roy, S. 2024. InVERGe: Intelligent Visual Encoder for Bridging Modalities in Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2028–2038.

Guo, Z.; Zhang, R.; Zhu, X.; Tang, Y.; Ma, X.; Han, J.; Chen, K.; Gao, P.; Li, X.; Li, H.; et al. 2023. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*.

Hartung, M. P.; Bickle, I. C.; Gaillard, F.; and Kanne, J. P. 2020. How to create a great radiology report. *Radiographics*, 40(6): 1658–1670.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19809–19818.

Hyland, S. L.; Bannur, S.; Bouzid, K.; Castro, D. C.; Ranjit, M.; Schwaighofer, A.; Pérez-García, F.; Salvatelli, V.; Srivastav, S.; Thieme, A.; et al. 2023. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*.

Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; and Carreira, J. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, 4651–4664. PMLR.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2607–2615.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Lee, H.; Kim, W.; Kim, J.-H.; Kim, T.; Kim, J.; Sunwoo, L.; and Choi, E. 2023. Unified chest x-ray and radiology report generation model with multi-view chest x-rays. *arXiv preprint arXiv:2302.12172*, 3(7): 8.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Yang, J.; Li, C.; and Liu, Z. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 19730–19742.

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.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024a. Bootstrapping Large Language Models for Radiology Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18635–18643.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Miura, Y.; Zhang, Y.; Tsai, E.; Langlotz, C.; and Jurafsky, D. 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*, 5288–5304.

Nguyen, D.; Chen, C.; He, H.; and Tan, C. 2023. Pragmatic radiology report generation. In *Machine Learning for Health (ML4H)*, 385–402. PMLR.

Nicolson, A.; Dowling, J.; and Koopman, B. 2023a. Improving chest X-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144: 102633.

Nicolson, A.; Dowling, J.; and Koopman, B. 2023b. Longitudinal Data and a Semantic Similarity Reward for Chest X-Ray Report Generation. *arXiv preprint arXiv:2307.09758*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pérez-García, F.; Sharma, H.; Bond-Taylor, S.; Bouzid, K.; Salvatelli, V.; Ilse, M.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Lungren, M. P.; et al. 2024. RAD-DINO: Exploring Scalable Medical Image Encoders Beyond Text Supervision. *arXiv preprint arXiv:2401.10815*.

Ramesh, V.; Chi, N. A.; and Rajpurkar, P. 2022. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, 456–473. PMLR.

Sanjeev, S.; Maani, F. A.; Abzhanov, A.; Papineni, V. R.; Almakky, I.; Papież, B. W.; and Yaqub, M. 2024. TiBiX: Leveraging Temporal Information for Bidirectional X-ray and Report Generation. *arXiv preprint arXiv:2403.13343*.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. 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)*, 1500–1519.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7433–7442.

Thawakar, O. C.; Shaker, A. M.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. 2024. XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 440–448.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *NEJM AI*, 1(3): AIoa2300138.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023a. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11558–11567.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023b. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3): 100033.

Wang, Z.; Tang, M.; Wang, L.; Li, X.; and Zhou, L. 2022. A medical semantic-assisted transformer for radiographic report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 655–664. Springer.

Woolson, R. F. 2005. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, 8.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Yan, S.; Cheung, W. K.; Tsang, I. W.; Chiu, K.; Tong, T. M.; Cheung, K. C.; and See, S. 2024. AHIVE: Anatomy-aware Hierarchical Vision Encoding for Interactive Radiology Report Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14324–14333.

Yuan, J.; Liao, H.; Luo, R.; and Luo, J. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, 721–729. Springer.