HC-LLM: Historical-Constrained Large Language Models for Radiology Report Generation

Tengfei Liu[1], Jiapu Wang[1], Yongli Hu[1]*, Mingjie Li[2], Junfei Yi[3], Xiaojun Chang[4], Junbin Gao[5], Baocai Yin[1]

[1]School of Information Science and Technology, Beijing University of Technology, Beijing, China
[2]Stanford University, Palo Alto CA 94305 USA
[3]School of Electrical and Information Engineering, Hunan University, Hunan, China
[4]School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[5]University of Sydney Business School, The University of Sydney, Camperdown, NSW 2006, Australia
{tfliu, jpwang}@emails.bjut.edu.cn, {huyongli,ybc}@bjut.edu.cn, lmj695@stanford.edu, yijunfei@hnu.edu.cn, cxj273@gmail.com, junbin.gao@sydney.edu.au

Abstract

Radiology report generation (RRG) models typically focus on individual exams, often overlooking the integration of historical visual or textual data, which is crucial for patient follow-ups. Traditional methods usually struggle with long sequence dependencies when incorporating historical information, but large language models (LLMs) excel at in-context learning, making them well-suited for analyzing longitudinal medical data. In light of this, we propose a novel Historical-Constrained Large Language Models (HC-LLM) framework for RRG, empowering LLMs with longitudinal report generation capabilities by constraining the consistency and differences between longitudinal images and their corresponding reports. Specifically, our approach extracts both time-shared and time-specific features from longitudinal chest X-rays and diagnostic reports to capture disease progression. Then, we ensure consistent representation by applying intra-modality similarity constraints and aligning various features across modalities with multimodal contrastive and structural constraints. These combined constraints effectively guide the LLMs in generating diagnostic reports that accurately reflect the progression of the disease, achieving state-of-the-art results on the Longitudinal-MIMIC dataset. Notably, our approach performs well even without historical data during testing and can be easily adapted to other multimodal large models, enhancing its versatility.

Introduction

Radiology report generation (RRG) is a crucial research area in medical AI, with numerous studies focused on reducing the heavy workload of radiologists. In clinical practice, an essential function of these reports is to document pathological changes in patients, aiding doctors in recalling and diagnosing disease progression. As a result, ground truth reports often include descriptions of historical information. However, most existing models (Chen et al. 2020; Liu et al. 2021a; Chen et al. 2021; Li et al. 2023a,c; Tanida et al. 2023; Liu, Tian, and Song 2023; Huang, Zhang, and Zhang 2023; Wang et al. 2023b,a; Liu et al. 2024b; Jin et al. 2024; Liu et al. 2024a; Shen et al. 2024; Li et al. 2024) rely on a single image as input, preventing them from accurately generating descriptions of prior references and thereby impacting their performance. This limitation is evident as current models struggle to achieve high scores on natural language generation (NLG) metrics. Therefore, as illustrated in Figure 1, this paper focuses on a more practical research problem: how to generate radiology reports from longitudinal data.

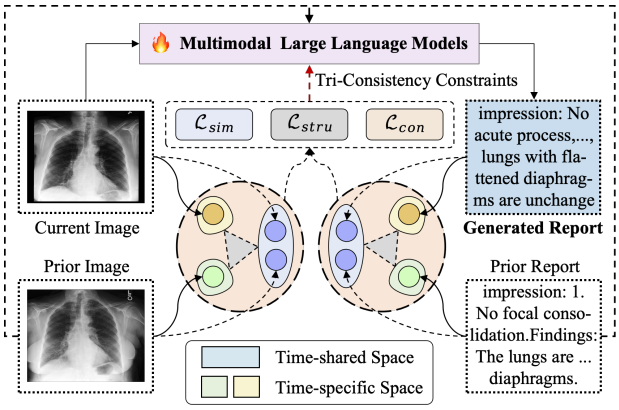

Figure 1: Illustration of the longitudinal report generation process. Unlike traditional methods that rely solely on the current chest X-ray, our approach emphasizes the effective utilization of historical diagnostic information to significantly enhance the accuracy of LLMs in RRG.

Zhu et al. (2023b) have made preliminary attempts in this direction by utilizing a memory mechanism to incorporate historical information for enhanced chest X-ray report generation. However, their approach still relies on traditional cross-attention mechanisms and requires the presence of historical data during testing, limiting its practicality. Recently, large language models (LLMs) have been successfully applied to traditional radiology report generation tasks (Jin et al. 2024; Liu et al. 2024a,b), and their inherent in-context learning abilities make them well-suited for analyzing longitudinal medical data. However, despite this potential, directly inputting longitudinal medical data into LLMs often

*Corresponding author

struggles to produce reports that accurately capture the progression of diseases over time. Therefore, our work focuses on a key challenge: How can historical diagnostic information be effectively utilized to enhance the radiology report generation capabilities of LLMs?

To address this challenge, we propose a Historical-Constrained Large Language Models (HC-LLM) framework for RRG. Specifically, considering changes in disease progression, which may involve the disappearance, stability, and emergence of conditions, we first extract time-shared (stability) and time-specific (disappearance, emergence) features from chest X-rays and diagnostic reports at two-time points. We then employ two types of constraints: intra-modality and inter-modality constraints, which indirectly guide LLM generation by aligning the shared and specific features of the generated and historical reports with those of the corresponding longitudinal images. This alignment ensures that the generated reports accurately capture disease progression, thereby enhancing overall effectiveness.

For intra-modality constraints, we apply similarity constraints to ensure the consistency of time-shared features within each modality, preserving the integrity of chest X-ray and diagnostic report characteristics over time. For inter-modality constraints, we implement multimodal contrastive constraint and multimodal structural constraint, respectively. The former is responsible for aligning time-shared and time-specific features of corresponding chest X-rays and reports while distancing non-matching pairs. This constraint also indirectly enhances the separation of features within the same modality, making the representation of time-shared and time-specific features more precise. The latter further regulates the spatial distribution of features by forming triangular structures, ensuring that the geometric relationships (e.g., distances and angles) within the triangles of image features correspond to those within text features. The combined effect of these constraints ensures that the generated reports more accurately reflect the progression of diseases, thereby enhancing their accuracy. Experimental results on the Longitudinal-MIMIC dataset demonstrate that our method achieves state-of-the-art performance on most NLG metrics, validating its effectiveness. Additionally, our method achieves superior results compared to other approaches without using historical information during testing and can be adapted to various multimodal large model frameworks, demonstrating strong applicability. We summarize contributions as follows:

- We propose an innovative HC-LLM framework that leverages historical diagnostic data to improve the adaptability and performance of LLMs in RRG.
- We propose tri-consistency constraints that can effectively enhance the consistency and specificity of generated reports with historical data, ensuring alignment with disease progression observed in sequential chest X-rays.
- The proposed framework achieves superior performance without relying on historical data during testing and can be easily integrated with various multimodal large models, demonstrating its strong applicability.
- Extensive evaluations on the Longitudinal-MIMIC dataset demonstrate that our method achieves state-of-the-art performance, underscoring its effectiveness and robustness in leveraging historical data to enhance radiology report generation with LLMs.

Related Works

Radiology Report Generation: Radiology report generation methodologies have evolved significantly from early CNN-RNN frameworks (Vinyals et al. 2015; Lu et al. 2017; Wang et al. 2024a,b) to the integration of advanced Transformer architectures (Li et al. 2022; Wang et al. 2023a; Huang, Zhang, and Zhang 2023). Initial approaches focused on cross-modal alignment, utilizing CNNs for image features and RNNs for text generation. The advent of Transformers brought enhanced cross-modal interactions and long-range dependency modeling. Researchers introduced memory modules (Chen et al. 2020, 2021; Qin and Song 2022; Cao et al. 2023; Shen et al. 2024), hierarchical alignment (You et al. 2021; Li et al. 2023c) and knowledge-guided enhancement techniques (Li et al. 2023b; Huang, Zhang, and Zhang 2023; Li et al. 2023a, 2024) to better capture multi-level interactions. Despite their advancements, these approaches are tailored for single chest X-ray report generation and are limited in processing multimodal inputs, which constrains their applicability in longitudinal radiology report generation.

Recently, the advent of LLMs has further brought significant advancements to RRG. For instance, Liu et al. (2024a) proposed bootstrapping LLMs with in-domain instance induction and coarse-to-fine decoding to enhance alignment with medical data. Jin et al. (2024) proposed PromptMRG, which enhances radiology report generation by using diagnosis-driven prompts and addresses disease imbalance with adaptive loss techniques. Although LLMs can technically process sequential data by feeding the sequence directly into the model, this approach often encounters issues such as hallucination and suboptimal performance. To overcome these challenges, we propose a novel HC-LLM framework for RRG. By leveraging historical information and incorporating various types of constraints, HC-LLM ensures that LLM-generated reports can accurately capture the disease's progression nature.

Longitudinal Radiology Report Generation: Recent advancements in RRG have increasingly focused on leveraging longitudinal information to improve the accuracy and relevance of generated reports. Current research can be divided into two main directions. The first direction involves using only historical chest X-ray information. For example, Karwande et al. (2022) introduced an anatomy-aware model to track longitudinal relationships between chest X-rays, effectively capturing disease progression. Bannur et al. (2023) proposed a self-supervised framework to model the longitudinal evolution of chest X-ray findings, enhancing the understanding of disease changes over time. Additionally, Serra et al. (2023) employed Faster R-CNN to create composite representations of longitudinal studies, highlighting anatomical changes. The second direction, which is more aligned with our focus, leverages both historical chest X-rays and

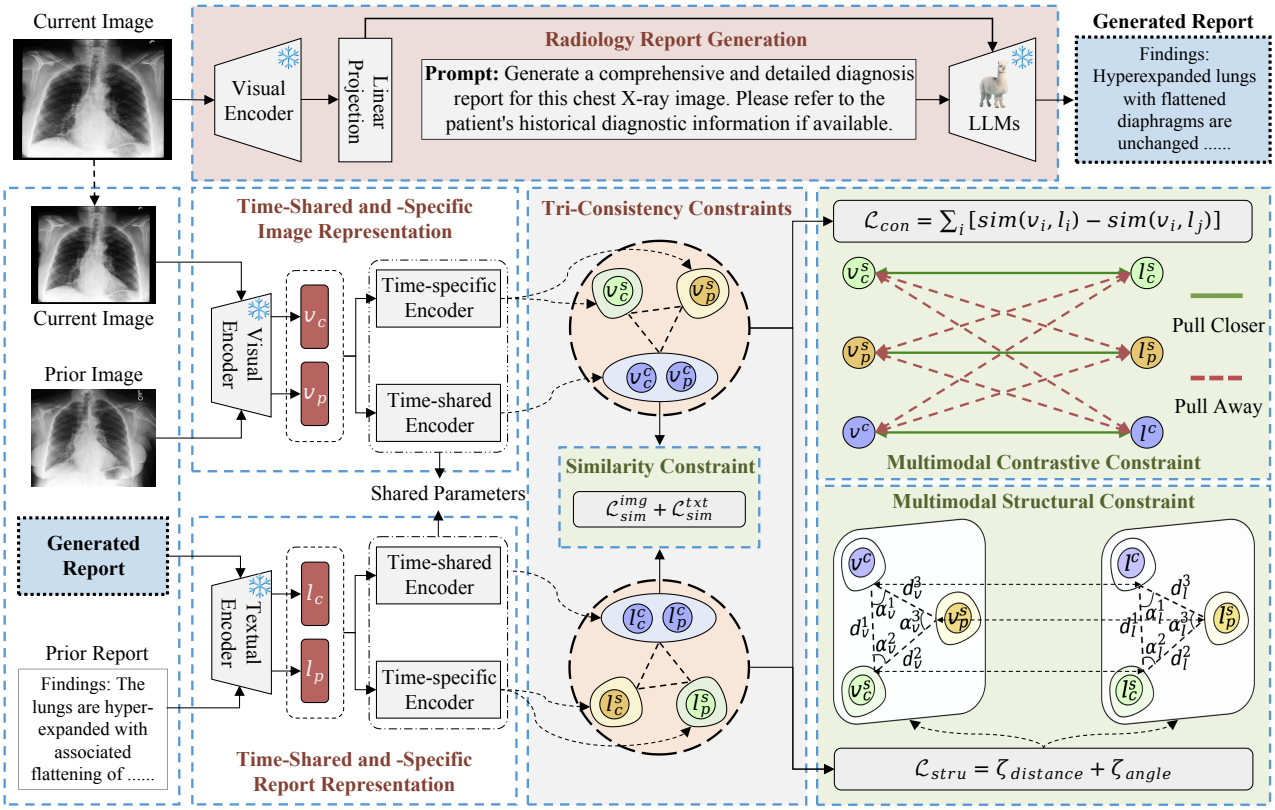

Figure 2: **Overview of the proposed framework:** First, the current chest X-ray is processed to generate a diagnostic report using a visual encoder and LLM. The framework then extracts time-shared and time-specific features from the current and prior chest X-rays, along with the generated and prior diagnostic reports. Then, similarity constraints are first applied to ensure consistent time-shared representation over time. Finally, multimodal contrastive and structural constraints are employed to align shared and specific features across modalities, ensuring the generated report accurately reflects disease progression.

diagnostic reports. This approach is crucial for capturing the full scope of disease progression and providing comprehensive context for current diagnoses. For instance, Zhu et al. (2023b) proposed a cross-attention-based multi-modal fusion framework to utilize patient record chronology, thereby improving report pre-filling tasks. Although these methods have made significant progress, they have not thoroughly explored the adaptation of LLMs to longitudinal medical data, often missing the complex progression of diseases, which significantly impacts the effectiveness of generated reports.

Method

Problem Formulation

The overall framework of HC-LLM is illustrated in Figure 2. The input comprises the current chest X-ray image (I_c) and the previous chest X-ray image (I_p) along with its corresponding diagnostic report (R_p). The objective is to generate a diagnostic report ($\hat{R}_c$) for I_c, that closely approximates the ground truth report (R_c). Formally, given:

$$\hat{R}_c \leftarrow \text{HC-LLM}(I_c, (I_p, R_p)). \tag{1}$$

Notably, our framework is flexible and supports two testing scenarios: 1) leveraging historical diagnostic information; 2) relying solely on the current chest X-ray image. Additionally, it can be extended to incorporate diagnostic information from multiple historical time points to assist in generating the current report. This flexibility enhances the practical applicability of our model in real-world clinical settings.

Radiology Report Generation

The overall workflow of the RRG is illustrated at the top part of Figure 2. This process consists of three main components: visual encoding, prompt generation, and report generation with the LLMs. Firstly, the visual encoder processes the current chest X-ray image I_c using the Swin Transformer (Liu et al. 2021b) $f_{ve}(\cdot)$, extracting latent visual features that capture the anatomical and pathological details from the radiograph. Formally, the visual feature extraction is defined as:

$$f_{ve}(I_c) = X = \{x_1, x_2, \ldots, x_S\}, \tag{2}$$

where $x_i \in \mathbb{R}^d$ is a feature patch, d denotes the feature dimension, and S is the number of patches. Next, for prompt generation, we define a general prompt p_g as shown in the middle section of the top part of Figure 2. It is important to note that the prompt can be adapted to include or exclude historical diagnostic information based on the input

provided during testing. Finally, the radiology report generation component utilizes a large language model f_{tg} to produce the diagnostic report $\hat{R}c$. Each report is represented as $\hat{R}_c = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_T\}$, where $\hat{r}_i \in \mathbb{V}$ is a token, T is the length of the report, and $\mathbb{V}$ represents the vocabulary. The decoding process is formulated as:

$$\hat{r}_t = f_{tg}(X, p_g, \hat{r}_{1:t-1}), \tag{3}$$

where $\hat{r}_t$ is the token to be predicted at token step t. The model is optimized based on the cross-entropy loss $\mathcal{L}_{RRG}$ from the final generated reports $\hat{R}_c$ and the gold standard R_c. The primary loss function is defined as:

$$\mathcal{L}_{RRG} = -\sum_{t=1}^{T} \log p(\hat{r}_t | \hat{r}_{1:t-1}, X, p_g). \tag{4}$$

Time-Shared and -Specific Representations

For chest X-rays and reports at two different time points, diseases often exhibit characteristics such as disappearance, stability, and emergence within each respective modality space. Therefore, we construct both time-shared and specific features for the chest X-rays and reports to capture these characteristics. For the images, we use the previously mentioned visual encoder f_{ve} and take the output at the `CLS` position as the representations of the two images, followed by a linear mapping layer to project these features into the text space, as shown below:

$$v_c, v_p = W_v \cdot f_{ve}(I_c)[\texttt{CLS}], W_v \cdot f_{ve}(I_p)[\texttt{CLS}], \tag{5}$$

where W_v represents the linear mapping layer for visual features. For the generated and historical reports, we separately input them into the text encoder f_{tg} and also take the output at the `CLS` position as their respective representations, as shown below:

$$l_c, l_p = f_{tg}(\hat{R}_c)[\texttt{CLS}], f_{tg}(R_p)[\texttt{CLS}]. \tag{6}$$

The time-shared and specific features are then extracted using dedicated encoders for both image and text modalities, as described below:

$$\begin{aligned} y_c^c, y_c^s &= E_c(LN(y_c), \theta^c), E_c^s(LN(y_c), \theta_c^s), y \in \{v, l\} \\ y_p^c, y_p^s &= E_c(LN(y_p), \theta^c), E_p^s(LN(y_p), \theta_p^s), y \in \{v, l\}, \end{aligned} \tag{7}$$

where y represents the modality, with y_c^c, y_p^c as shared features and y_c^s, y_p^s as specific features at current and prior times. $LN(\cdot)$ denotes the Layer Normalization. $E_c(\cdot)$ represents the shared encoder, while $E_c^s(\cdot)$ and $E_p^s(\cdot)$ represent the time-specific encoders for current and previous data, respectively. Specifically, using the same set of encoders for both image and text modalities could enhance multimodal alignment and integration, enabling more effective constraint application across modalities.

Tri-Consistency Constraints

Based on the time-shared and specific features of longitudinal chest X-rays and reports, we further introduce three constraints to enhance the performance of the LLMs in generating medical reports, ensuring that the generated reports accurately reflect the disease progression characteristics.

Similarity Constraint $\mathcal{L}_{sim}$. The similarity constraint is designed to align the time-shared features within each modality. Among various metric choices, we employ the Mean Squared Error (MSE) for this purpose, which measures the discrepancy by computing the average of the squares of the differences between corresponding values. The MSE loss is defined as follows:

$$\mathcal{L}_{sim}^{img} = \frac{1}{2}\sum \|v_c^c - v_p^c\|_2^2, \quad \mathcal{L}_{sim}^{txt} = \frac{1}{2}\sum \|l_c^c - l_p^c\|_2^2. \tag{8}$$

Despite its sensitivity to outliers, its simplicity and computational efficiency make it ideal for our similarity loss, enhancing the consistency of time-shared representations across different time points.

Multimodal Contrastive Constraint $\mathcal{L}_{con}$. The introduction of $\mathcal{L}_{con}$ serves two main purposes. Firstly, it aligns the shared and specific features of corresponding chest X-rays and reports, ensuring that the progression characteristics of diseases in images are consistent with those in reports. Secondly, by bringing the matching features between images and reports closer and distancing the non-matching features, it indirectly promotes the separation of the three features within the same modality, thereby enhancing their specific semantic information. Considering the semantic consistency of shared features and to facilitate the implementation of contrastive constraint, we perform average pooling on the two shared features within the same modality, as follows:

$$y^c = (y_c^c + y_p^c)/2, \quad y \in \{v, l\}. \tag{9}$$

Then, for the image sequence $\tilde{v} = [v^c, v_p^s, v_c^s]$ and text sequence $\tilde{l} = [l^c, l_p^s, l_c^s]$, we use the InfoNCE loss $\mathcal{L}_{con}$, which includes an image-to-text contrastive loss $\mathcal{L}_{i2t}$ and a text-to-image contrastive loss $\mathcal{L}_{t2i}$ to achieve the aforementioned objectives, denoted as:

$$\mathcal{L}_{con} = (\mathcal{L}_{i2t} + \mathcal{L}_{t2i})/2, \tag{10}$$

where the image-to-text contrastive loss $\mathcal{L}_{i2t}$ is formulated as:

$$\mathcal{L}_{i2t} = -\log \frac{\exp((\tilde{v}_i, \tilde{l}_i)/\tau)}{\sum_{k=1}^{3} \exp((\tilde{v}_i, \tilde{l}_k)/\tau)}, \tag{11}$$

where τ is the temperature hyper-parameter. Similarly, the text-to-image contrastive loss $\mathcal{L}_{t2i}$ is

$$\mathcal{L}_{t2i} = -\log \frac{\exp((\tilde{l}_i, \tilde{v}_i)/\tau)}{\sum_{k=1}^{3} \exp((\tilde{l}_i, \tilde{v}_k)/\tau)}. \tag{12}$$

By aligning the evolution of disease characteristics between images and text, we ensure that the generated reports more accurately reflect the longitudinal progression of medical conditions.

Multimodal Structural Constraint $\mathcal{L}_{stru}$. While multimodal contrastive constraint effectively pulls together matched features and pushes apart unmatched features, it does not sufficiently guarantee consistent structural relationships in the feature space. Thus, we further introduce multimodal structural constraint $\mathcal{L}_{stru}$, which ensures that the geometric relationships among features from images and reports remain consistent in the feature space. Following the

Dataset	Model	Year	Inputs	NLG metrics						CE metrics		
				BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	PREC	REC	F-1
Longitudinal -MIMIC	AoANet	2019	Single	0.272	0.168	0.112	0.080	0.115	0.249	0.437	0.249	0.317
	CNN+Trans	2019	Single	0.299	0.186	0.124	0.088	0.120	0.263	0.445	0.258	0.326
	R2Gen	2020	Single	0.302	0.183	0.122	0.087	0.124	0.259	0.500	0.305	0.379
	R2CMN	2021	Single	0.305	0.184	0.122	0.085	0.126	0.265	0.521	0.396	0.449
	R2GenRL	2022	Single	0.303	0.153	0.082	0.136	-	0.175	0.435	0.464	0.419
	CvT2DistilGPT2	2023	Single	0.365	0.226	0.151	0.107	0.143	0.275	0.443	0.369	0.379
	PromptMRG	2024	Single	0.370	0.219	0.141	0.098	0.144	0.266	0.519	**0.507**	0.482
	Prefilling	2023	Longitudinal	0.343	0.210	0.140	0.099	0.137	0.271	0.538	0.434	0.480
	R2GenGPT♠	2023	Single	0.358	0.224	0.150	0.103	0.235	0.269	0.228	0.151	0.168
	+ report	2023	Longitudinal	0.367	0.223	0.145	0.100	0.124	0.265	0.460	0.435	0.416
	+ image	2023	Longitudinal	0.332	0.194	0.125	0.082	0.145	0.237	0.341	0.267	0.277
	+ report & image	2023	Longitudinal	0.389	0.246	0.166	0.117	0.228	0.278	0.402	0.358	0.352
	HC-LLM(Ours)♠	-	Longitudinal	0.404	0.260	0.178	0.128	0.160	**0.287**	0.417	0.357	0.357
	BioMedGPT♦	2023	Single	0.365	0.230	0.155	0.111	0.085	0.266	0.269	0.242	0.237
	+ report	2023	Longitudinal	0.393	0.252	0.172	0.121	0.232	0.281	0.381	0.314	0.321
	+ image	2023	Longitudinal	0.356	0.225	0.149	0.102	0.133	0.265	0.252	0.176	0.194
	+ report & image	2023	Longitudinal	0.398	0.254	0.173	0.121	**0.279**	0.281	0.401	0.340	0.341
	HC-LLM(Ours)♦	-	Longitudinal	0.406	0.260	0.178	0.127	0.162	0.285	0.415	0.358	0.360
	MiniGPT4♣	2023	Single	0.375	0.231	0.150	0.099	0.135	0.266	0.193	0.112	0.133
	+ report	2023	Longitudinal	0.405	0.255	0.172	0.119	0.156	0.281	0.420	0.374	0.366
	+ image	2023	Longitudinal	0.365	0.226	0.149	0.100	0.146	0.266	0.182	0.129	0.141
	+ report & image	2023	Longitudinal	0.395	0.248	0.166	0.114	0.144	0.280	0.411	0.343	0.346
	HC-LLM(Ours)♣	-	Longitudinal	**0.416**	**0.276**	**0.193**	**0.142**	0.162	0.284	**0.617**	0.494	**0.498**

Table 1: Results of the NLG metrics (BLEU (BL), Meteor (MTR), Rouge-L (R-L)) and clinical efficacy (CE) metrics (Precision (PREC), Recall (REC) and F-1 score) on the *Longitudinal-MIMIC* dataset. Best results are highlighted in bold, and the second best are underlined. Identical symbols (i.e., ♠, ♦, ♣) in the table denote models using the same architecture.

methodology proposed by Park et al. (2019), we define the structural loss $\mathcal{L}_{stru}$ as a combination of distance-wise and angle-wise constraints to enhance the structural consistency of the representations.

Distance-wise loss: This constraint aligns the distances between features in the image space with those in the text space. Given a pair of feature representations $(t_i, t_j), t \in \{\tilde{v}, \tilde{l}\}$, the distance-wise function ψ_D measures the Euclidean distance between the two features as follows:

$$\psi_D(t_i, t_j) = \frac{1}{\mu}\|t_i - t_j\|_2, \tag{13}$$

where μ is a normalization factor for distance, defined as the average distance between pairs $\mathcal{X}^2$ within that modality:

$$\mu = \frac{1}{|\mathcal{X}^2|}\sum_{(t_i,t_j)\in\mathcal{X}^2}\|t_i - t_j\|_2. \tag{14}$$

The distance-wise constraint loss is defined as follows:

$$\mathcal{L}_{distance} = \sum_{i,j=1,i\neq j}^{3} l_\delta(\psi_D(\tilde{v}_i, \tilde{v}_j), \psi_D(\tilde{l}_i, \tilde{l}_j)), \tag{15}$$

where l_δ is the Huber loss, defined as:

$$l_\delta(x, y) = \begin{cases} \frac{1}{2}(x-y)^2, & \text{if } |x-y| \leq 1, \\ |x-y| - \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{16}$$

Angle-wise loss: This loss ensures that the angles between triplets of features in the image modality are consistent with those in the text modality. This is achieved by calculating the angles formed by three examples using cosine similarity:

$$\psi_A(t_i, t_j, t_k) = \frac{\langle t_i - t_j, t_i - t_k\rangle}{\|t_i - t_j\| \cdot \|t_i - t_k\|}, t \in \{\tilde{v}, \tilde{l}\}. \tag{17}$$

The angle-wise constraint is defined as follows:

$$\mathcal{L}_{angle} = \sum_{\substack{i,j,k\in\{1,2,3\} \\ i\neq j\neq k}} l_\delta(\psi_A(\tilde{v}_i, \tilde{v}_j, \tilde{v}_k), \psi_A(\tilde{l}_i, \tilde{l}_j, \tilde{l}_k)). \tag{18}$$

Thus, the final $\mathcal{L}_{stru}$ can be summarized as:

$$\mathcal{L}_{stru} = \mathcal{L}_{distance} + \mathcal{L}_{angle}. \tag{19}$$

By minimizing this structural loss, we ensure that the geometric relationships within the image features are mirrored in the text features, thereby reinforcing the structural consistency between both modalities.

Learning Objective

The overall learning of the model is performed by minimizing:

$$\mathcal{L}_{total} = \mathcal{L}_{RRG} + \beta_1(\mathcal{L}_{sim}^{img} + \mathcal{L}_{sim}^{txt}) + \beta_2\mathcal{L}_{con} + \beta_3\mathcal{L}_{stru}, \tag{20}$$

where β_1, β_2, β_3 are the hyperparameters that determine the contribution of each regularization component to the overall loss $\mathcal{L}_{total}$.

Experiments

Dataset: Building on the dataset presented in (Zhu et al. 2023b), we utilized the Longitudinal-MIMIC dataset, which

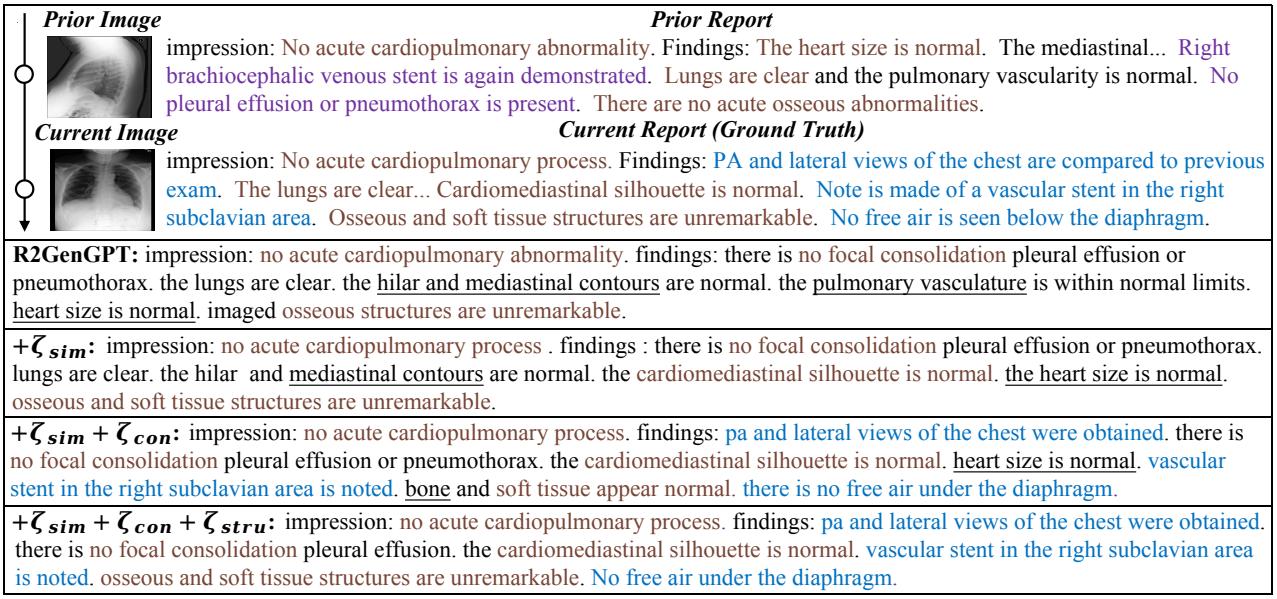

Figure 3: An illustration of reports generated by different models using longitudinal images and the historical report. Brown denotes common content, while purple and blue indicate time-specific content. Underlined text marks incorrect predictions.

$\mathcal{L}_{sim}$	$\mathcal{L}_{con}$	$\mathcal{L}_{stru}$	NLG metrics						CE metrics		
			BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	PREC	REC	F-1
✗	✗	✗	0.389	0.246	0.166	0.117	0.228	0.278	0.402	0.358	0.352
✗	✓	✓	0.403	0.255	0.172	0.119	**0.232**	0.282	**0.419**	**0.362**	**0.361**
✓	✗	✓	0.383	0.244	0.166	0.118	0.148	0.278	0.367	0.311	0.311
✓	✓	✗	0.399	0.252	0.171	0.121	0.230	0.280	0.417	0.355	0.356
✓	✓	✓	**0.404**	**0.260**	**0.178**	**0.128**	0.160	**0.287**	0.417	0.357	0.357

Table 2: Ablation study of each constraint on the dataset of *Longitudinal-MIMIC*.

is derived from MIMIC-CXR, for our evaluation. This dataset was constructed by selecting patients with at least two visit records, resulting in a comprehensive dataset of 26,625 patients and a total of 94,169 samples. Each sample used for model training included the current visit's chest X-ray (CXR) and report, as well as the previous visit's CXR and report. The dataset was divided into training (26,156 patients and 92,374 samples), validation (203 patients and 737 samples), and test (266 patients and 2,058 samples) sets.

Evaluation Metrics: We assess the performance of our model using both natural language generation (NLG) metrics and clinical efficacy (CE) metrics. For NLG, we utilize the BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2011), and ROUGE-L (Lin 2004) metrics. In accordance with the methodology proposed by Nicolson, Dowling, and Koopman (2023), our CE evaluation involves precision, recall, and F1 scores. These metrics are derived by converting generated reports into 14 disease classification labels using CheXbert (Smit et al. 2020).

Implementation Details: In this study, for the R2GenGPT (Wang et al. 2023b) framework, we selected the base version of the Swin Transformer[1] as the visual encoder and used the LLAMA2-7B[2] model as the primary language model for both R2GenGPT and MiniGPT4 (Zhu et al. 2023a) frameworks. BioMedGPT[3] (Luo et al. 2023) maintains consistency with the R2GenGPT image encoder and utilizes BioMedGPT-LM-7B as its language model. The coefficients were set to $\beta_1 = 1.0$, $\beta_2 = 0.8$, and $\beta_3 = 1.0$, respectively. The training process was executed on a single NVIDIA A800 80GB GPU using mixed precision for 5 epochs on the Longitudinal-MIMIC dataset, with a mini-batch size of 4 and a learning rate of 1e-4. For the testing phase, we employed a beam search strategy with a beam size of 3. Further implementation details can be found at https://github.com/TengfeiLiu966/HC-LLM.

Comparison with State-of-the-Art Methods

We compared our method with single time-point RRG methods (i.e., AoANet (Huang et al. 2019), CNN+Trans, R2Gen

[1]https://huggingface.co/microsoft/swin-base-patch4-window7-224

[2]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[3]https://huggingface.co/PharMolix/BioMedGPT-LM-7B

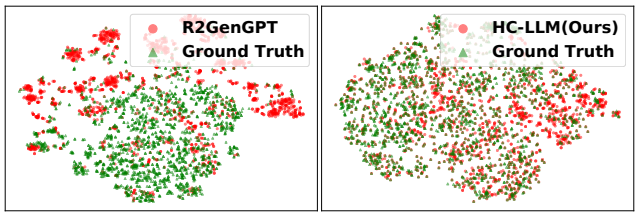

Figure 4: Visualization of feature distributions using t-SNE for the R2GenGPT and HC-LLM (Ours) models.

Models	BL-1	BL-2	BL-3	BL-4	MTR	R-L
PromptMRG	0.370	0.219	0.141	0.098	0.144	0.266
Prefilling	0.253	0.159	0.107	0.077	0.118	**0.269**
R2GenGPT	0.358	0.224	0.150	0.103	**0.235**	**0.269**
BioMedGPT	0.346	0.211	0.136	0.088	0.096	0.255
MiniGPT4	0.344	0.181	0.106	0.063	-	0.222
HC-LLM(Ours)	**0.371**	**0.231**	**0.154**	**0.107**	0.127	0.268

Table 3: Performance comparison without historical information during testing. The HC-LLM model operates within the R2GenGPT framework.

(Chen et al. 2020), R2CMN (Chen et al. 2021), R2GenRL (Qin and Song 2022), CvT2DistilGPT2 (Nicolson, Dowling, and Koopman 2023), PromptMRG (Jin et al. 2024)) and longitudinal RRG methods (i.e., Prefilling (Zhu et al. 2023b), R2GenGPT (Wang et al. 2023b), BioMedGPT (Luo et al. 2023), MiniGPT4 (Zhu et al. 2023a)). As shown in Table 1, our method achieves improvements in most metrics. Specifically, compared to single time-point methods, longitudinal models generally exhibit superior performance, demonstrating the importance of modeling longitudinal information for the RRG task. Additionally, for longitudinal models, traditional cross-attention methods fall short of LLM-based approaches due to the latter's superior semantic modeling capabilities. Nevertheless, LLM-based baselines still underperform compared to our proposed method, primarily because simply inputting longitudinal data into large language models does not fully utilize the unique longitudinal characteristics. In contrast, our model constrains the consistency of disease progression within longitudinal chest X-rays and reports, ensuring that the LLM-generated reports accurately reflect disease progression, thereby enhancing their accuracy. Additionally, we can observe that our method achieves performance improvements across different frameworks, demonstrating its strong applicability.

Model Analysis

Ablation Study. Table 2 presents an ablation study of each constraint on the Longitudinal-MIMIC dataset. Firstly, when the similarity constraint $\mathcal{L}_{sim}$ is removed, there is only a minor performance drop. This can be attributed to the fact that although removing the similarity constraint disrupts the alignment of shared features, the presence of contrastive and structural constraints still maintains a certain degree of cross-modal longitudinal consistency. Secondly, removing the contrastive constraint $\mathcal{L}_{con}$ results in a significant performance decrease. The core reason is that the lack of a contrastive constraint hinders the alignment of shared and specific features across modalities. Additionally, this indirectly prevents the promotion of separation between features within the same modality, thereby failing to ensure the accuracy of specific features, which ultimately impacts the overall model performance. Lastly, when the structural constraint $\mathcal{L}_{stru}$ is removed, there is also a noticeable performance decline, highlighting its importance in maintaining cross-modal longitudinal consistency and ensuring the effectiveness of the LLMs' outputs.

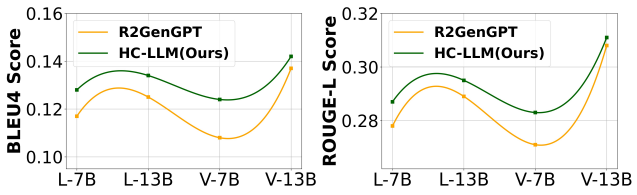

Figure 5: Performance comparison of BLEU-4 and ROUGE-L scores for R2GenGPT and HC-LLM(Ours) models across different LLMs (L: LLaMA2, V: Vicuna; 7B/13B: model sizes).

Qualitative results. To qualitatively demonstrate how historical information, under different constraints, better adapts LLMs to RRG, we perform a case study on the output reports generated with various combinations of constraints using the same longitudinal inputs. As shown in Figure 3, the reports at different time points indeed show disease disappearance (purple), stability (brown), and new occurrences (blue). When feeding longitudinal chest X-rays and historical reports directly to R2GenGPT, it generates more generic content, not fully utilizing longitudinal data to produce a targeted report for the current X-ray. Upon introducing similarity constraints, the generated report includes more common content, aligning with our expected outcome. With the addition of contrastive constraint, both common and unique contents are reflected in the generated report. This is mainly due to the contrastive constraint ensuring that the distinctive features in the generated report align with those in the current chest X-ray, effectively promoting the generation of specific content. Finally, by introducing structural constraint, we observe that the generated report's accuracy improves, and certain erroneous predictions are eliminated. The contrastive constraint helps maintain consistency, while the structural constraint significantly enhances this consistency, providing better regulation and adaptation to the LLMs' generative performance. To more intuitively display the distribution of features before and after applying constraints, we used t-SNE to visualize the distributions between the R2GenGPT, our method, and the actual reports. As shown in Figure 4, the distribution of our method aligns more closely with the actual reports, more directly confirming its superior generative performance.

Testing Performance without Historical Data. We further evaluate the longitudinal models using only the current chest X-ray during testing. As shown in Table 3, the performance

of traditional models drops significantly, demonstrating their limited applicability. While methods based on LLMs are relatively more stable, they also experience some performance decline. Notably, R2GenGPT, BioMedGPT and MiniGPT4 perform worse than single time-point PromptMRG when tested using only the current chest X-ray. This is primarily because it does not effectively utilize historical information to adapt LLMs to the RRG task. In contrast, our model outperforms the PromptMRG method. This superior performance is attributed to our training process, which better captures the evolutionary characteristics of diseases in longitudinal data and more effectively adapts LLMs to RRG.

Performance Analysis Under Different LLMs. Figure 5 shows the results of our method under different LLMs. As observed, regardless of whether LLaMA2-7B, LLaMA2-13B, Vicuna-7B, or Vicuna-13B is used, our method consistently achieves better results in both BLEU-4 and ROUGE-L metrics. This indicates that our model has good adaptability and robustness across various LLM architectures, leading to stable improvements in report generation quality. Notably, larger models exhibit improved performance, likely because they contain more general information and can better adapt to RRG with the help of our introduced constraints.

Conclusion

In this paper, we propose a novel HC-LLM framework that leverages historical diagnostic information to ensure that the reports generated by LLMs better align with the progression of diseases. Experimental results demonstrate that our method exhibits superior performance with both single chest X-ray data and longitudinal data during testing, proving its effectiveness. Additionally, our architecture can easily adapt to different multimodal large model frameworks and achieve substantial performance improvements, demonstrating its excellent applicability. This method provides a practical paradigm for adapting general LLMs to sequential data applications. Currently, HC-LLM only uses two-time point longitudinal data and has not yet explored more complex diagnostic data from multiple historical time points, which is key for understanding disease progression and could be explored in the future to further improve performance.

Acknowledgments

This work was supported by National Key R&D Program of China No.2021ZD0111902 and National Natural Science Foundation of China under Grant 62172022, Grant U21B2038, Grant 62476179, and Grant 62476015.

References

Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D. C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15016–15027.

Cao, Y.; Cui, L.; Zhang, L.; Yu, F.; Li, Z.; and Xu, Y. 2023. MMTN: multi-modal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 277–285.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 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.

Denkowski, M.; and Lavie, A. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 85–91.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4634–4643.

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 (CVPR)*, 19809–19818.

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 (AAAI)*, volume 38, 2607–2615.

Karwande, G.; Mbakwe, A. B.; Wu, J. T.; Celi, L. A.; Moradi, M.; and Lourentzou, I. 2022. Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 581–591.

Li, M.; Cai, W.; Verspoor, K.; Pan, S.; Liang, X.; and Chang, X. 2022. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20656–20665.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023a. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3334–3343.

Li, M.; Lin, H.; Qiu, L.; Liang, X.; Chen, L.; Elsaddik, A.; and Chang, X. 2024. Contrastive Learning with Counterfactual Explanations for Radiology Report Generation. arXiv:2407.14474.

Li, M.; Liu, R.; Wang, F.; Chang, X.; and Liang, X. 2023b. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1): 253–270.

Li, Y.; Yang, B.; Cheng, X.; Zhu, Z.; Li, H.; and Zou, Y. 2023c. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), 2863–2874.

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 (AAAI)*, volume 38, 18635–18643.

Liu, C.; Tian, Y.; and Song, Y. 2023. A systematic review of deep learning-based research on radiology report generation. arXiv:2311.14199.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 13753–13762.

Liu, R.; Li, M.; Zhao, S.; Chen, L.; Chang, X.; and Yao, L. 2024b. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, 8721–8730.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 375–383.

Luo, Y.; Zhang, J.; Fan, S.; Yang, K.; Wu, Y.; Qiao, M.; and Nie, Z. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. arXiv:2308.09442.

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

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 (ACL)*, 311–318.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3967–3976.

Qin, H.; and Song, Y. 2022. Reinforced cross-modal alignment for radiology report generation. In *Proceedings of the Association for Computational Linguistics (ACL)*, 448–458.

Serra, F. D.; Wang, C.; Deligianni, F.; Dalton, J.; and O'Neil, A. Q. 2023. Controllable chest x-ray report generation from longitudinal representations. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shen, H.; Pei, M.; Liu, J.; and Tian, Z. 2024. Automatic Radiology Reports Generation via Memory Alignment Network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 4776–4783.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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 (CVPR)*, 7433–7442.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164.

Wang, J.; Cui, Z.; Wang, B.; Pan, S.; Gao, J.; Yin, B.; and Gao, W. 2024a. IME: Integrating Multi-curvature Shared and Specific Embedding for Temporal Knowledge Graph Completion. In *Proceedings of the ACM on Web Conference 2024*, 1954–1962.

Wang, J.; Wang, B.; Gao, J.; Pan, S.; Liu, T.; Yin, B.; and Gao, W. 2024b. MADE: Multicurvature Adaptive Embedding for Temporal Knowledge Graph Completion. *IEEE Transactions on Cybernetics*.

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 (CVPR)*, 11558–11567.

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

You, D.; Liu, F.; Ge, S.; Xie, X.; Zhang, J.; and Wu, X. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 72–82.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592.

Zhu, Q.; Mathai, T. S.; Mukherjee, P.; Peng, Y.; Summers, R. M.; and Lu, Z. 2023b. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 189–198.