

Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation

Kang Liu¹ Zhuoqi Ma^{1*} Xiaolu Kang¹ Yunan Li¹ Kun Xie¹ Zhicheng Jiao² Qiguang Miao¹

¹Xidian University, China ²Brown University, USA

{kangliu, 22031212472}@stu.xidian.edu.cn, {zhuoqima, yunanli, xiekun, qgmiao}@xidian.edu.cn, zhicheng_jiao@brown.edu

Abstract

Automated radiology report generation offers an effective solution to alleviate radiologists' workload. However, most existing methods focus primarily on single or fixed-view images to model current disease conditions, which limits diagnostic accuracy and overlooks disease progression. Although some approaches utilize longitudinal data to track disease progression, they still rely on single images to analyze current visits. To address these issues, we propose enhanced contrastive learning with **Multi-view Longitudinal data** to facilitate chest X-ray **Report Generation**, named **MLRG**. Specifically, we introduce a multi-view longitudinal contrastive learning method that integrates spatial information from current multi-view images and temporal information from longitudinal data. This method also utilizes the inherent spatiotemporal information of radiology reports to supervise the pre-training of visual and textual representations. Subsequently, we present a tokenized absence encoding technique to flexibly handle missing patient-specific prior knowledge, allowing the model to produce more accurate radiology reports based on available prior knowledge. Extensive experiments on MIMIC-CXR, MIMIC-ABN, and Two-view CXR datasets demonstrate that our MLRG outperforms recent state-of-the-art methods, achieving a 2.3% BLEU-4 improvement on MIMIC-CXR, a 5.5% F1 score improvement on MIMIC-ABN, and a 2.7% F1 RadGraph improvement on Two-view CXR.

1. Introduction

Chest X-ray (CXR) is a widely employed diagnostic tool in clinical practice, primarily for evaluating the lungs, heart, pleura, and skeletal structures. It is critical for diagnosing conditions, such as pneumonia, fracture, pneumothorax, pleural effusion, and cardiomegaly [18]. To ensure effective communication across departments and between

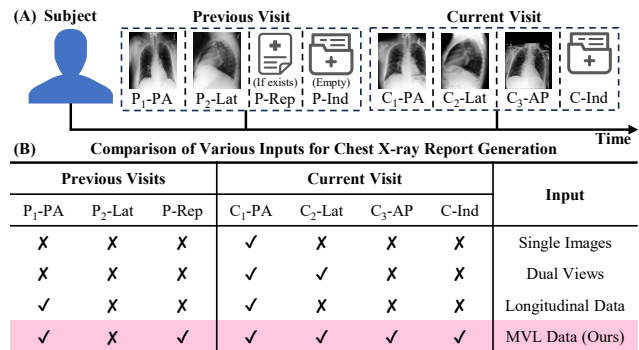


Figure 1. (A) shows medical historical data of a subject (patient) over time. (B) compares inputs for RRG, with *AP* and *PA* as frontal views, and *Lat* and *Rep* as a lateral view and its report. *Ind* and *MVL Data* are “INDICATION” and multi-view longitudinal data.

physicians and patients, radiologists manually document detailed reports based on their interpretation of CXR images. However, this process is both expertise-dependent and time-consuming [29, 50]. As the demand for imaging studies continues to grow, the workload associated with manual report generation may intensify, potentially impacting medical efficiency and compromising patient care quality [2, 49]. To mitigate these challenges, radiology report generation (RRG) [13, 43] has emerged as a promising solution. By automatically analyzing imaging data from X-ray [27], CT [15], or pathology [14], RRG generates clinical findings using factual terminology [30, 56] and descriptive language. This automation aids radiologists by providing high-quality draft reports [29], improving diagnostic efficiency.

In clinical practice, radiologists typically conduct comprehensive evaluations using multi-view images from the current visit, incorporate patient medical histories (i.e., longitudinal data) to track disease progression, and integrate patient-specific prior knowledge to assist in diagnosis and report generation. However, most existing RRG methods [8, 32, 35] focus solely on single images when generating reports and struggle to effectively differentiate between views, such as posteroanterior (PA), anteroposterior (AP),

*Corresponding author. The code is available at <https://github.com/mk-runner/MLRG>

lateral, or left anterior oblique. These views exhibit inherent differences; for example, although both PA and AP views are frontal images, geometric variations can cause cardiac enlargement in the AP view, potentially impacting diagnostic accuracy. To address this issue, some studies [9, 59] have introduced dual-view report generation, as illustrated in Figure 1. Empirical results [8, 9] reveal that incorporating dual views enhances the quality of generated reports. Nevertheless, these methods merely distinguish between frontal and lateral views, neglecting more subtle differences across multiple views. Moreover, both single-image and dual-view methods focus solely on the images from the current visit, disregarding the descriptions of disease progression found in radiology reports. This limitation may lead to model hallucinations. To combat this problem, some studies [22, 49, 65] have sought to leverage longitudinal data to model disease progression, as shown in Figure 1. However, these methods still rely on a single image to characterize the current visit, limiting diagnostic accuracy. Additionally, some patients may lack “INDICATION”, “previous report”, or “previous image” due to their first visit or improper data storage. This variability challenges the flexible use of available data to generate accurate reports.

To mimic radiologists’ diagnostic pipeline and address these challenges, we propose a two-stage MLRG for chest X-ray report generation. In Stage 1, the key part is our proposed multi-view longitudinal contrastive learning approach, which utilizes the inherent spatiotemporal information in radiology reports to supervise the pre-training of visual and textual representations. Specifically, we incorporate learnable position embeddings for each view to identify differences across varying numbers of views. We then employ a multi-view longitudinal fusion network that flexibly integrates spatial information from current multi-view images and temporal information from longitudinal data. Subsequently, we learn visual and textual representations by leveraging agreements between multi-view longitudinal data (see Figure 1) and their corresponding reports. In Stage 2, we introduce a tokenized absence encoding technique to handle missing patient-specific prior knowledge (i.e., “INDICATION” and “previous report”). This allows the multi-modal fusion network to adapt flexibly to the presence or absence of such data, improving the accuracy of the generated reports. Extensive experiments on MIMIC-CXR [20], MIMIC-ABN [34], and Two-view CXR [33] datasets confirm the superiority of MLRG in producing clinically accurate reports. Our contributions are stated as follows:

- We propose a novel multi-view longitudinal contrastive learning method that flexibly integrates multi-view longitudinal data and leverages the inherent spatiotemporal information from reports to supervise the pre-training of visual and textual representations.
- We introduce a tokenized absence encoding technique to

handle missing patient-specific prior knowledge. This technique enables the model to adapt flexibly to scenarios with or without such data, ensuring the text generator can utilize available prior knowledge effectively.

- Our MLRG shows competitive results compared to various state-of-the-art methods across three public datasets: MIMIC-CXR, MIMIC-ABN, and Two-view CXR.

2. Related Work

Radiology report generation (RRG). RRG is akin to image captioning [24, 61], but requires generating detailed content with specialized medical terminology. Existing RRG methods consist of a vision encoder (like ResNet101 [9, 17, 59], CvT [35, 36], or ViT [31, 53]) and a text generator (such as Memory-driven Transformer [8, 30], MiniGPT-4 [28], DistilGPT2 [27, 35], or LLaMA [23, 31, 52]). To improve clinical accuracy in RRG, researchers have incorporated various techniques or prior knowledge, including knowledge graphs [58, 64], cross-modal alignment [6, 30], region-guided frameworks [45, 62], warm starting [35], patient-specific “INDICATION” [29, 33], disease labels [59], and disease progression [16, 49]. However, these methods rely on single-image or fixed-view data, missing the comprehensive insights supplied by multi-view longitudinal data. To address this issue, we propose the MLRG method, which flexibly captures spatiotemporal features from multi-view longitudinal data and generates radiology reports based on patient-specific prior knowledge.

Medical vision-language models. These models aim to learn generalized medical visual representations by maximizing agreements between image-report pairs. MGCA [48] presents multi-granularity cross-modal alignment, harnessing agreements at the instance, pathological region, and disease levels. KAD [63] enhances visual representation using established knowledge graphs. MedCLIP [54] expands training sets by decoupling images and reports, reducing false negatives through semantic matching loss. BioViL-T [1] captures disease progression by analyzing longitudinal data. Despite notable improvements in tasks like medical image classification and image-text retrieval, the utilization of multi-view longitudinal data remains limited, constraining diagnostic accuracy. Therefore, we present a multi-view longitudinal contrastive learning approach that utilizes the inherent spatiotemporal information of radiology reports to guide the pre-training of visual and textual representations.

Enhancing medical image analysis via multi-view data. Multi-view learning [55] empowers models to derive shared and complementary insights from multiple views of the same subject. FMVP [32] treats single-view images and auxiliary inputs (i.e., disease labels and medical concepts) as multi-view data to assist the text generator in producing radiology reports. However, its reliance on additional annotated disease labels limits broader applicability. CXRMate

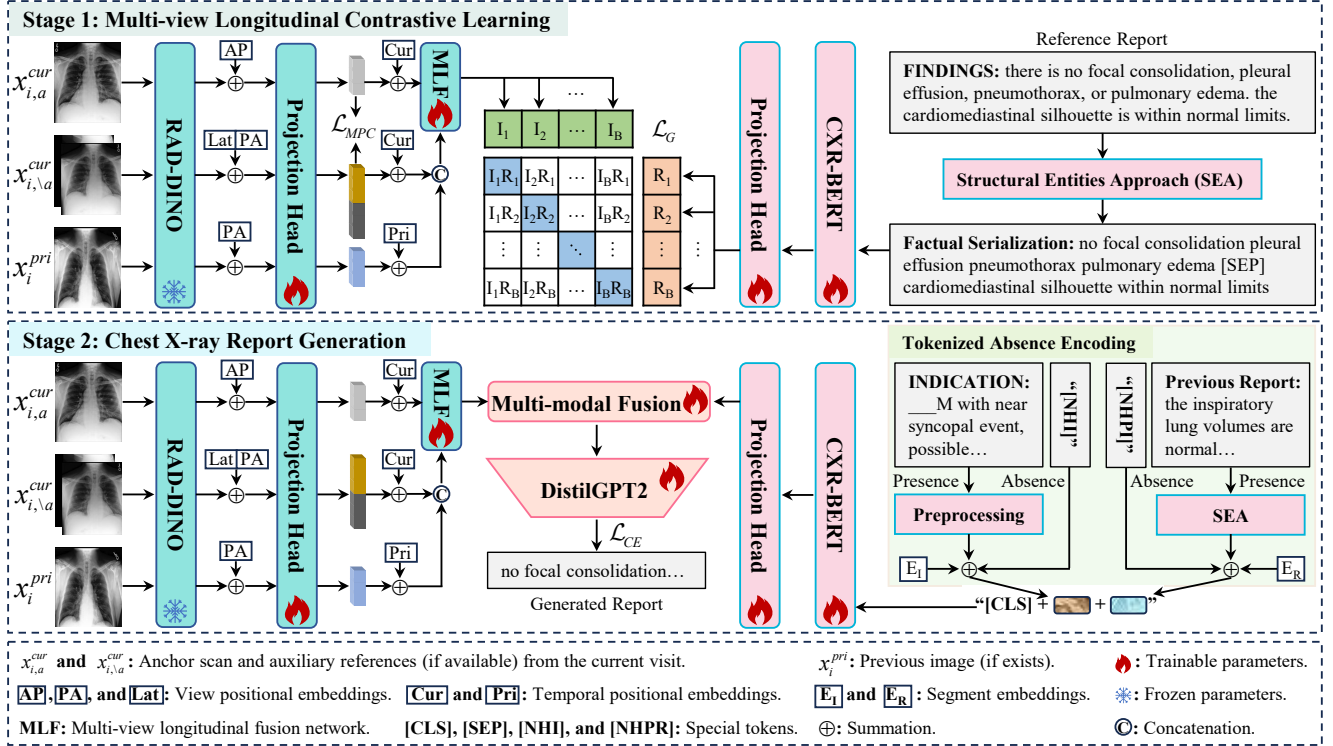


Figure 2. Overview of our proposed MLRG, including a vision encoder (RAD-DINO [39]), a text encoder (CXR-BERT [3]), and a text generator (DistilGPT2 [41]). MLRG first learns visual features through multi-view longitudinal contrastive learning and then generates radiology reports based on patient-specific prior knowledge.

[36] synthesizes previous reports based on previous images and integrates them with current multi-view images to produce final reports. However, this method overlooks subtle differences across views and can introduce additional noise from previous reports. In response, we propose the MLRG approach, which captures inter-view differences and leverages spatiotemporal information in reports to guide the pre-training, all without relying on additional manual labels.

3. Method

Figure 2 presents an overview of our proposed MLRG. In Stage 1, we introduce a multi-view longitudinal contrastive learning approach that leverages inherent spatiotemporal information from radiology reports to supervise the pre-training of visual and textual representations. In Stage 2, we propose a tokenized absence encoding technique to handle missing patient-specific prior knowledge, ensuring the generation of more coherent and accurate radiology reports based on available prior knowledge.

3.1. Problem Formation

Let $\mathcal{D}_{tr} = \{(x_i^{pri}, y_i^{pri}, z_i, X_i^{cur}, y_i^{cur})\}_{i=1}^n$ be the training set, where n denotes the total number of visits. Each visit consists of a frontal previous image x_i^{pri} (which may

be absent), a previous report y_i^{pri} (which may be absent), an “INDICATION” z_i (which may be absent), m_i current images (views) X_i^{cur} , and a reference report y_i^{cur} . Notably, the number of current multi-view images m_i may vary across visits. Our goal is to learn the function $F_\theta(\cdot)$ that maps $(x_i^{pri}, y_i^{pri}, z_i, X_i^{cur})$ to y_i^{cur} on the training set \mathcal{D}_{tr} , such that $F_\theta(x_i^{pri}, y_i^{pri}, z_i, X_i^{cur}) \rightarrow y_i^{cur}$. We then utilize the learned function $F_\theta(\cdot)$ to generate a radiology report based on current multi-view images, the previous image, and patient-specific prior knowledge (i.e., y_i^{pri} and z_i).

3.2. Multi-view Longitudinal Contrastive Learning

Visual features extraction. We employ RAD-DINO [39], a vision transformer model [12] trained solely on chest X-rays using DINOv2 [37], as the vision encoder. The feature maps from the last hidden state are treated as visual features $V \in \mathbb{R}^{M \times p \times d_1}$, where $M = \sum_{i=1}^B m_i$ denotes the total number of images in the mini-batch. Here, B , p , and d_1 represent the batch size, the number of patches, and the feature dimension, respectively.

Textual features extraction. Inspired by FSE [30], we first adopt the structural entities approach [30] to extract factual serialization, which comprises exclusively clinical descriptions from radiology reports, as shown in Figure 2.

This approach enables the model to focus on aligning images with factual serialization. We then consider CXR-BERT [3], a language model tailored for chest X-rays, as the text encoder. This is followed by a simple projection head that generates textual features $\mathbf{R} \in \mathbb{R}^{B \times t \times d}$, where t denotes the number of tokens, and d is the hidden size.

Multi-positive contrastive learning between current multi-view images to improve the consistency of visual features. In clinical practice, radiologists often select some representative images as primary references, with other images as auxiliary support. Therefore, we treat each image from current multi-view images X_i^{cur} as an anchor scan $x_{i,a}^{cur}$ while considering the remaining images as auxiliary references $x_{i,\setminus a}^{cur} = \{x_{i,j}^{cur} | j \neq a, x_{i,j}^{cur} \in X_i^{cur}\}$, where $a \in [1, m_i]$. To identify differences among views, we incorporate learnable view positional embeddings $\mathbf{E}_v \in \mathbb{R}^{M \times 1 \times d_1}$ into visual features. This is followed by a simple projection head $P_v(\cdot)$ that maps the features to a specific dimension d . These processes are formulated as follows:

$$\mathbf{V} = P_v(\mathbf{V} + \mathbf{E}_v) \in \mathbb{R}^{M \times p \times d}. \quad (1)$$

To enhance the consistency of visual features, we employ multi-positive contrastive learning [46] to maximize the similarity between images from the same visit while minimizing the similarity to images from different visits. Specifically, we first exclude visits with only one image, as they do not provide positive pairs (Notably, these visits are still used for subsequent cross-modal alignment). Following this, we calculate the predicted categorical distribution $\mathbf{q} \in \mathbb{R}^{K \times (K-1)}$ to estimate the similarity between images:

$$\mathbf{q}_i = \frac{\exp(\mathbf{v}_i \cdot \mathbf{v}_{\setminus i}^T / \tau_1)}{\sum_{j=1, j \neq i}^M \exp(\mathbf{v}_i \cdot \mathbf{v}_j^T / \tau_1)}, \text{ s.t. } m_i \neq 1, \quad (2)$$

where $K = \sum_{i=1, m_i \neq 1}^M m_i$ represents the total number of multi-view images in the mini-batch, and $\tau_1 \in \mathbb{R}^+$ is a temperature parameter. $\mathbf{v}_i \in \mathbb{R}^{1 \times d}$ refers to the global visual feature of the i^{th} image, while $\mathbf{v}_{\setminus i} \in \mathbb{R}^{(K-1) \times d}$ denotes global visual features of all multi-view images except the i^{th} image. Next, we compute the ground-truth categorical distribution $\mathbf{p} \in \mathbb{R}^{K \times (K-1)}$ by assigning the same labels to images from the same visit, formulated as:

$$\mathbf{p}_i = \frac{\mathbb{I}_{match}(\mathbf{v}_i, \mathbf{v}_{\setminus i})}{\sum_{j=1, m_j \neq 1}^M \mathbb{I}_{match}(\mathbf{v}_i, \mathbf{v}_j)}, \quad (3)$$

where $\mathbb{I}_{match}(\cdot, \cdot)$ is an indicator function that determines whether two visual features originate from the same visit. Although the number of current views m_i may vary across visits, the different number of non-zero elements in \mathbf{p}_i account for this variability. Finally, the multi-positive con-

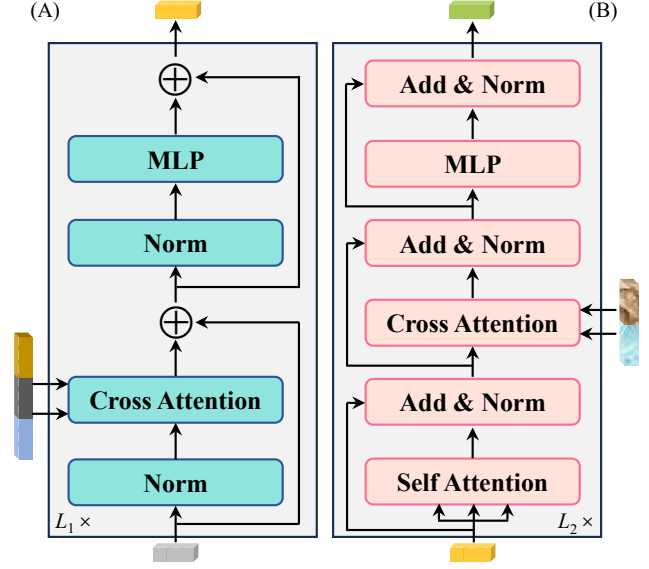


Figure 3. (A) represents the multi-view longitudinal fusion (MLF) network. (B) denotes the multi-modal fusion network.

trastive (MPC) loss is calculated using the cross-entropy between \mathbf{q} and \mathbf{p} , represented as:

$$\mathcal{L}_{MPC} = -\frac{1}{K} \sum_{i=1, m_i \neq 1}^M \mathbf{p}_i \log \mathbf{q}_i. \quad (4)$$

Multi-view longitudinal fusion network. Due to the varying number of current multi-view images and the absence of previous images for some patients, integrating this information flexibly presents certain challenges. To address this issue, we design the multi-view longitudinal fusion (MLF) network, as illustrated in Figure 3(A). We select the most recent previous visit to model temporal information, as it typically holds the highest reference value. To distinguish different time points, we integrate temporal positional embeddings into visual features, as depicted in Figure 2. Subsequently, the spatiotemporal features $\mathbf{V}^{st} = \{\mathbf{v}_1^{st}, \mathbf{v}_2^{st}, \dots, \mathbf{v}_B^{st}\}$ are extracted using the MLF network:

$$\mathbf{v}_i^{st} = \text{MLF}(\mathbf{v}_{i,a}^{cur}, [\mathbf{v}_{i,\setminus a}^{cur}, \mathbf{v}_i^{pri}]) \in \mathbb{R}^{p \times d}, \quad (5)$$

where the anchor scan $\mathbf{v}_{i,a}^{cur}$ functions as the query, and the concatenation of auxiliary references and previous image, $[\mathbf{v}_{i,\setminus a}^{cur}, \mathbf{v}_i^{pri}]$, serves as the key and value. Although the number of images for the current visit varies and some patients may lack previous images, we process one sample at a time, allowing for flexible adaptation to these changes.

Instance-wise cross-modal alignment. Radiology reports not only describe the current visit's condition but may also include comparisons with the patient's medical history. Therefore, relying solely on the current multi-view images

Split	MIMIC-CXR					MIMIC-ABN					Two-view CXR				
	#Img	#Rep	%Ind	%PI	%PR	#Img	#Rep	%Ind	%PI	%PR	#Img	#Rep	%Ind	%PI	%PR
Train	239,998	150,957	66.4	60.5	60.5	69,641	34,763	64.6	52.7	52.7	181,312	90,656	67.1	51.6	51.6
Val	2,113	1,182	65.4	61.3	0.0	586	263	62.7	50.6	0.0	1,778	889	72.1	36.2	0.0
Test	3,852	2,343	57.3	87.7	0.0	844	378	56.3	89.2	0.0	3,000	1500	68.9	55.8	0.0

Table 1. Statistics of the three datasets for the training, validation, and test sets. “#Img” and “#Rep” denote the number of images and reports, while “%Ind”, “%PI”, and “%PR” represent the ratios of “INDICATION”, “previous image”, and “previous report”, respectively.

and corresponding reports for cross-modal alignment could lead to model hallucinations. To address this, We first extract spatiotemporal features from multi-view longitudinal data using the multi-view longitudinal fusion network (see Figure 3(A)). Subsequently, we utilize the inherent spatiotemporal information from radiology reports to supervise the pre-training of visual and textual representations. Inspired by CLIP [40] and MGCA [48], we employ instance-wise cross-modal alignment to learn uni-modal representations. Specifically, we compute the image-to-text predicted categorical distribution $\mathbf{q}^{v2r} \in \mathbb{R}^{B \times B}$, defined as:

$$\mathbf{q}^{v2r} = \frac{\exp(\bar{\mathbf{V}} \cdot \bar{\mathbf{R}}^T / \tau_2)}{\sum_{j=1}^B \exp(\bar{\mathbf{V}} \cdot \bar{\mathbf{R}}_j^T / \tau_2)}, \quad (6)$$

where $\bar{\mathbf{V}}$ and $\bar{\mathbf{R}}$ denote global features of spatiotemporal features \mathbf{V}^{st} and textual features \mathbf{R} , respectively. Similarly, we can also obtain the symmetric text-to-image predicted categorical distribution \mathbf{q}^{r2v} . We consider spatiotemporal and textual features extracted from the same visit as positive pairs. Accordingly, the image-to-text ground-truth categorical distribution $\mathbf{p}^{v2r} \in \mathbb{R}^{B \times B}$ is defined as:

$$\mathbf{p}_{i,j}^g = \frac{\mathbb{I}_{\text{identical}}(y_i^{cur}, y_j^{cur})}{\sum_{k=1}^B \mathbb{I}_{\text{identical}}(y_i^{cur}, y_k^{cur})}, \quad (7)$$

where $\mathbb{I}_{\text{identical}}(\cdot, \cdot)$ denotes an indicator function that determines whether two radiology reports are identical. Finally, the cross-modal alignment loss is formulated as:

$$\mathcal{L}_G = -\frac{1}{B} \sum_{i=1}^B (\mathbf{p}_i^g \log \mathbf{q}_i^{v2r} + \mathbf{p}_i^g \log \mathbf{q}_i^{r2v}). \quad (8)$$

Overall objective in Stage 1. We train our MLRG by jointly optimizing \mathcal{L}_{MPC} and \mathcal{L}_G , formulated as: $\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{MPC} + \mathcal{L}_G$.

3.3. Chest X-ray Report Generation

Integrating patient-specific prior knowledge into the text generator. Radiologists commonly refer to patient-specific prior knowledge, such as the “INDICATION”

(which outlines the visit reasons or symptoms) and the “previous report” (which provides the patient’s medical history), when drafting radiology reports. However, these data may be absent for some patients due to de-identification or incomplete records. To combat this issue, we propose a tokenized absence encoding technique to handle missing patient-specific prior knowledge, as shown in Figure 2. Specifically, for missing “INDICATION” and “previous report”, we utilize special tokens, “[NHI]” and “[NHPR]”, to simulate their presence. For existing “INDICATION”, we apply a preprocessing strategy from SEI [29] to remove de-identification noise (e.g., *-year-old*, *---*, and **). For available “previous report”, we employ the structural entities approach [30] to extract factual serialization, enabling the model to focus on clinically relevant details. We then combine the cleaned “INDICATION” with factual serialization extracted from the “previous report” into a cohesive paragraph, supplemented by segment embeddings to help the model distinguish between different sentence meanings. Finally, we utilize the multi-modal fusion network [10] (illustrated in Figure 3(B)) to flexibly integrate patient-specific prior knowledge into the spatiotemporal features \mathbf{V}^{st} , allowing the text generator to generate more accurate radiology reports based on available prior knowledge.

Report generation. We start by initializing the projection heads, MLF network, and CXR-BERT using the model trained in Stage 1. Following prestigious works [35, 49], we treat the DistilGPT2 [41], initialized by CGPT2 [35], as the text generator. We then minimize the cross-entropy loss \mathcal{L}_{CE} to ensure that the generated reports closely align with the reference reports.

4. Experiments

4.1. Experimental Settings

Datasets. 1) **MIMIC-CXR** [20] is a large-scale, publicly available dataset comprising paired chest X-rays and radiology reports. Each pair contains a varying number of images compared to others, and all pairs for a patient are organized chronologically, facilitating the construction of multi-view longitudinal data. 2) **MIMIC-ABN** [34] is a subset of MIMIC-CXR, focusing solely on radiology reports that describe abnormal clinical findings. 3) **Two-view**

Dataset	Input	Method	Venue	NLG Metrics \uparrow						CE Metrics \uparrow				
				B-1	B-2	B-3	B-4	MTR	R-L	RG	P	R	F1	
M-CXR	SI	SA [56]	EMNLP'23	-	0.184	-	-	-	-	-	0.228	-	-	0.394
	SI	MET [53]	CVPR'23	0.386	0.250	0.169	0.124	0.152	0.291	-	0.364	0.309	0.311	
	SI	KiUT [17]	CVPR'23	0.393	0.243	0.159	0.113	0.160	0.285	-	0.371	0.318	0.321	
	SI	CoFE [26]	ECCV'24	-	-	-	0.125	0.176	<u>0.304</u>	-	0.489	0.370	0.405	
	SI	MAN [42]	AAAI'24	0.396	0.244	0.162	0.115	0.151	0.274	-	0.411	0.398	0.389	
	SI	B-LLM [28]	AAAI'24	<u>0.402</u>	<u>0.262</u>	<u>0.180</u>	0.128	<u>0.175</u>	0.291	-	0.465	0.482	<u>0.473</u>	
	SI	DCG [27]	ACMMM'24	0.397	0.258	0.166	0.126	0.162	0.295	-	0.441	0.414	0.404	
	SI	Med-LLM [31]	ACMMM'24	-	-	-	0.128	0.161	0.289	-	0.412	0.373	0.395	
	SI+Ind	SEI [29]	MICCAI'24	0.382	0.247	0.177	<u>0.135</u>	0.158	0.299	<u>0.249</u>	<u>0.523</u>	0.410	0.460	
	MVD	FMVP [32]	TMM'23	0.389	0.236	0.156	0.108	0.150	0.284	-	0.332	0.383	0.336	
	Long	HERGen [49]	ECCV'24	0.395	0.248	0.169	0.122	0.156	0.285	-	-	-	-	
	MVL	CXRMate [36]	arXiv'23	0.361	0.223	0.150	0.108	0.159	0.263	0.238	0.495	0.367	0.422	
	MVL	MLRG(Ours)	-	0.411	0.277	0.204	0.158	0.176	0.320	0.291	0.549	<u>0.468</u>	0.505	
	-	Δ (%) \uparrow	-	+0.9	+1.5	+2.4	+2.3	+0.1	+1.6	+4.2	+2.6	-1.4	+3.2	
M-ABN	SI	R2Gen ^b [8]	EMNLP'20	0.253	0.144	0.092	0.063	0.106	0.229	0.179	0.444	0.425	0.434	
	SI	CMN ^b [9]	ACL'21	0.256	0.147	0.095	0.066	0.110	0.230	0.183	<u>0.466</u>	<u>0.454</u>	<u>0.460</u>	
	SI+Ind	SEI ^b [29]	MICCAI'24	<u>0.267</u>	<u>0.157</u>	<u>0.104</u>	<u>0.073</u>	<u>0.114</u>	<u>0.231</u>	<u>0.191</u>	<u>0.466</u>	0.408	0.435	
	MVL	MLRG(Ours)	-	0.332	0.199	0.132	0.094	0.136	0.248	0.219	0.513	0.517	0.515	
	-	Δ (%) \uparrow	-	+6.5	+4.2	+2.8	+2.1	+2.2	+1.7	+2.8	+4.7	+6.3	+5.5	
T-CXR	DV	R2Gen ^b [8]	EMNLP'20	0.346	0.219	0.153	0.113	0.141	0.302	0.267	0.478	0.329	0.390	
	DV	CMN ^b [9]	ACL'21	0.387	0.241	0.166	0.122	0.151	0.310	0.268	0.496	0.336	0.401	
	DV+Ind	SEI ^b [29]	MICCAI'24	<u>0.409</u>	<u>0.263</u>	<u>0.186</u>	<u>0.140</u>	<u>0.168</u>	<u>0.320</u>	<u>0.301</u>	<u>0.522</u>	<u>0.447</u>	<u>0.481</u>	
	MVL	MLRG(Ours)	-	0.417	0.276	0.200	0.154	0.178	0.331	0.328	0.532	0.474	0.501	
	-	Δ (%) \uparrow	-	+0.8	+1.3	+1.4	+1.4	+1.0	+1.1	+2.7	+1.0	+2.7	+2.0	

Table 2. Comparison with SOTA methods on MIMIC-CXR (M-CXR), MIMIC-ABN (M-ABN), and Two-view CXR (T-CXR) datasets. Δ denotes the performance difference between MLRG and the best peer methods. ^b signifies results reproduced using official codes, while other results are sourced from original publications. The best and second-best values are emphasized in **bold** and underlined, respectively.

CXR [33] aggregates visits with two current images from both MIMIC-CXR and IU X-ray [11]. Notably, as the IU X-ray does not include previous visits, both previous images and reports are unavailable. We adhere to official splits for these datasets and summarize the sample counts for the training, validation, and test set in Table 1. In line with [5, 8, 25, 29, 45], we treat the “FINDINGS” section in radiology reports as the reference reports.

Evaluation metrics. Following prior works [4, 8, 49, 51, 53], we evaluate the effectiveness of our MLRG using both natural language generation (NLG) and clinical efficacy (CE) metrics. NLG metrics, which assess the linguistic similarities between generated and reference reports, include BLEU-n (B-n), METEOR (MTR), and ROUGE-L (R-L). For CE metrics, we utilize CheXpert [18] to label the generated reports with 14 observations (see Appendix Table A1) and compute the micro-average Precision (P), Recall (R), and F1 score (F1) based on ground truths. CE metrics also include F1 RadGraph (RG) [19], which evaluates

the overlap of clinical entities and their relations, aligning more closely with radiologists than B-3 and F1 metrics [60]. All metrics are computed using official libraries [7, 19, 44], with higher values indicating better performance.

Implementation details. Both τ_1 and τ_2 are set to 0.5. The number of blocks L_1 and L_2 in Figure 3 are set to 3 and 1, respectively. Each dataset is configured to generate a maximum of 100 tokens. We identify the best model as the one with the highest sum of BLEU-4, F1 RadGraph, and F1 score on the validation set, and we report its performance on the test set. Additional details about epochs, learning rates, and other settings can be found in the Appendix A.1.

4.2. Main Results

We compare our MLRG with 14 state-of-the-art methods: R2Gen [8], CMN [9], SA [56], MET [53], KiUT [17], CoFE [26], MAN [42], B-LLM [28], DCG [27], Med-LLM [31], SEI [29], FMVP [32], HERGen [49], and CXRMate [36]. Results are presented in Table 2, where “SI”, “Ind”, “MVD”, “Long”, “MVL”, and “DV” represent different in-

Model	M/S	F/R	PI	Stage 1		Stage 2		NLG metrics \uparrow						CE metrics \uparrow	
				\mathcal{L}_G	\mathcal{L}_{MPC}	Ind	PR	B-1	B-2	B-3	B-4	MTR	R-L	RG	F1
(a)	M	F	✓	✗	✗	✓	✓	0.346	0.235	0.173	0.136	0.154	0.305	0.258	0.373
(b)	M	F	✓	✓	✓	✗	✗	0.385	0.239	0.162	0.118	0.155	0.283	0.238	0.479
(c)	M	F	✗	✓	✗	✓	✓	0.384	0.257	0.188	0.146	0.165	0.310	0.267	0.455
(d)	M	F	✗	✓	✓	✓	✓	0.395	0.257	0.183	0.138	0.167	0.302	0.278	0.503
(e)	M	F	✓	✓	✓	✓	✗	0.392	0.265	0.195	0.153	0.171	0.316	0.281	0.476
(f)	M	F	✓	✓	✓	✗	✓	0.387	0.240	0.163	0.118	0.156	0.281	0.243	0.484
(g)	M	R	✓	✓	✓	✓	✓	0.403	0.267	0.193	0.148	0.172	0.309	0.287	0.510
(h)	S	F	✓	✓	✗	✓	✓	0.400	0.269	0.196	0.151	0.173	0.314	0.289	0.498
MLRG	M	F	✓	✓	✓	✓	✓	0.411	0.277	0.204	0.158	0.176	0.320	0.291	0.505

Table 3. Ablation study on the MIMIC-CXR dataset. “M/S” refers to methods that utilize current **Multi-view** images or **Single** images as input. “F/R” indicates alignment based on either **Factual** serialization or **Report**. “PI”, “PR”, and “Ind” represent **Previous Images**, **Previous Reports**, and “INDICATION”, respectively. The best values are emphasized in **bold**.

Setting	%	B-2 \uparrow	B-4 \uparrow	MTR \uparrow	R-L \uparrow	RG \uparrow
w/ Ind	57.8	0.295	0.174	0.184	0.332	0.318
w/o Ind	42.2	0.253	0.137	0.166	0.302	0.254
w/ MV	70.7	0.282	0.161	0.179	0.323	0.301
w/o MV	29.3	0.264	0.150	0.171	0.310	0.266
w/ MVL	61.4	0.282	0.160	0.178	0.322	0.300
w/o MVL	38.6	0.270	0.155	0.174	0.316	0.276

Table 4. Breakdown of MLRG’s metrics on the MIMIC-CXR test set, categorized by (a) inclusion of indications (Ind), (b) inclusion of current multi-view images (MV), (c) inclusion of multi-view longitudinal data (MVL).

put types: single images, “INDICATION”, multi-view data, longitudinal data, multi-view longitudinal data, and dual views, respectively. We observe that our MLRG achieves SOTA performance across most metrics, with particular strength in B-4, RG, and F1. This suggests that MLRG excels in generating both coherent and accurate radiology reports. Although MLRG shows slightly lower Recall than B-LLM [28], its F1 and other metrics are significantly better. In Appendix Table A3, we also show our MLRG’s ability to generate “FINDINGS” and “IMPRESSION” sections.

4.3. Ablation Study

Table 3 presents an ablation study on the MIMIC-CXR [20] dataset, analyzing the effect of different components on model performance.

Effect of multi-view longitudinal contrastive learning (Stage 1). In Table 3, (a) represents a report generation scheme based solely on patient-specific prior knowledge, excluding Stage 1. Results reveal that our MLRG significantly exceeds (a), highlighting the critical role of multi-view longitudinal contrastive learning in enhancing the accuracy and coherence of generated reports. Additionally,

we observe that both \mathcal{L}_G ((c) vs. (a)) and \mathcal{L}_{MPC} ((d) vs. (c)) have a positive impact on model performance.

Effect of patient-specific prior knowledge (Stage 2). As shown in Table 3, MLRG significantly outperforms (b), which lacks patient-specific prior knowledge, emphasizing the importance of incorporating such knowledge to improve the coherence and clinical accuracy of generated reports. Moreover, the independent integration of “INDICATION” ((e) vs. (b)) and “previous report” ((f) vs. (b)) contributes positively to model performance.

Effect of current multi-view images. Table 3 demonstrates that generating reports using current multi-view images outperforms those derived from single images (MLRG vs. (h)), highlighting the effectiveness of multi-view images in modeling the current disease conditions.

Effect of previous images. As shown in Table 3, MLRG shows a clear advantage over (d), indicating that MLF network (see Figure 3) effectively integrates previous images. This capability allows the model to track disease progression, thereby generating more clinically accurate reports.

4.4. Case Study

Model benefits from current multi-view images, multi-view longitudinal data, and “INDICATION”. Table 4 compares performance on test subsets with and without these data (multi-view images, multi-view longitudinal data, and indications). We observe that including these data significantly improves NLG and RG metrics. This suggests that the multi-view longitudinal contrastive learning method effectively integrates current multi-view images and multi-view longitudinal data, capturing semantically rich visual representations. Moreover, our multi-modal fusion network effectively leverages “INDICATION” to help the text generator produce more accurate radiology reports.

Clinical accuracy of 14 observations. Appendix Tables A1 and A2 show the clinical accuracy of 14 observations

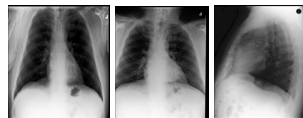

 <table border="1" data-bbox="162 315 462 388"> <thead> <tr> <th>B-4↑</th> <th>RG↑</th> <th>Ind</th> <th>PI</th> <th>PR</th> </tr> </thead> <tbody> <tr> <td>0.608/0.000</td> <td>0.838/0.323</td> <td>✓</td> <td>✓</td> <td>×</td> </tr> </tbody> </table>	B-4↑	RG↑	Ind	PI	PR	0.608/0.000	0.838/0.323	✓	✓	×	<p>Reference report: the lungs are well expanded and clear. cardiomediastinal and hilar contours are unremarkable. there is no pleural effusion or pneumothorax. sternotomy wires are again noted, with fracture of the two upper wires unchanged from prior exam.</p>	<p>Generated report by SEI: frontal and lateral views of the chest were obtained . the patient is status post median sternotomy and cabg. no focal consolidation, pleural effusion, or evidence of pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.</p>	<p>Generated report by MLRG (Ours): the lungs are well expanded and clear. cardiomediastinal and hilar contours are unremarkable. there is no pleural effusion or pneumothorax. sternotomy wires are intact.</p>
B-4↑	RG↑	Ind	PI	PR									
0.608/0.000	0.838/0.323	✓	✓	×									
 <table border="1" data-bbox="162 535 462 609"> <thead> <tr> <th>B-4↑</th> <th>RG↑</th> <th>Ind</th> <th>PI</th> <th>PR</th> </tr> </thead> <tbody> <tr> <td>0.804/0.288</td> <td>0.769/0.258</td> <td>✓</td> <td>✓</td> <td>×</td> </tr> </tbody> </table>	B-4↑	RG↑	Ind	PI	PR	0.804/0.288	0.769/0.258	✓	✓	×	<p>Reference report: as compared to the previous radiograph, the patient has received a new right internal jugular vein catheter. the course of the catheter is unremarkable, the tip of the catheter projects over the lower svc. there is no evidence of complications, notably no pneumothorax. otherwise unchanged radiographic appearance.</p>	<p>Generated report by SEI: as compared to the previous radiograph, the patient has been extubated and the nasogastric tube has been removed. the right internal jugular vein catheter and the right internal jugular vein catheter are in unchanged position. the lung volumes have slightly decreased . there is no evidence of pneumothorax. the lung volumes remain low . moderate cardiomegaly persists.</p>	<p>Generated report by MLRG (Ours): as compared to the previous radiograph, the patient has received a right internal jugular vein catheter. the course of the catheter is unremarkable, the tip of the catheter projects over the mid svc. there is no evidence of complications, notably no pneumothorax. otherwise, the radiograph is unchanged.</p>
B-4↑	RG↑	Ind	PI	PR									
0.804/0.288	0.769/0.258	✓	✓	×									

Figure 4. Generated reports examples on the MIMIC-CXR test set. Each “A/B” cell refers to “MLRG/SEI”. Sentences in the reference report are highlighted in unique colors to clarify alignment with descriptions in the generated reports. Matching content in generated reports is shown in the same color, while correct temporal descriptions and failure descriptions of our MLRG are in **bold** and underlined.

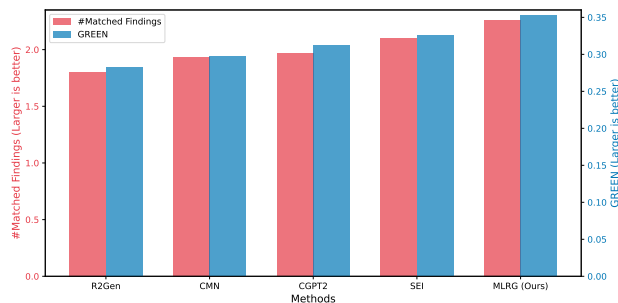


Figure 5. Comparison with baselines on MIMIC-CXR using LLMs. “#Matched Findings” denotes the number of matched findings between generated and reference reports.

labeled by CheXpert [18] across three datasets. Results indicate that our MLRG outperforms SEI [29] on most observations. Although not specifically tailored for imbalanced observations, MLRG still slightly exceeds the baseline on challenging observations like *Pneumothorax* and *Fracture*.

Qualitative analysis. Figure 4 presents examples of generated reports using SEI [29] and our MLRG, with additional examples in Appendix Figure A1. A greater number of colors in the generated report indicates broader coverage of clinical findings, while a longer color bar reflects more accurate and detailed descriptions of specific findings. Results indicate that 1) Our MLRG provides radiologists with higher-quality draft reports compared to SEI [29]; 2) Our MLRG exhibits a certain ability to describe disease progression, as evidenced in case 2 with the statement “the patient has received a right internal jugular vein catheter”.

Evaluation using large language models (LLMs). Figure 5 illustrates the performance of our generated reports

evaluated with GREEN [38], a fine-tuned LLaMA 2 [47] designed to identify clinically significant errors and count matched findings. Results demonstrate that our MLRG outperforms R2Gen [8], CMN [9], CGPT2 [35], and SEI [29] in both #Matched Findings and GREEN score, further confirming the advantage of our MLRG in generating coherent and clinically accurate radiology reports. For further details, please refer to the Appendix A.5.

5. Conclusion

In this paper, we introduced the MLRG method for chest X-ray report generation. We first proposed a multi-view longitudinal contrastive learning approach that leveraged the inherent spatiotemporal information from radiology reports to guide the pre-training of visual and textual representations. This approach not only captured differences among views but also flexibly extracted spatial features from current multi-view images and temporal features from longitudinal data, effectively leveraging spatiotemporal information in reports for pre-training. Subsequently, we presented a tokenized absence encoding technique to handle missing patient-specific prior knowledge. This technique allowed the multi-modal fusion network to adapt flexibly to scenarios with or without such data, ensuring the text generator can utilize available prior knowledge effectively. Extensive experiments on MIMIC-CXR, MIMIC-ABN, and Two-view datasets demonstrated that our MLRG outperforms existing SOTA methods in generating coherent and clinically accurate radiology reports, making it a strong contender for chest X-ray report generation. Future work will focus on using saliency maps [57] to learn region-based features and predict uncertainty [21] to improve model reliability.

Acknowledgments

The work was jointly supported by the National Science and Technology Major Project (No. 2022ZD0117103), the National Natural Science Foundations of China (Nos. 62272364 and 62202360), the provincial Key Research and Development Program of Shaanxi (No. 2024GH-ZDXM-47), the Research Project on Higher Education Teaching Reform of Shaanxi Province (No. 23JG003), the Fundamental Research Funds for the Central Universities (No. ZYTS24090), the Innovation Fund of Xidian University, and the High-Performance Computing Platform of Xidian University.

References

- [1] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023. 2
- [2] Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. 1
- [3] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 3, 4
- [4] Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. 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 (CVPR)*, pages 14194–14204, 2024. 6
- [5] Yiming Cao, Lizhen Cui, Lei Zhang, Fuqiang Yu, Zhen Li, and Yonghui Xu. Mmtm: Multi-modal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 277–285, 2023. 6
- [6] Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9494–9509, 2024. 2
- [7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 6
- [8] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020. 1, 2, 6, 8
- [9] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 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)*, pages 5904–5914, 2021. 2, 6, 8
- [10] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23346–23356, 2023. 5
- [11] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 6
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3
- [13] Li Guo, Anas M. Tahir, Dong Zhang, Z. Jane Wang, and Rabab K. Ward. Automatic medical report generation: Methods and applications, 2024. 1
- [14] Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–199. Springer, 2024. 1
- [15] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024. 1
- [16] Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. RECAP: towards precise radiology report generation via dynamic disease progression reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2134–2147, 2023. 2
- [17] Z. Huang, X. Zhang, and S. Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023. 2, 6
- [18] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins,

- David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597, 2019. 1, 6, 8
- [19] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems*, 2021. 6
- [20] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019. 2, 5, 7
- [21] Xiaolu Kang, Zhuoqi Ma, Kang Liu, Yunan Li, and Qiguang Miao. Multi-scale information sharing and selection network with boundary attention for polyp segmentation, 2024. 8
- [22] Gaurang Karwande, Amarachi Mbakawe, Joy T. Wu, Leo A. Celi, Mehdi Moradi, and Ismini Lourentzou. Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part I*, pages 581–591. Springer, 2022. 2
- [23] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 2
- [25] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023. 6
- [26] Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsadik, and Xiaojun Chang. Contrastive learning with counterfactual explanations for radiology report generation. In *Computer Vision – ECCV 2024*, pages 162–180, Cham, 2025. Springer Nature Switzerland. 6
- [27] Xiao Liang, Yanlei Zhang, Di Wang, Haodi Zhong, Ronghan Li, and Quan Wang. Divide and conquer: Isolating normal-abnormal attributes in knowledge graph-enhanced radiology report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 4967–4975, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 6
- [28] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18635–18643, 2024. 2, 6, 7
- [29] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhushi Zhong, Zhicheng Jiao, Grayson Baird, Harrison Bai, and Qiguang Miao. Structural entities extraction and patient indications incorporation for chest x-ray report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 433–443, Cham, 2024. Springer Nature Switzerland. 1, 2, 5, 6, 8
- [30] Kang Liu, Zhuoqi Ma, Mengmeng Liu, Zhicheng Jiao, Xiaolu Kang, Qiguang Miao, and Kun Xie. Factual serialization enhancement: A key innovation for chest x-ray report generation, 2024. 1, 2, 3, 5
- [31] Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 8721–8730, New York, NY, USA, 2024. Association for Computing Machinery. 2, 6
- [32] Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. From observation to concept: A flexible multi-view paradigm for medical report generation. *IEEE Transactions on Multimedia*, 26:5987–5995, 2024. 1, 2, 6
- [33] Qiguang Miao, Kang Liu, Zhuoqi Ma, Yunan Li, Xiaolu Kang, Ruixuan Liu, Tianyi Liu, Kun Xie, and Zhicheng Jiao. Evoke: Elevating chest x-ray report generation via multi-view contrastive learning and patient-specific knowledge, 2025. 2, 6
- [34] Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian J. McAuley. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1954–1960, 2020. 2, 5
- [35] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. 1, 2, 5, 8
- [36] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation, 2023. 2, 3, 6
- [37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3
- [38] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics*:

- EMNLP 2024*, pages 374–390, Miami, Florida, USA, 2024. Association for Computational Linguistics. 8
- [39] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Rad-dino: Exploring scalable medical image encoders beyond text supervision, 2024. 3
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 5
- [41] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 3, 5
- [42] Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4776–4783, 2024. 6
- [43] Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, pages 1–20, 2024. 1
- [44] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 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. 6
- [45] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 2, 6
- [46] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 4
- [47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 8
- [48] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhati, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*, pages 33536–33549, 2022. 2, 5
- [49] Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal data. In *Computer Vision – ECCV 2024*, pages 183–200, Cham, 2025. Springer Nature Switzerland. 1, 2, 5, 6
- [50] Xinyi Wang, Graziela Figueredo, Ruizhe Li, Wei Emma Zhang, Weitong Chen, and Xin Chen. A survey of deep learning-based radiology report generation using multimodal data, 2024. 1
- [51] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2433–2442, 2021. 6
- [52] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023. 2
- [53] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023. 2, 6
- [54] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383, 2023. 2
- [55] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning, 2013. 2
- [56] Benjamin Yan, Ruochen Liu, David E. Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe O’Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. Style-aware radiology report generation with radgraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14676–14688, 2023. 1, 6
- [57] Honglong Yang, Hui Tang, and Xiaomeng Li. Fita: Fine-grained image-text aligner for radiology report generation, 2024. 8
- [58] Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510, 2022. 2
- [59] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023. 2
- [60] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. 6
- [61] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23465–23476, 2023. 2
- [62] Ke Zhang, Yan Yang, Jun Yu, Jianping Fan, Hanliang Jiang, Qingming Huang, and Weidong Han. Attribute prototype-guided iterative scene graph for explainable radiology report

- generation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. [2](#)
- [63] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023. [2](#)
- [64] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Loddon Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12910–12917, 2020. [2](#)
- [65] Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M. Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 189–198, Cham, 2023. Springer Nature Switzerland. [2](#)